Permissionless Finance

From Perpetual Futures to the On-Chain Global Market — A Market Microstructure Perspective

Eric Cheung

2026

Contents

Acknowledgments **i**

Disclaimer **ii**

Preface **iii**
Why This Book Exists . . . iii
Who This Book Is For . . . iii
What You Will Gain . . . iv
How This Book Is Organized . . . iv
The Landscape in 2026 . . . v
Why I Believe in This Future . . . v

About the Author **vi**

Part I. Foundations **1**

The Distinctiveness of Digital Assets Compared with Traditional Assets **2**
1.1 The trust architecture of traditional finance . . . 3
1.1.1 The complete journey of a trade . . . 3
1.1.2 The five layers of trust . . . 6
1.1.3 The trust tax: a unified view of institutional cost . . . 9
1.1.4 The rationale for the trust tax . . . 11
1.2 The six distinctive properties and the compression of the trust tax . . . 12
1.2.1 Permissionlessness: compressing the entry tax . . . 12
1.2.2 Atomic divisibility: compressing the granularity barrier . . . 14
1.2.3 Settlement finality: compressing the time tax . . . 15
1.2.4 Around-the-clock trading: compressing temporal fragmentation . . . 16
1.2.5 Smart-contract execution: compressing the intermediary tax . . . 17
1.2.6 On-chain transparency: the opacity tax and the transparency paradox 18
1.3 The composability thesis . . . 20
1.3.1 The interaction matrix of the six properties . . . 20
1.3.2 Financial primitives impossible in traditional finance . . . 21
1.3.3 The dark side of composability . . . 22
1.4 Microstructure reconstruction and the transformation of the risk surface . 24

1.4.1 The reconstruction of the three microstructure elements 24
1.4.2 The transformation of the risk surface 25
1.4.3 Three trust paradigms 27
1.5 Chapter summary 29
References 30

Derivatives Markets and the Rise of Perpetual Futures **33**
2.1 The trust evolution of derivatives 34
2.1.1 Personalized trust and forward contracts 34
2.1.2 Institutionalized trust and the futures revolution 35
2.1.3 Options and nonlinear payoffs 35
2.1.4 A unified functional framework 36
2.1.5 Algorithmic trust and perpetual futures 37
2.2 The three structural conflicts 37
2.2.1 Temporal mismatch 38
2.2.2 Liquidity fragmentation 38
2.2.3 Trust mismatch 39
2.2.4 Ruling out the alternatives 39
2.3 The core mechanisms of perpetual futures 42
2.3.1 The funding rate 42
2.3.2 Mark price and the liquidation mechanism 45
2.3.3 Inverse contracts and linear contracts 47
2.3.4 The cost of abolishing the expiry date 48
2.4 Market landscape and ecosystem evolution 49
2.4.1 Market size and price discovery 49
2.4.2 The democratization of leverage 50
2.4.3 Leverage-cycle dynamics 51
2.4.4 Risks unique to DEXs 55
2.4.5 Regulatory evolution 57
2.5 Chapter summary 58
References 60

The Distinctive Value of Digital Assets and the Future of Finance **62**
3.1 An economic reexamination of the four distinctive properties 63
3.1.1 Permissionlessness and contestable markets 64
3.1.2 Around-the-clock continuous trading and the transformation of constraints 66
3.1.3 The scope of atomic settlement 67
3.1.4 The power and cost of programmability 70
3.2 The institutional convergence thesis 71
3.2.1 The on-chain deployment of traditional finance 71
3.2.2 The institutionalization of on-chain finance 73
3.2.3 The competitive compression of the trust tax 73
3.2.4 The institutional-convergence significance of the BUIDL fund . . . 75
3.3 The trust-tax compression curve and the three-horizon framework 76
3.3.1 The compression-difficulty gradient 76

3.3.2 H1: Stablecoin settlement and crypto-native derivatives 79
3.3.3 H2: Tokenized securities and institutional DeFi 80
3.3.4 H3: Programmable money and autonomous finance 81
3.3.5 Differences in the trust needs of the three horizons 82
3.4 Stablecoins: a structural migration of the settlement layer 83
3.4.1 The meaning of $33 trillion in settlement volume 83
3.4.2 The network effects and escape velocity of stablecoins 85
3.4.3 The upper-layer effects after the settlement layer is established . . 86
3.4.4 The structural risks of stablecoins 86
3.5 Asset tokenization and on-chain settlement 87
3.5.1 The economic logic of asset tokenization 87
3.5.2 The architectural advantages of on-chain settlement 89
3.5.3 Three surmountable challenges 90
3.6 The impossibility boundaries 92
3.6.1 The sovereignty boundary 92
3.6.2 The physical-world interface 93
3.6.3 The governance paradox 94
3.6.4 Maximizing within boundaries 95
3.7 Chapter summary 95
References 97

Part II. The Ecosystem **100**

Participants and Behavior in Perpetual Futures Trading **101**
4.1 The participant classification framework 102
4.1.1 Functional classification 102
4.1.2 The two ultimate sources of profit 103
4.1.3 Behavioral classification 106
4.1.4 The hidden costs of the negative-sum game 108
4.1.5 Limitations of the classification 109
4.2 Behavioral analysis of the six core participant types 109
4.2.1 Market makers 109
4.2.2 Informed traders 113
4.2.3 Arbitrageurs 114
4.2.4 Mechanism parasites 116
4.2.5 Speculators and futile traders 119
4.2.6 Liquidators 120
4.3 The game network among participants 123
4.3.1 Market makers and informed traders 123
4.3.2 Market makers and mechanism parasites 124
4.3.3 Profit-motivated traders and futile traders 125
4.3.4 Speculators and the liquidation mechanism 126
4.3.5 The complete map of profit flows 128
4.4 The threefold reshaping by the digitally native environment 130
4.4.1 The fluidization of roles 131

4.4.2 The algorithmization of the game . . . 132
4.4.3 The transparency-fragility paradox . . . 134
4.4.4 The compounding effect of the threefold reshaping . . . 135
4.5 AI agents: the rise of a seventh participant type . . . 136
4.5.1 From automation to autonomous agency . . . 137
4.5.2 AI agents' competition for the two profit sources . . . 137
4.5.3 Differentiated impact on the six participant types . . . 139
4.5.4 Three entirely new forms of the game . . . 140
4.5.5 Revisions to the analytical framework . . . 142
4.6 Comparison with traditional finance . . . 143
4.6.1 The invariant underlying logic . . . 144
4.6.2 The changed modes of implementation . . . 144
4.6.3 Entirely new species and phenomena . . . 146
4.7 Chapter summary . . . 146
References . . . 147

Service Providers and the Ecosystem of Perpetual Futures Trading 150
5.1 The risk substitution theorem and the service provider's trilemma . . . 151
5.1.1 The functional dimensions of the trade lifecycle . . . 151
5.1.2 The risk substitution theorem . . . 153
5.1.3 The service provider's trilemma . . . 155
5.1.4 The verifiability ladder . . . 157
5.2 The functional-separation architecture of traditional finance . . . 160
5.2.1 The multi-intermediary flow of a futures trade . . . 160
5.2.2 The institutional rationale for functional separation . . . 162
5.2.3 The CCP default waterfall mechanism . . . 163
5.2.4 The efficiency cost of functional separation . . . 166
5.3 The functional bundling of centralized exchanges . . . 167
5.3.1 The CEX's bundling of five roles . . . 168
5.3.2 The definition and composition of the bundling tax . . . 170
5.3.3 The efficiency advantages of the bundling model . . . 172
5.3.4 Comparing the CEX and CCP risk waterfalls . . . 173
5.3.5 The limits of proof of reserves . . . 175
5.3.6 The FTX and Bybit failure cases . . . 176
5.4 On-chain trading models . . . 178
5.4.1 Self-custody and atomic settlement . . . 178
5.4.2 A comparison of three execution models . . . 180
5.4.3 App-chain architecture and validator centralization . . . 182
5.4.4 Architecture audit and verifiability rating . . . 184
5.4.5 The on-chain risk waterfall . . . 186
5.5 Cross-architecture risk isomorphism . . . 189
5.5.1 Adverse selection and impermanent loss . . . 189
5.5.2 Margin cascades and liquidation cascades . . . 190
5.5.3 Market fragmentation and cross-chain fragmentation . . . 191
5.5.4 Closing gaps and low-liquidity amplification . . . 192
5.5.5 The cross-modal risk transmission path . . . 193

5.6 Effects on market microstructure . . . 196
5.6.1 The reconstruction of price discovery . . . 196
5.6.2 The evolution of liquidity supply . . . 197
5.6.3 Structural differences in transaction costs . . . 198
5.6.4 The distribution of systemic risk . . . 199
5.7 Chapter summary . . . 200
References . . . 201

The Evolution and Trade-offs of Perpetual Futures Market Structure 206
6.1 The trade lifecycle . . . 207
6.1.1 Centralized exchanges . . . 208
6.1.2 General-purpose public-chain DEXs . . . 210
6.1.3 Application-chain DEXs . . . 213
6.1.4 Structural comparison . . . 216
6.2 The competition-mode migration thesis . . . 218
6.2.1 Latency competition in CEXs . . . 218
6.2.2 Information competition in general-purpose public-chain DEXs . . . 219
6.2.3 Mechanism competition in application-chain DEXs . . . 220
6.2.4 A unified comparison of the competition modes . . . 221
6.2.5 The drivers of competition-mode migration . . . 224
6.3 Strategic MEV . . . 226
6.3.1 Atomic MEV and strategic MEV . . . 226
6.3.2 Typical forms . . . 229
6.3.3 Defending against strategic MEV . . . 232
6.3.4 The path toward institutionalizing MEV . . . 233
6.4 The price-discovery migration hypothesis . . . 235
6.4.1 Structural drivers . . . 236
6.4.2 Critical mass and network effects . . . 238
6.4.3 A cautious reading of the current evidence . . . 238
6.4.4 The shaping of trader behavior . . . 240
6.5 The evolution of risk distribution . . . 242
6.5.1 The risk visibility-manageability proposition . . . 242
6.5.2 Statistical properties of risk . . . 244
6.5.3 Cross-paradigm risk transmission . . . 247
6.6 The paradigm endgame . . . 249
6.6.1 Coexistence and the competitive landscape . . . 250
6.6.2 Convergence dynamics . . . 251
6.6.3 Expanding the impossible triangle . . . 253
6.6.4 Open problems . . . 257
6.7 Chapter summary . . . 260
References . . . 260

Part III. Microstructure 265

Orders and the Language of Perpetual Futures Trading 266

7.1 The signaling function of orders . . . 267
7.1.1 The trading function and the signaling function . . . 267
7.1.2 The signal-strength spectrum . . . 269
7.1.3 The microfoundations of adverse selection . . . 272
7.1.4 The intent-execution gap . . . 274
7.2 Basic order types . . . 277
7.2.1 Market orders . . . 277
7.2.2 Limit orders . . . 278
7.2.3 Stop-loss and take-profit orders . . . 278
7.2.4 The maker-taker fee model and order selection . . . 280
7.3 Advanced order types . . . 282
7.3.1 Time-control orders . . . 283
7.3.2 Size-hiding orders . . . 283
7.3.3 Execution-constraint orders . . . 284
7.3.4 Algorithmic orders: TWAP and VWAP . . . 284
7.3.5 Conditional orders . . . 287
7.4 The strategic dimensions of perpetual futures . . . 288
7.4.1 The leverage dimension . . . 289
7.4.2 The liquidation dimension . . . 290
7.4.3 The time dimension . . . 291
7.4.4 The choice of margin mode . . . 293
7.4.5 A unified decision framework . . . 294
7.5 The commitment property of on-chain orders . . . 296
7.5.1 The privacy of CEX orders . . . 297
7.5.2 The public-commitment property of on-chain orders . . . 298
7.5.3 The MEV exposure surface of orders . . . 298
7.5.4 The order-protection mechanisms of application chains . . . 302
7.6 The execution-architecture spectrum and the reshaping of order semantics 303
7.6.1 A comparison of the three execution architectures . . . 304
7.6.2 The order lifecycle . . . 306
7.6.3 The order-cancellation mechanism . . . 308
7.6.4 Conditional-order triggering . . . 310
7.6.5 The order-matching mechanism . . . 313
7.6.6 State-management constraints . . . 315
7.7 The order strategies of market participants . . . 317
7.7.1 The order strategies of market makers . . . 317
7.7.2 The order strategies of arbitrageurs . . . 318
7.7.3 The order strategies of speculators . . . 319
7.7.4 The order strategies of hedgers . . . 320
7.7.5 The order strategies of liquidators . . . 320
7.7.6 Participant identification based on order flow . . . 322
7.8 A three-dimensional framework for risk analysis . . . 325
7.8.1 Defining the three dimensions . . . 325
7.8.2 Applying the three-dimensional risk matrix . . . 326
7.8.3 Practical implications for risk management . . . 328
7.9 Order-flow analysis . . . 329

7.9.1 The composition of order flow . . . 329
7.9.2 Order-flow toxicity and VPIN . . . 331
7.9.3 The distinctive signals of perpetual futures order flow . . . 332
7.9.4 From orders to the order book . . . 335
7.10 Chapter summary . . . 335
References . . . 336

The Perpetual Futures Order Book and Market Microstructure 339
8.1 The order book as an information machine . . . 340
8.1.1 The structure and rules of the CLOB . . . 341
8.1.2 The five signal channels of the order book . . . 343
8.1.3 Signal timeliness and information decay . . . 346
8.1.4 From single signals to state diagnosis . . . 347
8.2 The depth illusion . . . 348
8.2.1 Flickering liquidity and ghost quotes . . . 348
8.2.2 Conditional depth . . . 349
8.2.3 The three-layer decomposition of depth . . . 350
8.2.4 The amplification mechanisms of the depth illusion . . . 352
8.2.5 Case study: the Hyperliquid POPCAT phantom order-wall incident 355
8.3 The distinctiveness of the perpetual futures order book . . . 357
8.3.1 The order book transmission mechanism of the liquidation cascade 357
8.3.2 The cyclical influence of the funding rate . . . 360
8.3.3 Order book fragility under high leverage . . . 360
8.3.4 The divergence signal between mark price and traded price . . . 362
8.3.5 Open interest as a harbinger of latent order flow . . . 364
8.4 Topological features of the order book . . . 365
8.4.1 Symmetry and asymmetry . . . 366
8.4.2 Slope and convexity/concavity . . . 366
8.4.3 Order book morphology . . . 367
8.4.4 The mapping between morphology and market state . . . 369
8.5 The four fingerprints of the perpetual futures order book . . . 370
8.5.1 The depth cliff . . . 370
8.5.2 The depth tide . . . 372
8.5.3 The liquidity circadian rhythm . . . 373
8.5.4 The asymmetric response of the spread . . . 374
8.5.5 The superposition effect of the four fingerprints . . . 375
8.6 Order book resilience . . . 378
8.6.1 The operational definition and measurement of resilience . . . 378
8.6.2 The structural disadvantage in resilience . . . 380
8.6.3 A comparison of resilience across architectures . . . 383
8.6.4 Resilience and market quality . . . 384
8.7 The evolution of matching-engine architecture . . . 385
8.7.1 Centralized exchanges . . . 386
8.7.2 General-purpose-chain DEXs . . . 387
8.7.3 The application-chain architecture . . . 387
8.7.4 A multidimensional comparison of the three architectures . . . 390

8.8 Continuous auction and batch auction . . . 392
8.8.1 The speed race and latency arbitrage . . . 392
8.8.2 The formation of toxic order flow and adverse selection . . . 393
8.8.3 The frequent batch auction . . . 394
8.8.4 The dual-flow batch auction . . . 395
8.8.5 Injective's FBA practice . . . 397
8.9 The information ecosystem of on-chain and off-chain order books . . . 398
8.9.1 The characteristics of the CEX order book . . . 399
8.9.2 The characteristics of the on-chain CLOB . . . 399
8.9.3 The democratization of data access . . . 400
8.9.4 The on-chain data-analysis ecosystem . . . 400
8.10 Chapter summary . . . 402
References . . . 403

Hyperliquid's Microstructure Innovations 406
9.1 The semantic ordering mechanism . . . 407
9.1.1 The limitations of traditional paradigms . . . 407
9.1.2 Definition and implementation . . . 409
9.1.3 Effects on market makers . . . 412
9.1.4 Effects on arbitrageurs . . . 413
9.1.5 Theoretical foundations . . . 414
9.1.6 A comparison of MEV mitigation approaches . . . 415
9.1.7 The interaction of semantic ordering with the transaction lifecycle 416
9.2 The cancellation priority mechanism . . . 417
9.2.1 Hedging adverse selection . . . 418
9.2.2 A comparison with the IEX speed bump . . . 419
9.2.3 Structural effects . . . 420
9.3 Discrete-time matching: an inherent property of on-chain matching . . . 422
9.3.1 Continuous- versus discrete-time games . . . 422
9.3.2 The displacement of speed competition . . . 423
9.3.3 A comparison with frequent batch auctions . . . 424
9.3.4 Synergy with semantic ordering . . . 425
9.3.5 The weakening of the speed race . . . 426
9.3.6 Block time as a hyperparameter . . . 428
9.3.7 Boundary effects and MonarchEV . . . 429
9.4 Synergies and trade-offs . . . 430
9.4.1 Transparency effects . . . 430
9.4.2 The price discovery lag . . . 431
9.4.3 The JELLY incident: attack chain and mechanism analysis . . . 432
9.4.4 Risk transmission and counterfactual analysis . . . 433
9.4.5 Legal and governance implications of the emergency intervention . 434
9.4.6 The boundary of protection . . . 434
9.4.7 Validator economics and centralization risk . . . 435
9.4.8 The core trade-offs . . . 436
9.5 Microstructure effects and the liquidity flywheel . . . 437
9.5.1 The restructuring of market-making costs and spread compression 437

9.5.2 The HLP Vault and the democratization of liquidity supply 438
9.5.3 The liquidity flywheel and ecosystem sustainability 439
9.5.4 Empirical data and market performance 440
9.6 Chapter summary 443
References 444

Part IV. Pricing Mechanisms **447**

A Game-Theoretic Analysis of Funding Rates **448**
10.1 Mechanisms and definitions 449
10.1.1 The three-price system 449
10.1.2 Deconstructing the funding rate formula 451
10.1.3 Cross-exchange parameter differences 453
10.2 The three-layer decomposition model 457
10.2.1 The financing cost layer 457
10.2.2 The sentiment premium layer 459
10.2.3 The institutional friction layer 459
10.2.4 Macro interest rate resonance and the economics of arbitrage . . 460
10.2.5 Empirical methods for separating the three layers 462
10.3 The anchoring strength spectrum 464
10.3.1 Defining anchoring strength 464
10.3.2 Arbitrage activity and anchoring strength 466
10.3.3 Conditions that impede arbitrage 466
10.3.4 The AHJ framework: applications and limitations 468
10.3.5 The anchoring paradox 468
10.4 The three-layer game model 470
10.4.1 The anchoring game 470
10.4.2 The convergence game 473
10.4.3 The strategic game 474
10.4.4 State dependence and the decay of anchoring efficacy 475
10.4.5 The structural boundaries of arbitrageurs 477
10.5 Rate polarization and positive feedback 479
10.5.1 Trigger conditions for positive feedback 479
10.5.2 A behavioral finance explanation 481
10.5.3 Termination paths of polarization 482
10.5.4 Rate polarization and the leverage cycle 483
10.5.5 The reconfiguration of participant behavior 484
10.6 The spectrum of strategic behavior 485
10.6.1 A three-tier classification 485
10.6.2 Equilibrium characteristics under normal markets 487
10.6.3 The ALPACA funding rate squeeze case 488
10.7 The settlement frequency game 491
10.7.1 From passive reaction to active management 492
10.7.2 The game-theoretic implications of different settlement frequencies 492
10.7.3 The PIPPIN case analysis 494

10.7.4 Microstructure effects of dynamic settlement . . . 495
10.7.5 The costs of high-frequency settlement . . . 496
10.8 Funding rate as a state variable . . . 497
10.8.1 The information content of the rate level . . . 497
10.8.2 The information content of rate volatility . . . 499
10.8.3 The cross-asset funding rate surface . . . 501
10.8.4 Comparison with traditional sentiment indicators . . . 504
10.8.5 Empirical evidence and limitations of predictive power . . . 507
10.8.6 Rate financialization: Ethena and the funding rate swap market . 509
10.9 Chapter summary . . . 510
References . . . 512

The Endogenous Reflexivity of Leverage and Liquidation 516
11.1 The transmission chain of leverage and liquidation . . . 517
11.1.1 Margin and the leverage constraint . . . 517
11.1.2 The mark-price mechanism . . . 522
11.1.3 The liquidation engine . . . 524
11.1.4 Margin modes . . . 528
11.1.5 Differences in cascades across market architectures . . . 531
11.2 The theoretical foundations of reflexivity . . . 533
11.2.1 Soros's theory of reflexivity . . . 533
11.2.2 The liquidity spiral model . . . 535
11.2.3 The applicability to perpetual futures . . . 536
11.3 The liquidation reflexivity equation . . . 540
11.3.1 Liquidation order flow . . . 540
11.3.2 The three-step mapping . . . 540
11.3.3 The divergence condition . . . 541
11.3.4 Three sources of nonlinearity . . . 544
11.3.5 The cascade risk index . . . 546
11.4 The coupling of the funding rate and leverage . . . 549
11.4.1 The margin-erosion effect . . . 550
11.4.2 The failure of basis arbitrage . . . 552
11.4.3 The compound rate-liquidation positive feedback . . . 555
11.5 The four phases of a liquidation cascade . . . 557
11.5.1 The trigger phase . . . 559
11.5.2 The propagation phase . . . 560
11.5.3 The amplification phase . . . 562
11.5.4 The termination phase . . . 564
11.6 The impossible trinity of liquidation design . . . 566
11.6.1 The three design objectives . . . 567
11.6.2 The conflict among objectives . . . 567
11.6.3 The four liquidation modes . . . 569
11.6.4 Strategic liquidation . . . 570
11.6.5 The nested double trinity . . . 572
11.7 The time-compression effect . . . 574
11.7.1 A cross-market comparison . . . 574

11.7.2 Four driving forces . . . 576
11.7.3 The mismatch between decision-making and execution . . . 577
11.7.4 Implications for risk control and regulation . . . 578
11.7.5 The structural protections of traditional margin markets . . . 579
11.8 The information paradox and mechanism-induced volatility . . . 581
11.8.1 The information paradox of liquidation . . . 581
11.8.2 The adverse-selection dilemma of market makers . . . 582
11.8.3 Mechanism-induced volatility . . . 583
11.9 A reflexivity dissection of the "10.10 event" . . . 584
11.9.1 The system's pre-crisis fragility . . . 584
11.9.2 The trigger and the cascade . . . 586
11.9.3 Cross-asset contagion and collateral collapse . . . 588
11.9.4 Empirical validation of the four-phase model . . . 591
11.10 Chapter summary . . . 593
References . . . 595

The Political Economy of Insurance Funds 598
12.1 The inevitability of shortfalls . . . 599
12.1.1 Four frictions in the liquidation process . . . 599
12.1.2 Ex ante pooling and ex post allocation . . . 604
12.1.3 Internalizing negative externalities and supplying a public good . 607
12.2 Auto-deleveraging: the final mechanism of risk socialization . . . 610
12.2.1 Ranking algorithms . . . 611
12.2.2 The trilemma . . . 614
12.2.3 Formal analysis and empirical testing . . . 617
12.2.4 ADL as a contrarian indicator . . . 620
12.3 The confidence-as-liquidity hypothesis . . . 621
12.3.1 The transmission mechanism . . . 621
12.3.2 The positive feedback loop . . . 623
12.3.3 Empirical evidence of the confidence multiplier . . . 624
12.3.4 Confidence management as risk management . . . 627
12.3.5 The cliff effect . . . 628
12.4 The governance dilemma and the governance spectrum . . . 631
12.4.1 The limits of rule rigidity . . . 632
12.4.2 The guarantor of last resort and discretion . . . 632
12.4.3 Oracle risk . . . 634
12.4.4 The theory of incomplete contracts . . . 636
12.4.5 Information asymmetry and signaling games . . . 637
12.4.6 The governance spectrum G0–G3 . . . 637
12.5 An evaluation framework and cross-platform comparison . . . 641
12.5.1 Constructing the four-dimensional evaluation system . . . 641
12.5.2 Platform comparison analysis . . . 644
12.6 Case study: the JELLY incident . . . 648
12.6.1 The HLP architecture . . . 648
12.6.2 Timeline of the incident . . . 650
12.6.3 The boundaries of decentralized governance . . . 655

12.7 Chapter summary . . . 657
References . . . 659

Part V. Information and Efficiency 662

The Process of Price Discovery and Information Incorporation 663
13.1 The theoretical foundations of price discovery . . . 664
13.1.1 The three-dimensional information space of crypto markets . . . 664
13.1.2 The Kyle model . . . 668
13.1.3 Adverse selection and the bid-ask spread . . . 670
13.1.4 Order flow information metrics . . . 672
13.1.5 The distinctive features of crypto markets . . . 674
13.2 The three-stage price discovery model . . . 675
13.2.1 The expression stage . . . 675
13.2.2 The aggregation stage . . . 676
13.2.3 The transmission stage . . . 677
13.2.4 Total efficiency . . . 678
13.3 The venue gravity model . . . 681
13.3.1 Informed traders' venue preferences . . . 682
13.3.2 The four dimensions of gravity . . . 682
13.3.3 Trading volume and the gravity advantage . . . 684
13.3.4 Funding rates and basis . . . 686
13.3.5 The dynamics of the gravity model . . . 688
13.3.6 The gravity divergence between CEX and DEX perpetual futures 690
13.4 The information flow topology . . . 692
13.4.1 Constructing the topology . . . 692
13.4.2 Measurement tools . . . 693
13.4.3 The empirical landscape . . . 694
13.4.4 Three types of leaders . . . 696
13.4.5 Arbitrage friction and transmission delay . . . 697
13.4.6 The state dependence of the topology . . . 699
13.5 The three pathways of mechanical noise . . . 701
13.5.1 Liquidation-driven price displacement . . . 701
13.5.2 The distortionary effects of funding rate extremes . . . 705
13.5.3 The EMA delay of the mark price . . . 707
13.5.4 The superposition effects of the three pathways . . . 708
13.5.5 Cross-market transmission of noise . . . 711
13.6 The ceiling-floor model . . . 713
13.6.1 The ceiling . . . 713
13.6.2 The floor . . . 714
13.6.3 The state dependence of efficiency . . . 714
13.6.4 How to raise the ceiling and lower the floor . . . 716
13.6.5 A comparison of three markets . . . 717
13.7 Obstacles to price discovery . . . 720
13.7.1 Liquidation cascades and price overshoot . . . 721

13.7.2 Wash trading and spoofing . . . 726
13.7.3 Behavioral biases and narrative-driven trading . . . 728
13.7.4 The superposition of four types of obstacles . . . 730
13.8 Price discovery on on-chain order book DEXs . . . 732
13.8.1 The changing price discovery environment . . . 732
13.8.2 The dual effects of transparency: information symmetrization and the MEV problem . . . 735
13.8.3 The three native information channels of crypto markets . . . 736
13.8.4 A comparison of two architectures . . . 737
13.8.5 Sequencers and MEV protection . . . 741
13.9 Chapter summary . . . 743
References . . . 745

The Governance of Price Benchmarks — Indices, Mark Prices, and Oracles 749
14.1 The price lineage and dual specifications . . . 750
14.1.1 The transaction-price layer . . . 751
14.1.2 The benchmark-price layer . . . 752
14.1.3 The on-chain price layer . . . 752
14.1.4 The information-refining chain . . . 753
14.1.5 Computational and institutional specifications . . . 754
14.2 Price benchmarks as implicit governance . . . 754
14.2.1 Three engineered prices . . . 755
14.2.2 The four dimensions of power in price computation . . . 756
14.2.3 The historical lessons of LIBOR . . . 757
14.2.4 Conflicts of interest in crypto markets . . . 758
14.2.5 The inevitability of a multi-price system . . . 759
14.3 The five governance links of index construction . . . 760
14.3.1 Overview of the five governance links . . . 760
14.3.2 Source admission . . . 762
14.3.3 Weighting schemes . . . 763
14.3.4 Outlier handling . . . 764
14.3.5 Update smoothing . . . 765
14.3.6 Stress degradation . . . 765
14.3.7 Microstructure consequences . . . 766
14.4 Mark price and liquidation triggering . . . 768
14.4.1 The institutional status of the mark price . . . 768
14.4.2 The mark price trilemma . . . 769
14.4.3 The latency effect of EMA smoothing . . . 770
14.4.4 The mark price trap . . . 772
14.4.5 Liquidation hunting and institutional isolation . . . 773
14.4.6 Platform comparison . . . 774
14.4.7 The intensity of liquidation cascades . . . 776
14.5 Oracle architecture and the information bottleneck . . . 778
14.5.1 The oracle trilemma . . . 778
14.5.2 Push-based architecture . . . 779

14.5.3 Pull-based architecture . . . 779
14.5.4 App-chain native oracles . . . 781
14.5.5 The cross-layer information bottleneck . . . 782
14.5.6 Latency and information asymmetry . . . 783
14.5.7 Oracle extractable value . . . 784
14.5.8 The cost-benefit of oracle manipulation . . . 786
14.6 Feedback loops and systemic transmission . . . 787
14.6.1 Positive feedback and the liquidation spiral . . . 787
14.6.2 Negative feedback and the funding rate . . . 788
14.6.3 Oracle synchronization risk . . . 790
14.6.4 The four transmission paths . . . 791
14.7 Attack surfaces, failure modes, and lines of defense . . . 793
14.7.1 A unified attack-surface analysis . . . 793
14.7.2 Five failure modes . . . 794
14.7.3 Three lines of defense . . . 795
14.7.4 The attack-defense matrix . . . 796
14.7.5 Composite attacks and a case retrospective . . . 797
14.8 Case retrospective: the May 19, 2021 event . . . 799
14.8.1 Background and timeline reconstruction . . . 800
14.8.2 Mapping failure points across the layers . . . 801
14.8.3 The simultaneous activation of the four paths . . . 803
14.8.4 Case comparison: systemic crash versus targeted attack . . . 806
14.8.5 Post-event improvements and governance reflection . . . 807
14.9 Chapter summary . . . 808
References . . . 811

Informational Efficiency in Digital Asset Markets **815**
15.1 The analytical toolkit . . . 816
15.1.1 The efficient market hypothesis and crypto markets . . . 816
15.1.2 The adaptive markets hypothesis . . . 820
15.1.3 The efficiency spectrum framework . . . 822
15.2 Testing weak-form efficiency . . . 826
15.2.1 Testing methodology . . . 826
15.2.2 The efficiency evolution of major assets . . . 829
15.2.3 Perpetual-futures-native variables . . . 832
15.2.4 The discovery and disappearance of calendar effects . . . 834
15.2.5 The convergence of cross-exchange spreads . . . 836
15.2.6 Overall assessment . . . 840
15.3 Testing semi-strong-form efficiency . . . 841
15.3.1 The event study method . . . 842
15.3.2 The stratification of information processing costs . . . 843
15.3.3 A quantitative analysis of information processing costs . . . 848
15.3.4 A retrospective on the Bybit attack . . . 851
15.4 Institutional efficiency . . . 855
15.4.1 Mean reversion in the funding rate . . . 856
15.4.2 Efficiency fluctuations within the settlement cycle . . . 859

15.4.3 Mark price and oracle latency . . . 861
15.4.4 Price discovery leadership . . . 864
15.5 The FTX collapse: a case study . . . 868
15.5.1 Reconstructing the event timeline . . . 869
15.5.2 High efficiency in processing on-chain information . . . 873
15.5.3 The market-cap illusion of FTT . . . 875
15.5.4 Contagion and overreaction . . . 879
15.5.5 The liquidation cascade . . . 882
15.5.6 Summary of the FTX event . . . 885
15.6 The evolution of efficiency . . . 888
15.6.1 The frontier era (2009–2017) . . . 891
15.6.2 The institutionalization era (2018–2023) . . . 892
15.6.3 The regulated era (2024–) . . . 894
15.6.4 Future drivers . . . 896
15.7 Chapter summary . . . 897
References . . . 901

Part VI. Arbitrage and Anomalies 905

Arbitrage and the Transmission Mechanisms of Market Efficiency 906
16.1 The theoretical foundations of arbitrage . . . 907
16.1.1 The persistence paradox of the law of one price . . . 907
16.1.2 From *point convergence* to *flow anchoring* . . . 909
16.1.3 The arbitrage infrastructure hypothesis . . . 911
16.1.4 The Grossman-Stiglitz paradox . . . 914
16.2 The structure of the arbitrage ecosystem . . . 916
16.2.1 A five-tier map of participants . . . 916
16.2.2 Ethena and arbitrage-as-a-service . . . 920
16.2.3 The participant interaction network . . . 921
16.2.4 The protocolization of arbitrage . . . 923
16.3 A taxonomy of arbitrage . . . 924
16.3.1 The classification framework . . . 924
16.3.2 An overview of the six arbitrage types . . . 925
16.3.3 The complexity gradient . . . 927
16.4 Spot-perpetual basis arbitrage . . . 929
16.4.1 The bidirectional operating mechanism . . . 929
16.4.2 Decomposing the returns . . . 930
16.4.3 The Ethena effect . . . 932
16.4.4 Empirical evidence for price anchoring . . . 934
16.5 Cross-exchange arbitrage . . . 936
16.5.1 The life cycle of a price spread . . . 936
16.5.2 The evolution of the latency race . . . 937
16.5.3 The Makarov-Schoar effect . . . 939
16.5.4 The long-run compression of spreads . . . 940
16.6 Triangular arbitrage . . . 942

16.6.1 Theoretical foundations . . . 942
16.6.2 Construction and execution logic . . . 944
16.6.3 Triangular arbitrage dynamics in the post-ETF era . . . 947
16.6.4 Term-structure integration . . . 949
16.7 Funding rate arbitrage . . . 950
16.7.1 The tradability of the funding rate . . . 951
16.7.2 The funding rate surface framework . . . 951
16.7.3 The Hyperliquid case . . . 953
16.7.4 Financing-cost calibration . . . 956
16.8 Cross-asset statistical arbitrage . . . 958
16.8.1 Pairs trading in perpetual futures . . . 958
16.8.2 Differences between perpetual and spot statistical arbitrage . . . 961
16.8.3 Maintaining pricing consistency . . . 962
16.9 CEX-DEX cross-tier arbitrage . . . 964
16.9.1 The sources of the efficiency gap . . . 964
16.9.2 Oracle latency and OEV . . . 966
16.9.3 The efficiency function and the LP-loss problem . . . 968
16.9.4 The evolution of DEX anti-arbitrage design . . . 969
16.10 Arbitrage and market quality . . . 971
16.10.1 Spread compression . . . 972
16.10.2 Liquidity integration . . . 973
16.10.3 Price-discovery leadership . . . 975
16.10.4 The rate-stabilization function . . . 976
16.11 The arbitrage-efficiency transmission model . . . 977
16.12 Chapter summary . . . 984
References . . . 984

The Risks of Arbitrage — Why Risk-Free Arbitrage Is Not Risk-Free 987
17.1 The theoretical framework of arbitrage risk . . . 988
17.1.1 The theory of the limits of arbitrage . . . 989
17.1.2 The distinctive risk profile of perpetual futures . . . 990
17.1.3 The arbitrage constraint topology . . . 991
17.1.4 The temporal structure of risk . . . 994
17.2 Margin and forced liquidation . . . 996
17.2.1 How margin constrains the scale of arbitrage . . . 997
17.2.2 The cascading-liquidation mechanism . . . 998
17.2.3 The temporal constraint of no expiry date . . . 1000
17.2.4 Cross-exchange margin fragmentation . . . 1002
17.2.5 A quantitative framework for the margin constraint . . . 1004
17.3 Funding rate reversal . . . 1005
17.3.1 The negative funding rate trap . . . 1006
17.3.2 The tail risk of funding-rate polarization . . . 1007
17.3.3 Ethena and sustained negative funding . . . 1009
17.3.4 The cross-amplification of funding and margin . . . 1010
17.4 Counterparty and platform risk . . . 1011
17.4.1 Lessons from the FTX collapse . . . 1011

17.4.2 Withdrawal restrictions and frozen funds . . . 1013
17.4.3 Insurance-fund depletion and auto-deleveraging . . . 1014
17.4.4 The broken-leg scenario . . . 1016
17.5 Execution and timing risk . . . 1017
17.5.1 Slippage, latency, and partial fills . . . 1018
17.5.2 The coordination problem of triangular arbitrage . . . 1019
17.5.3 CEX-DEX cross-chain execution risk . . . 1021
17.5.4 High-frequency competition and technical risk . . . 1023
17.6 Liquidity depletion . . . 1024
17.6.1 The endogeneity of liquidity . . . 1025
17.6.2 The stampede mechanism . . . 1026
17.6.3 Liquidity stratification . . . 1029
17.6.4 Liquidity evaporation in a crisis . . . 1030
17.7 Smart contract and oracle risk . . . 1033
17.7.1 Smart-contract vulnerabilities . . . 1033
17.7.2 Oracle manipulation . . . 1035
17.7.3 Cross-chain bridge security risk . . . 1036
17.7.4 Protocol-upgrade and governance risk . . . 1037
17.7.5 Sequencer and validator risk . . . 1038
17.8 Regulatory and compliance risk . . . 1039
17.8.1 Cross-jurisdictional compliance costs . . . 1039
17.8.2 Access restrictions and route blockage . . . 1040
17.8.3 The uncertainty of tax treatment . . . 1040
17.8.4 The shock of sudden regulatory events . . . 1041
17.9 The risk resonance model . . . 1041
17.9.1 A resonance analysis of the May 2022 crisis . . . 1042
17.9.2 The risk transmission matrix . . . 1045
17.9.3 The nonlinear amplification effect . . . 1047
17.9.4 The phase-transition critical point . . . 1049
17.10 A risk anatomy of the Ethena protocol . . . 1051
17.11 Chapter summary . . . 1055
References . . . 1056

Arbitrage Failure and Market Anomalies 1059
18.1 An analytical framework for anomalies . . . 1060
18.1.1 The definition and identification of anomalies . . . 1061
18.1.2 The constraint–anomaly mapping . . . 1063
18.1.3 The anomaly persistence spectrum . . . 1065
18.1.4 An overview of the six categories of anomalies . . . 1067
18.2 Basis anomalies . . . 1069
18.2.1 The persistence of positive basis in bull markets . . . 1070
18.2.2 The stickiness of negative basis after panics . . . 1071
18.2.3 The asymmetric mean reversion of the basis . . . 1073
18.2.4 The Ethena effect and protocolized arbitrage . . . 1075
18.3 Funding-rate anomalies . . . 1076
18.3.1 Long-term positive bias and structural risk compensation . . . 1077

18.3.2 The clustering effect of extreme values . . . 1080
18.3.3 The settlement-cycle effect . . . 1082
18.3.4 The persistent splitting of cross-exchange rates . . . 1084
18.4 Cross-exchange anomalies . . . 1086
18.4.1 The residual spread among leading exchanges . . . 1087
18.4.2 The core-periphery pricing gradient . . . 1088
18.4.3 Regulatory barriers and segmented pricing zones . . . 1090
18.5 Cross-asset anomalies . . . 1092
18.5.1 The "inefficiency zone" of altcoin perpetual futures . . . 1092
18.5.2 The long-cycle drift in BTC-ETH relative pricing . . . 1094
18.6 Term-structure anomalies . . . 1096
18.6.1 The implied-rate deviation between perpetual funding and dated-futures basis . . . 1096
18.6.2 CME/ETF and offshore-perpetual dual-track pricing . . . 1098
18.7 CEX–DEX anomalies . . . 1100
18.7.1 Oracle latency and persistent spreads . . . 1101
18.7.2 The systematic deviation of DEX funding rates . . . 1103
18.7.3 The structural arbitrage residual . . . 1105
18.8 The crisis dynamics of anomalies . . . 1107
18.8.1 The nonlinear amplification of anomalies . . . 1108
18.8.2 The contagion path among anomalies . . . 1109
18.8.3 Case: the LUNA/UST crisis of May 2022 . . . 1111
18.8.4 The asymmetry of eruption and repair . . . 1113
18.9 The evolution of anomalies and implications for market design . . . 1115
18.9.1 Fading anomalies . . . 1117
18.9.2 Persistent anomalies . . . 1118
18.9.3 Emerging anomalies . . . 1119
18.9.4 From anomalies to institutional improvement . . . 1121
18.10 Chapter summary . . . 1125
References . . . 1127

Part VII. Liquidity **1129**

Liquidity Provision and Strategic Games **1130**
19.1 The economics of liquidity provision . . . 1131
19.1.1 The nature of market making . . . 1132
19.1.2 Adverse selection and inventory management . . . 1133
19.1.3 Institutional cost . . . 1134
19.1.4 The profit equation and survival conditions . . . 1137
19.2 Adverse selection . . . 1139
19.2.1 The applicability of the Glosten-Milgrom model . . . 1140
19.2.2 Sources of informed trading in perpetual futures . . . 1142
19.2.3 Amplification mechanisms of adverse selection . . . 1144
19.2.4 Toxicity quantification and risk monitoring . . . 1146
19.3 Inventory risk . . . 1148

19.3.1 Adapting classical inventory models to perpetual futures 1149
19.3.2 The inventory dilemma of no expiration 1151
19.3.3 Funding rates and inventory holding costs 1153
19.3.4 The nonlinear relationship between volatility and inventory cost . 1157
19.4 The toxic-flow spectrum 1159
19.4.1 Informed-trading toxicity 1161
19.4.2 Arbitrage-extraction toxicity 1162
19.4.3 Liquidation-cascade toxicity 1163
19.4.4 MEV-attack toxicity 1166
19.4.5 Toxicity superposition and synergistic dynamics 1167
19.5 Market making on CEX CLOBs 1169
19.5.1 The market-maker ecosystem 1169
19.5.2 The quoting strategy framework 1171
19.5.3 Cross-exchange market making and position management 1173
19.5.4 The symbiotic relationship between market makers and exchanges 1175
19.5.5 How perpetual-futures market making differs from traditional market making 1178
19.6 Liquidity provision on DEXs 1181
19.6.1 The on-chain CLOB paradigm 1181
19.6.2 The market-making vault paradigm 1182
19.6.3 Comparing the two paradigms 1183
19.7 Competitive landscape and evolution 1184
19.7.1 The three-dimensional competition in technology, capital, and information 1184
19.7.2 Concentration and systemic risk 1186
19.7.3 Incentive design and market-maker behavior 1188
19.7.4 Future directions of evolution 1189
19.8 Chapter summary 1191
References 1192

The Endogeneity and Fragility of Liquidity **1194**
20.1 Theoretical foundations of liquidity endogeneity 1195
20.1.1 Exogenous versus endogenous liquidity 1195
20.1.2 The reflexivity triangle of liquidity 1196
20.1.3 The four liquidity amplifiers of perpetual futures 1198
20.1.4 The wedge between apparent and true liquidity 1200
20.2 The volatility-liquidity spiral 1200
20.2.1 The market maker's volatility-response function 1200
20.2.2 The transmission mechanism of spread widening 1202
20.2.3 Trigger conditions and the critical point of the spiral 1203
20.2.4 Cross-asset spillover effects 1205
20.2.5 A mathematical characterization of the spiral 1205
20.3 The leverage-liquidation-liquidity cascade 1206
20.3.1 The endogeneity of leverage 1206
20.3.2 Inelastic demand for liquidity 1208
20.3.3 The dual-shock mechanism 1208

20.3.4 The nonlinear cascade and state transitions 1210
20.4 Funding-rate distortion and position squeezes 1213
20.4.1 The distortion of the rate signal 1213
20.4.2 Position squeezes 1214
20.4.3 The stabilization limit of arbitrageurs 1215
20.4.4 Loop coupling and compound crises 1217
20.5 The split between apparent and true liquidity 1219
20.5.1 Ghost liquidity 1219
20.5.2 Hidden liquidity 1221
20.5.3 The synchronized retreat of market makers 1222
20.5.4 The failure of liquidity metrics 1223
20.6 A five-stage framework of liquidity evaporation 1224
20.6.1 Stage one: the microstructural trigger 1226
20.6.2 Stage two: the initial ebb of liquidity 1226
20.6.3 Stage three: the intervention of the liquidation cascade 1227
20.6.4 Stage four: the liquidity vacuum 1228
20.6.5 Stage five: contagion spillover 1229
20.7 The cross-dimensional contagion of liquidity 1229
20.7.1 Cross-asset contagion 1230
20.7.2 Cross-exchange contagion 1233
20.7.3 Cross-layer contagion 1234
20.7.4 Cross-product contagion 1235
20.8 Institutional roots and quantification 1235
20.9 Chapter summary 1239
References 1241

The Public-Goods Dilemma and Incentive Design of Liquidity Governance 1244
21.1 The core problem of liquidity governance 1245
21.1.1 The tragedy of the commons 1245
21.1.2 Two institutional approaches 1246
21.1.3 A four-dimensional evaluation framework 1247
21.2 The principal-agent model of CEX liquidity procurement 1250
21.2.1 Information asymmetry 1250
21.2.2 The structure of the agreement 1252
21.2.3 Adverse selection 1254
21.2.4 Moral hazard 1256
21.2.5 Proprietary market making 1257
21.2.6 Structural deficiencies 1258
21.3 The three-layer architecture of DEX liquidity engineering 1259
21.3.1 The paradigm shift 1259
21.3.2 The capital layer 1260
21.3.3 The strategy layer 1260
21.3.4 The distribution layer 1261
21.3.5 Overall operation 1262
21.3.6 Two routes 1263

21.4 Case study: the HLP vault . . . 1264
21.4.1 Index-style investing . . . 1264
21.4.2 Sources of return . . . 1265
21.4.3 Risk exposure . . . 1267
21.4.4 Historical performance . . . 1269
21.4.5 Comparison with traditional products . . . 1272
21.4.6 Governance dilemmas . . . 1273
21.4.7 Division of labor among vaults . . . 1275
21.5 A survival simulation for the independent market maker . . . 1277
21.5.1 Simulation setup . . . 1277
21.5.2 The revenue side . . . 1278
21.5.3 The cost side . . . 1279
21.5.4 Scenario analysis . . . 1280
21.5.5 The profitability boundary . . . 1281
21.5.6 Active versus passive . . . 1282
21.5.7 Judging effectiveness . . . 1283
21.6 Comparison and coexistence of the two paradigms . . . 1284
21.6.1 Systematic comparison . . . 1284
21.6.2 The democratization of liquidity . . . 1286
21.6.3 The pattern of coexistence . . . 1286
21.6.4 Future trends . . . 1287
21.7 Chapter summary . . . 1288
References . . . 1289

Part VIII. Volatility **1294**

A Four-Source Decomposition of Volatility and Excess Volatility **1295**
22.1 Theoretical foundations of volatility decomposition . . . 1296
22.1.1 The necessity of volatility decomposition . . . 1296
22.1.2 A review of traditional decomposition methods . . . 1298
22.1.3 The four-source decomposition framework . . . 1299
22.1.4 The concept of a volatility budget . . . 1302
22.2 Informational volatility . . . 1304
22.2.1 Macro information shocks . . . 1304
22.2.2 On-chain information and idiosyncratic events . . . 1306
22.2.3 Informed trading and volatility decomposition . . . 1306
22.2.4 Leverage amplification and round-the-clock continuity . . . 1307
22.2.5 Measuring informational volatility . . . 1309
22.3 Liquidity volatility . . . 1310
22.3.1 Illiquidity and spurious volatility . . . 1310
22.3.2 Market-maker retreat and the volatility positive-feedback loop . . . 1312
22.3.3 Time structure . . . 1313
22.3.4 The asset dimension gradient . . . 1314
22.3.5 Measurement methods . . . 1316
22.4 Mechanism-induced volatility . . . 1317

22.4.1 Liquidation cascades . . . 1317
22.4.2 Funding-rate settlement pulses . . . 1319
22.4.3 The lag effect of the mark-price mechanism . . . 1320
22.4.4 Attribution of volatility differences . . . 1321
22.4.5 Measuring mechanism-induced volatility . . . 1322
22.5 Transmitted volatility . . . 1323
22.5.1 Spillover effects within the crypto market . . . 1324
22.5.2 BTC dominance and beta volatility . . . 1325
22.5.3 Arbitrage transmission between spot and perpetual . . . 1326
22.5.4 Macro transmission from traditional finance . . . 1326
22.5.5 Measuring transmitted volatility . . . 1328
22.6 Four-source interaction and regime switching . . . 1329
22.6.1 The nonlinear interaction of the four sources . . . 1329
22.6.2 The state dependence of the volatility budget . . . 1331
22.6.3 Triggers and early warning of regime switching . . . 1333
22.6.4 The liquidity-volatility spiral in the crisis state . . . 1333
22.6.5 Practical implications . . . 1334
22.7 The excess volatility of perpetual futures . . . 1335
22.7.1 Definition and measurement . . . 1335
22.7.2 Source attribution . . . 1336
22.7.3 The heterogeneity of excess volatility . . . 1338
22.7.4 Welfare implications . . . 1339
22.8 Chapter summary . . . 1341
References . . . 1342

The Microstructural Mechanisms of Volatility Clustering 1347
23.1 The empirical facts of volatility clustering . . . 1348
23.1.1 Definition and measurement . . . 1348
23.1.2 Cross-market quantitative comparison . . . 1350
23.1.3 Nonlinear clustering features . . . 1351
23.2 Slow variables and the fragility baseline . . . 1352
23.2.1 The macroeconomic environment . . . 1352
23.2.2 The level of systemic leverage . . . 1353
23.2.3 Liquidity depth . . . 1354
23.2.4 The pre-programming of fragility . . . 1355
23.3 Positive feedback loops and shock amplification . . . 1357
23.3.1 The intertemporal transmission of shocks . . . 1357
23.3.2 The liquidity loop . . . 1357
23.3.3 The leverage loop . . . 1359
23.3.4 The behavioral loop . . . 1360
23.3.5 Activation and termination conditions . . . 1361
23.4 The state machine and volatility regime transitions . . . 1361
23.4.1 The model framework . . . 1362
23.4.2 The transition matrix . . . 1363
23.4.3 Transition asymmetry . . . 1365
23.4.4 The regulation by slow variables . . . 1367

23.5 The institutional amplifiers of perpetual futures . . . 1368
23.5.1 Funding rates . . . 1368
23.5.2 Automated liquidation . . . 1369
23.5.3 Around-the-clock trading . . . 1371
23.5.4 Synergistic effects . . . 1372
23.6 The accumulation of fragility during low-volatility periods . . . 1373
23.6.1 The endogenous accumulation of leverage . . . 1374
23.6.2 Liquidity concentration . . . 1376
23.6.3 Empirical relationships . . . 1377
23.6.4 Early-warning signals . . . 1378
23.7 Case study: the life cycle of the May 2021 crash . . . 1381
23.7.1 The state of the slow variables . . . 1381
23.7.2 The trigger and initial shock . . . 1382
23.7.3 The cascade and crisis . . . 1383
23.7.4 Theoretical validation . . . 1385
23.8 Chapter summary . . . 1388
References . . . 1390

Volatility Forecasting and Decision-Making 1394
24.1 The set of alternative volatility signals for perpetual futures . . . 1395
24.1.1 The absence of implied volatility . . . 1395
24.1.2 The funding rate as a volatility signal . . . 1396
24.1.3 Liquidation heatmaps . . . 1398
24.1.4 Changes in order book shape . . . 1399
24.1.5 Cross-exchange and cross-asset spillover signals . . . 1399
24.1.6 On-chain signals . . . 1400
24.2 Volatility forecasting models: adapting classical methods . . . 1400
24.2.1 Realized volatility . . . 1401
24.2.2 GARCH models . . . 1401
24.2.3 HAR models . . . 1402
24.2.4 Machine-learning methods . . . 1403
24.2.5 Model comparison and combination . . . 1404
24.3 A source-decomposed volatility forecasting framework . . . 1405
24.3.1 Forecasting informational volatility . . . 1407
24.3.2 Forecasting liquidity volatility . . . 1408
24.3.3 Forecasting mechanical volatility . . . 1409
24.3.4 Forecasting transmission volatility . . . 1411
24.3.5 Aggregating the four-source forecast . . . 1412
24.4 Applying volatility in market-making decisions . . . 1416
24.4.1 Volatility and the optimal spread . . . 1416
24.4.2 State identification and strategy switching . . . 1417
24.4.3 Distinguishing liquidation volatility from informational volatility . . . 1418
24.4.4 Risk management for protocol market-making vaults . . . 1420
24.5 Applying volatility in directional trading and portfolio management . . . 1420
24.5.1 Volatility-adjusted position sizing . . . 1421
24.5.2 Volatility timing . . . 1423

24.5.3 Indirect volatility trading in perpetual futures 1424
24.5.4 Portfolio-level tail-risk management 1425
24.5.5 Designing stress-test scenarios 1426
24.6 Applying volatility in systemic risk management 1428
24.6.1 Dynamic margin calibration 1428
24.6.2 Adjusting liquidation parameters 1429
24.6.3 Liquidity early-warning systems 1430
24.6.4 The volatility dashboard 1432
24.7 The limits of volatility forecasting 1433
24.7.1 The decay of predictability 1433
24.7.2 Model-homogeneity risk 1435
24.7.3 Overfitting and out-of-sample failure 1437
24.7.4 Black swan events 1438
24.8 Chapter summary 1439
References 1441

Part IX. Engineering: From Diagnosis to Construction 1446

A Five-Dimensional Evaluation of Perpetual Futures Market Quality 1447
25.1 Theoretical foundations of market quality 1448
25.1.1 The economic meaning of market quality 1448
25.1.2 Traditional theories of market quality 1449
25.1.3 The distinctive challenges of perpetual futures 1451
25.1.4 The architecture of the five-dimensional model 1452
25.2 Transaction costs 1454
25.2.1 Explicit costs 1455
25.2.2 The bid-ask spread and market impact 1455
25.2.3 The holding cost of the funding rate 1456
25.2.4 Measuring total transaction cost 1457
25.2.5 Cross-platform comparison and evolution 1458
25.3 Price discovery 1460
25.3.1 Measurement methods 1460
25.3.2 The leadership relationship between perpetual futures and spot . 1461
25.3.3 Cross-asset differences 1462
25.3.4 Excess volatility 1462
25.4 Resilience 1464
25.4.1 Definition and measurement framework 1464
25.4.2 The spread-recovery half-life 1465
25.4.3 The speed of order-book reconstruction 1466
25.4.4 The capacity to withstand liquidation cascades 1467
25.4.5 The resilience function of the funding rate 1469
25.4.6 Cross-exchange contagion 1470
25.4.7 Cross-platform comparison and evolution 1471
25.5 Fairness 1472
25.5.1 A multilayered structure 1473

25.5.2 Execution fairness . . . 1474
25.5.3 Institutional fairness . . . 1475
25.5.4 The fairness of DEXs . . . 1476
25.6 Accessibility . . . 1477
25.6.1 Definition and measurement framework . . . 1477
25.6.2 The entry barriers of CEXs . . . 1478
25.6.3 The permissionless access and technical barriers of DEXs . . . 1478
25.6.4 The feedback loop between accessibility and market quality . . . 1479
25.6.5 Cross-platform comparison and evolution . . . 1479
25.7 Integrated assessment . . . 1480
25.7.1 Construction methodology . . . 1480
25.7.2 A cross-sectional comparison of CEXs . . . 1481
25.7.3 A comparison of CEXs and DEXs . . . 1482
25.7.4 Dynamic evolution . . . 1483
25.8 Regulatory implications . . . 1483
25.8.1 The correspondence with regulatory objectives . . . 1484
25.8.2 The applicability of existing frameworks . . . 1485
25.8.3 The design of regulatory tools . . . 1486
25.8.4 Regulatory arbitrage and international coordination . . . 1487
25.9 Technological change and institutional evolution . . . 1488
25.9.1 AI market makers . . . 1488
25.9.2 On-chain derivatives . . . 1489
25.9.3 Cross-chain liquidity integration . . . 1490
25.9.4 Long-term equilibrium . . . 1490
25.10 Chapter summary . . . 1492
References . . . 1493

The Economics of Regulation Between Innovation and Stability 1498
26.1 The legitimacy of regulation . . . 1499
26.1.1 Four categories of market failure . . . 1500
26.1.2 The boundary of self-repair . . . 1501
26.1.3 The costs of regulation . . . 1501
26.1.4 The net-benefit analysis framework . . . 1502
26.2 Regulatory targets and prioritization . . . 1504
26.2.1 Systemic risk . . . 1504
26.2.2 Market manipulation . . . 1505
26.2.3 Consumer protection . . . 1506
26.2.4 Market integrity . . . 1507
26.2.5 Prioritization . . . 1508
26.3 The regulatory toolbox . . . 1510
26.3.1 Leverage limits . . . 1510
26.3.2 Disclosure and transparency . . . 1512
26.3.3 Market-maker obligations . . . 1513
26.3.4 Capital requirements and asset segregation . . . 1513
26.3.5 The five-dimensional impact matrix . . . 1514
26.4 The jurisdiction problem . . . 1517

26.4.1 The impossible trinity . . . 1517
26.4.2 The global regulatory map . . . 1519
26.4.3 The economic consequences of fragmentation . . . 1520
26.4.4 The path to coordination . . . 1521
26.5 Functional regulation versus entity-based regulation . . . 1523
26.5.1 The limits of entity-based regulation . . . 1523
26.5.2 The principle of functional regulation . . . 1524
26.5.3 The functional decomposition of perpetual futures . . . 1525
26.5.4 Implementation challenges . . . 1527
26.6 Regulating DeFi perpetual futures . . . 1527
26.6.1 The spectrum of decentralization . . . 1528
26.6.2 The failure of traditional tools . . . 1529
26.6.3 New regulatory tools . . . 1530
26.6.4 Case study: Hyperliquid . . . 1531
26.7 Empirical evidence on regulation . . . 1532
26.7.1 Empirical evidence on leverage limits . . . 1532
26.7.2 The market-quality effects of compliance requirements . . . 1534
26.7.3 The cost of regulatory uncertainty . . . 1534
26.7.4 A comprehensive assessment of five-dimensional market quality . 1535
26.8 The political economy of regulation . . . 1537
26.8.1 The theory of regulatory capture . . . 1537
26.8.2 Stakeholder analysis . . . 1538
26.8.3 Regulatory competition . . . 1539
26.8.4 Factors affecting regulatory quality . . . 1540
26.9 Chapter summary . . . 1540
References . . . 1541

A Thought Experiment on Perpetual Futures Reshaping Global Asset Pricing 1547
27.1 The universality of the perpetual futures mechanism . . . 1548
27.1.1 The funding rate anchoring mechanism . . . 1549
27.1.2 Applicability across asset classes . . . 1550
27.1.3 The economics of synthetic exposure . . . 1551
27.1.4 The concept of an Everything Perpetual . . . 1552
27.1.5 The order-book microstructure of centralized exchanges . . . 1552
27.2 The historical logic of pricing-power migration . . . 1553
27.2.1 Three migrations of pricing power . . . 1553
27.2.2 The drivers of migration . . . 1554
27.2.3 The candidacy of perpetual futures . . . 1555
27.3 Thought experiment I: foreign exchange perpetual futures . . . 1555
27.3.1 Structural inefficiencies of the foreign exchange market . . . 1555
27.3.2 The design of foreign exchange perpetual futures . . . 1556
27.3.3 Potential advantages . . . 1556
27.3.4 Core obstacles . . . 1557
27.3.5 Feasibility assessment . . . 1558
27.4 Thought experiment II: equity index perpetual futures . . . 1558

27.4.1 The success and limits of the ETF . . . 1558
27.4.2 The design of index perpetual futures . . . 1559
27.4.3 Potential advantages . . . 1561
27.4.4 Core obstacles . . . 1561
27.4.5 Near-term value . . . 1562
27.5 Scenario analysis: price transmission in TSLA perpetual futures . . . 1562
27.5.1 The traditional pricing ecosystem of TSLA . . . 1563
27.5.2 The price-transmission model . . . 1564
27.5.3 A five-stage scenario simulation . . . 1565
27.5.4 Sensitivity analysis . . . 1567
27.5.5 Summary of the thresholds . . . 1568
27.5.6 Case simulation: price transmission of an after-hours event . . . 1569
27.6 Thought experiment III: commodities and emerging assets . . . 1571
27.6.1 Crude oil, gold, and agricultural commodities . . . 1571
27.6.2 Carbon credits and real estate indices . . . 1572
27.6.3 Prediction markets . . . 1573
27.6.4 Technical feasibility and institutional obstacles . . . 1574
27.7 The dynamics of pricing-power migration . . . 1576
27.7.1 Stage one: crypto-native assets . . . 1576
27.7.2 Stage two: shadow pricing . . . 1577
27.7.3 Stage three: institutional recognition . . . 1577
27.7.4 Stage four: infrastructure integration . . . 1578
27.7.5 Preconditions and timeline . . . 1579
27.8 The ultimate constraints . . . 1580
27.8.1 The ultimate picture . . . 1580
27.8.2 The settlement needs of the physical world . . . 1580
27.8.3 The financial power of sovereign states . . . 1581
27.8.4 The oracle problem . . . 1582
27.8.5 Human trust structures . . . 1582
27.9 Anchoring in reality . . . 1583
27.9.1 Early practice of the Everything Perpetual . . . 1583
27.9.2 The bridging role of the Bitcoin ETF . . . 1583
27.9.3 The evolutionary roadmap . . . 1584
27.10 Chapter summary . . . 1585
References . . . 1585

From the Monolithic Black Box to Layered Architecture 1589
28.1 A trust anatomy of the CEX monolithic black box . . . 1591
28.1.1 The five trust assumptions in a trade's life cycle . . . 1591
28.1.2 The trust stack model . . . 1593
28.1.3 FTX's extreme failure mode . . . 1595
28.1.4 Trust concentration and systemic risk . . . 1596
28.2 The trust-performance frontier . . . 1598
28.2.1 The performance cost of a fully on-chain architecture . . . 1598
28.2.2 The systemic risk of a fully off-chain architecture . . . 1599
28.2.3 Characterizing the trust-performance frontier . . . 1599

28.2.4 The logic of layered architecture 1601
28.3 The layered trust architecture of traditional finance 1602
28.3.1 The historical evolution of functional separation 1602
28.3.2 The CCP and clearing-layer trust 1603
28.3.3 The DTCC and custody-layer trust 1604
28.3.4 The CFTC/SEC and governance-layer trust 1605
28.3.5 An overall assessment of the traditional trust architecture 1606
28.3.6 Functional inheritance from the CCP to on-chain clearing 1607
28.4 Design principles of layered architecture 1608
28.4.1 De-trusting the asset-custody layer 1608
28.4.2 The verifiability of the matching and execution layer 1609
28.4.3 The clearing and settlement layer: from reckoning and posting to default handling 1610
28.4.4 Manipulation-resistant design of the data and oracle layer 1612
28.4.5 The procedural legitimacy of the governance layer 1613
28.4.6 Modular design of the interfaces between layers 1613
28.5 A trust-stack audit of current architectures 1614
28.5.1 The trust stack of a typical CEX 1614
28.5.2 The trust stack of AMM-type DEXs 1616
28.5.3 The trust stack of app-chain DEXs 1618
28.5.4 The trust stack of hybrid architectures 1621
28.6 Hyperliquid's architectural choices 1623
28.6.1 The app-chain and the on-chain order book 1623
28.6.2 An analysis of the trust state at each layer 1625
28.6.3 The performance-latency trade-off 1626
28.6.4 Hyperliquid's residual trust 1628
28.6.5 The generalizability of the app-chain model 1630
28.7 The engineering judgment of sufficient decentralization 1631
28.7.1 Defining 'sufficient' 1631
28.7.2 The trust needs of different users 1632
28.7.3 A gradation of trust 1633
28.7.4 The path of progressive decentralization 1635
28.8 Modularity, composability, and the trust market 1636
28.8.1 How layered architecture reshapes the five dimensions of market quality 1636
28.8.2 New forms of risk in a layered architecture 1639
28.8.3 Architecture as governance 1640
28.8.4 The trust market and composability 1642
28.9 Chapter summary 1643
References 1643

Designing an Ideal Decentralized Clearing and Settlement Architecture 1649
29.1 The overall architecture of a decentralized clearinghouse 1651
29.1.1 A single deterministic state machine 1651
29.1.2 The clearing stage: novation, multilateral netting, and recomputation 1654

29.1.3 The settlement stage: marking to market, funding, and finality . 1655
29.1.4 Default management: liquidation and the loss waterfall 1657
29.1.5 Concentrated adjudicative authority, open verification 1659
29.1.6 Inter-layer transmission and systemic fragility 1661
29.2 The missing layers of the default waterfall 1663
29.2.1 The sources of default-waterfall depth 1663
29.2.2 The layer gaps in the on-chain waterfall 1665
29.2.3 Liquidity providers as the counterparty: outsourcing the first loss 1667
29.2.4 The procyclical failure of reflexive-token collateral 1669
29.2.5 Cross-protocol mutualization and the common-mode risk of restaking 1671
29.2.6 The current absence of an equally deep waterfall on-chain 1672
29.3 Orderly degradation without a lender of last resort 1675
29.3.1 The absence of a lender of last resort and the shift in design goals 1675
29.3.2 Risk-pool isolation 1677
29.3.3 Active shock management 1679
29.3.4 Other degradation mechanisms 1681
29.3.5 The three costs of orderly degradation 1683
29.3.6 The incentive failure of liquidation execution 1685
29.4 The boundary of market incentives and the constraint of residual discretion1687
29.4.1 Market incentives as a substitute for regulatory compulsion . . . 1687
29.4.2 Procyclicality and predictability attacks 1690
29.4.3 Incomplete contracts and the guarantor of last resort 1692
29.4.4 The three constraints on residual discretion 1694
29.4.5 The JELLY incident and the boundary of the purely algorithmic clearinghouse 1696
29.5 The ideal-architecture checklist and the Hyperliquid gap audit 1698
29.5.1 Four normative principles 1698
29.5.2 The Hyperliquid gap audit 1701
29.5.3 Unmet items and the frontier of the field 1705
29.5.4 The interdependence of the four principles 1706
29.5.5 From risk governance to rule governance 1708
Chapter summary 1708
References 1710

Embedded Compliance and Automated Governance 1714
30.1 The theoretical foundations of embedded compliance 1715
30.1.1 The limits of the traditional compliance model 1715
30.1.2 The core concept of embedded compliance 1717
30.1.3 The three-tier compliance spectrum 1718
30.1.4 From regulatory principles to engineering implementation 1720
30.2 Hard-coded compliance 1720
30.2.1 When hard-coding applies 1720
30.2.2 Hard-coding asset segregation 1721
30.2.3 Hard-coding the leverage cap 1724
30.2.4 The cost of hard-coding 1725

30.3 Parameterized compliance 1726
30.3.1 When parameterization applies 1726
30.3.2 Parameterizing the margin ratio 1727
30.3.3 Parameterizing rate parameters 1730
30.3.4 Parameterizing market-making obligations 1731
30.3.5 The governance mechanism for parameter adjustment 1733
30.4 Reporting-based compliance 1733
30.4.1 When reporting-based compliance applies 1734
30.4.2 Transaction monitoring and anomaly detection 1734
30.4.3 Automating proof of reserves 1736
30.4.4 Real-time risk disclosure 1737
30.4.5 The regulatory interface 1737
30.5 The technical reconciliation of privacy and compliance 1739
30.5.1 The structural conflict between compliance and privacy 1739
30.5.2 Compliance applications of zero-knowledge proofs 1740
30.5.3 Designing compliance credentials 1742
30.5.4 Engineering challenges and progress 1743
30.6 Engineering the governance layer 1744
30.6.1 Immutability versus the need to upgrade 1744
30.6.2 The dual architecture of timelock and emergency powers 1744
30.6.3 The proxy contract pattern and governance constraints 1746
30.6.4 Governance attacks and defenses 1747
30.6.5 The governance roadmap of progressive decentralization 1749
30.6.6 A governance audit of Hyperliquid 1751
30.7 Mapping regulatory requirements to technical implementations 1753
30.7.1 The technical implementation of the six principles 1753
30.7.2 The synergy of regulation and technology 1755
30.7.3 Compliance functions that technology cannot replace 1756
30.8 Chapter summary 1757
References 1758

Part X. The Future **1762**

When AI Becomes Market Infrastructure **1763**
31.1 A four-stage model of AI penetration into microstructure 1765
31.1.1 AI assistance 1766
31.1.2 AI execution 1767
31.1.3 AI autonomy 1768
31.1.4 AI adversarial competition 1769
31.1.5 Assessing the stage of penetration 1770
31.2 The new economics of AI market making 1772
31.2.1 Recalibrating the profit equation 1773
31.2.2 The changing nature of adverse selection 1774
31.2.3 The dimensional explosion of inventory management 1775
31.2.4 Adversarial market making 1776

31.2.5 Structural strengths and weaknesses . . . 1777
31.3 AI and liquidity dynamics . . . 1778
31.3.1 Mitigating the tragedy of the liquidity commons . . . 1779
31.3.2 The fragility of strategy homogenization . . . 1780
31.3.3 Synchronized behavior and heterogeneous decision-making . . . 1782
31.3.4 AI adaptive market making . . . 1783
31.4 AI and volatility structure . . . 1785
31.4.1 Revising the four-source decomposition . . . 1785
31.4.2 AI flash crashes . . . 1787
31.4.3 The AI effect on volatility clustering . . . 1789
31.4.4 The effectiveness of volatility forecasting . . . 1791
31.5 AI and market quality . . . 1792
31.5.1 Transaction cost . . . 1793
31.5.2 Price discovery . . . 1793
31.5.3 Resilience . . . 1794
31.5.4 Fairness . . . 1795
31.5.5 Accessibility . . . 1796
31.6 The governance challenges of AI agents . . . 1798
31.6.1 Regulating non-human participants . . . 1798
31.6.2 Liability attribution and compliance credentials . . . 1799
31.6.3 Native compliance and formal verification . . . 1800
31.6.4 AI and DAO governance . . . 1801
31.7 From market participant to market infrastructure . . . 1802
31.7.1 The profit of participant AI trends toward zero . . . 1803
31.7.2 The value space of infrastructure AI . . . 1804
31.7.3 Comparing operating environments . . . 1805
31.7.4 Value-capture mechanisms . . . 1807
31.8 The six core AI-driven systems . . . 1808
31.8.1 Dynamic risk parameters . . . 1810
31.8.2 Intelligent matching engine . . . 1813
31.8.3 Predictive liquidation . . . 1815
31.8.4 Dynamic pricing . . . 1818
31.8.5 Adaptive compliance monitoring . . . 1820
31.8.6 Liquidity-routing optimization . . . 1821
31.9 Semantic oracles . . . 1823
31.9.1 The capability boundary of traditional oracles . . . 1823
31.9.2 Semantic oracle architecture . . . 1824
31.9.3 Unlocking perpetual futures on everything . . . 1826
31.9.4 The trust problem of semantic verification . . . 1828
31.10 From code is law to model is law . . . 1830
31.10.1 The strengths and limitations of code is law . . . 1831
31.10.2 The potential and risks of model is law . . . 1832
31.10.3 The hybrid architecture . . . 1834
31.10.4 Model auditing . . . 1836
31.11 Chapter summary . . . 1838
References . . . 1839

The Global Vision of Autonomous World Finance **1844**
32.1 The three-layer stack of autonomous finance . . . 1845
32.1.1 The asset layer . . . 1846
32.1.2 The infrastructure layer . . . 1848
32.1.3 The intelligence layer . . . 1849
32.1.4 Synergies . . . 1850
32.2 The structural constraints on financial inclusion . . . 1853
32.2.1 Genuine demand . . . 1853
32.2.2 The inclusive functions of perpetual futures . . . 1854
32.2.3 The inclusive properties of the infrastructure . . . 1855
32.2.4 The democratization of AI financial advisers . . . 1857
32.2.5 Real constraints . . . 1858
32.3 The globalization of capital-allocation efficiency . . . 1860
32.3.1 The current fragmentation of capital allocation . . . 1860
32.3.2 Around-the-clock continuous pricing . . . 1861
32.3.3 Frictionless capital flows . . . 1862
32.3.4 Welfare analysis and distributional effects . . . 1864
32.4 New systemic risks . . . 1865
32.4.1 The globalization of the reflexivity triangle . . . 1865
32.4.2 A global AI flash crash . . . 1867
32.4.3 Oracle single-point-of-failure risk . . . 1868
32.4.4 Too interconnected to fail . . . 1869
32.5 Three irreducible tensions . . . 1871
32.5.1 Efficiency versus stability . . . 1872
32.5.2 Permissionlessness versus consumer protection . . . 1872
32.5.3 Global integration versus sovereign plurality . . . 1873
32.5.4 The epistemology of tension management . . . 1873
32.6 From vision to roadmap . . . 1874
32.6.1 Evolutionary paths . . . 1874
32.6.2 Key uncertainties . . . 1876
32.6.3 Responsible optimism . . . 1877
32.7 Chapter summary . . . 1877
References . . . 1878

Summary and Research Outlook **1881**
33.1 Core propositions . . . 1882
33.2 Theoretical threads running through the book . . . 1885
33.3 The confidence spectrum of the conclusions . . . 1887
33.4 Unresolved problems and the research agenda . . . 1892
33.5 Disciplinary contributions and epistemological legacy . . . 1896
33.6 Recommendations for different readers . . . 1898
References . . . 1900

Acknowledgments

For Yan and Ya — my wife and my daughter — who are the reason any of this is worth building.

Disclaimer

This document is an independent survey and educational resource prepared solely by the author. The views, opinions, and conclusions expressed herein are those of the author alone.

This work does not contain any proprietary, confidential, or trade-secret information. All referenced material is drawn from publicly available sources, including peer-reviewed publications, open-access preprints, official documentation, and open-source repositories.

The content is provided "as is" without warranty of any kind, express or implied. The author makes no representations regarding the accuracy, completeness, or suitability of the information for any particular purpose. Readers should independently verify any claims, formulas, or implementation details before applying them in production systems.

Feedback welcome. If you find any mistakes or inaccuracies, or have suggestions for improvement, you are encouraged to send feedback to permissionlessfi@gmail.com.

AI Disclosure. Large language models were used as a research and drafting aid. All content was edited and verified by the author on a best-effort basis.

License. This work is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). You are free to share (copy and redistribute) and adapt (remix, transform, and build upon) this material for any purpose, including commercially, provided you give appropriate credit, provide a link to the license, and indicate if changes were made. Full license text: https://creativecommons.org/licenses/by/4.0/

Preface

Why This Book Exists

I arrived at this subject twice.

The first time was in traditional finance and trading, where I learned that a market is never just a place where prices meet — it is a trust machine two centuries in the making, built on central counterparties, multi-layered clearing, and regulation, and yet it never resolved one fundamental contradiction: institutionalized trust is itself the greatest single point of failure. From the collapse of Lehman Brothers to the implosion of FTX, every crisis asks the same question — can we build a financial system that does not depend on intermediary credit? The second time was in blockchain, where I watched a generation of engineers try to answer it, rebuilding those same functions — custody, matching, clearing, settlement, price discovery — from first principles, in the open, with code standing in for the institutions the rest of finance takes for granted. The perpetual future sat at the center of that effort: the instrument that, more than any other, became the native language of these new markets, and the clearest proof so far that price discovery itself can be permissionless.

We live in an age when information has never been cheaper. Ask a model, search a forum, read a hundred threads, and fragments of the answer will appear in seconds. Yet if you want a single, systematic account of how permissionless perpetual markets actually work — from the anatomy of a single order to the macro question of whether these markets might one day help price the world — you will not find it. The knowledge exists, but it is scattered across exchange documentation, academic papers written in different dialects, audit reports, and the tacit intuition of practitioners who rarely write things down. Abundance of information is not the same as structure. This book is my attempt to supply the structure: to connect the microstructure of a single trade to the architecture of a global, on-chain financial system, and to do it with the seriousness the subject deserves.

Who This Book Is For

This book is written for the reader who wants to understand *why*, not only *how*:

- **Traders and practitioners** who already use these markets and want a rigorous model of the machinery beneath the interface — funding, liquidation, order books, liquidity, and the reflexive dynamics that turn ordinary days into cascades.

- **Builders and engineers** designing exchanges, protocols, oracles, and risk systems, who need to see how each design choice propagates from the microstructure outward.
- **Researchers and students** of market microstructure, financial economics, and distributed systems, who want a coherent framework and an evidence map rather than a collection of anecdotes.
- **Policymakers and the intellectually curious**, who sense that something structurally new is happening and want an honest account of both its promise and its fragilities.

You do not need to be a quant or a Solidity developer. You do need a willingness to think carefully.

What You Will Gain

By the end, you should be able to trace a single order from intention to execution; explain how funding rates, liquidation, and margin jointly hold a perpetual market to reality — and how, under stress, they fail; reason about price discovery, information efficiency, arbitrage, liquidity, and volatility as one connected system rather than isolated topics; and evaluate, regulate, and ultimately imagine *building* a better version of these markets. The aim is not to give you predictions to trade on, but a set of durable mental models that outlive any particular protocol or cycle. Along the way you will meet several analytical frameworks of my own — among them a four-source decomposition of volatility and a five-dimensional evaluation of market quality — offered not as final answers but as tools you can pick up, test, and sharpen.

How This Book Is Organized

The book has thirty-three chapters grouped into ten parts, and those parts build in five layers.

- **The foundation (Part I, Chapters 1–3)** asks what digital assets and perpetual futures actually are, and why they deserve to exist at all.
- **The ecosystem (Parts II–III, Chapters 4–9)** turns to who trades these instruments, the infrastructure they depend on, and how a single trade becomes a price — culminating in Hyperliquid as a worked case study of an on-chain order book.
- **The theoretical core (Parts IV–VIII, Chapters 10–24)** is the heart of the book: pricing mechanisms (funding, liquidation, margin), information and efficiency, arbitrage and anomalies, liquidity, and volatility — the deep operating logic of the market.
- **The engineering (Part IX, Chapters 25–30)** follows a "diagnose first, then prescribe" arc: measuring market quality and the economics of regulation, then designing trust, decentralized clearing, and automated governance — closing with a thought experiment on what happens if perpetual futures come to price the world.

- **The future (Part X, Chapters 31–33)** looks at AI as market infrastructure, the possibility of autonomous on-chain finance, and where all of this may be heading — ending not with a conclusion but with the next question.

Each part can be read on its own, but the arguments compound: the stop-loss order in Chapter 7 becomes the liquidation cascade of Chapter 11 and the volatility of Chapter 22; the three futures sketched in Chapter 3 are the ones judged in Chapter 33. Read straight through and you will feel the loops close.

The Landscape in 2026

As I write, perpetual futures have become the dominant instrument by volume in digital-asset markets — and the model is spilling beyond crypto, with perpetuals now written on equities, commodities, foreign exchange, and more. Meanwhile, the line between centralized and decentralized venues is visibly dissolving. On-chain order books that seemed like research toys a few years ago now clear meaningful flow; oracles, clearing, and margin logic increasingly live in public code; and regulators across major jurisdictions are moving from denial to design. It is a moment of unusual plasticity — the rules of this market are still being written, which is precisely why understanding its microstructure matters now rather than later. The choices made in the next few years about liquidation engines, oracle governance, and incentive design will harden into the plumbing of a global system. And a deeper shift is already visible at the edges: algorithms are moving from being tools of the market to being its infrastructure — quoting, hedging, clearing, and increasingly enforcing the rules themselves. Code defines those rules, algorithms execute them, and consensus verifies them; the outline of an autonomous financial system is emerging from theory into practice.

Why I Believe in This Future

I will be honest about my bias: I believe in the permissionless future. Not as an ideology and not as a promise that everything will go well — this book spends many of its chapters on exactly how these markets break — but as a direction: a long migration from permission to permissionlessness, from institutional trust to algorithmic trust, traced through the microstructure where that migration is actually decided. A world in which the core functions of finance are open, programmable, globally reachable, and auditable by anyone is, to me, a better world than the one we inherited, and I have chosen to spend my working life helping build it. This book is the most useful thing I know how to contribute to that project: an attempt to understand the thing clearly, so that we can build it well.

About the Author

Eric Cheung writes from a vantage point that spans both sides of this subject. Over a decade — from a PhD into hands-on R&D — he has come to understand how traditional finance actually works and where its problems lie. He is a trader himself, and he brings deep, practical experience in blockchain research and development. That combination — close enough to the markets to know how they behave under stress, and close enough to the technology to know how it is actually built — runs through every chapter. He believes in the future of permissionless finance.

Part I
Foundations

Core question: What exactly are digital assets and perpetual futures, and by what right do they exist in the world?

The Distinctiveness of Digital Assets Compared with Traditional Assets

In January 2024, the U.S. Securities and Exchange Commission (SEC) approved the first spot Bitcoin exchange-traded funds (ETFs). Net inflows over the first full year (2024) reached approximately $35 billion; by the end of 2025, assets under management had climbed into the hundreds of billions of dollars, and 13F filings showed that institutional investors had become major holders [1]. The immediate catalyst was Grayscale's 2023 courtroom victory over the SEC: the court ruled that the SEC's rejection of Grayscale's application to convert its Bitcoin trust into an ETF lacked a reasonable basis, forcing the agency to approve the product under judicial pressure rather than through any genuine shift in its regulatory stance. The regulatory trajectory for digital assets thus remains deeply uncertain. The seventh annual global crypto hedge fund report, jointly published by AIMA and PwC, found that 55% of the hedge funds surveyed had added digital assets to their portfolios, nearly double the under-30% share of roughly two years earlier [2]. Together these figures mark the passage of digital assets from fringe experiment to the core of global institutional asset allocation. Yet in November 2022 the collapse of FTX erased more than $8 billion in customer assets [3]. Hundreds of billions of dollars in institutional inflows, set against the sudden evaporation of billions, define the tension of the present moment: institutional allocation has reached systemic scale, yet the safety infrastructure has not matured to match it. That tension frames the chapter's central question. Is the difference between digital assets and traditional assets merely a set of technical features to be enumerated, or is it a more fundamental reorganization of the trust paradigm?

Answering it requires first understanding, in precise terms, the institutional logic of traditional finance. This chapter opens with the full journey of an ordinary stock trade, following the sequential involvement of at least nine independent institutions, a settlement cycle longer than 24 hours, and successive layers of capital lockup. From that journey it distills the five-layer functional architecture of trust in traditional finance: asset custody, trade execution, clearing and settlement, data verification, and rule governance. Each layer has suffered a defining failure, and each failure has prompted a new round of institutional repair. The institutional costs abstracted from the journey and its failures, namely settlement delay, compliance overhead, entry barriers, information asymmetry, and cross-border friction, appear unrelated but share a single root.

In the absence of technological trust, society must build trust through institutions; because those institutions are costly, their cost passes systematically to market participants. This chapter unifies these dispersed costs under a single analytical concept, the trust tax, and uses it to show, dimension by dimension, how the six distinctive properties of digital assets (permissionlessness, atomic divisibility, embedded settlement finality, around-the-clock continuous trading, smart-contract execution, and on-chain transparency) compress that systemic cost from different directions.

Once these properties are examined individually, a deeper structural feature emerges: they interact multiplicatively. Financial primitives impossible in traditional finance, such as flash loans, maximal extractable value (MEV), and protocol-level arbitrage, all emerge from several properties acting at once. Composability is at once a channel for innovation and a channel for contagion. On-chain transparency, even as it removes old information barriers, creates a new asymmetry rooted in the capacity to process information, so that fairness itself must be redefined. More fundamentally, this transformation changes not only how markets operate but the statistical character of risk itself. Traditional finance faces a risk surface of low-frequency but systemic shocks; digital assets face one of high-frequency shocks of highly variable magnitude, and the two demand fundamentally different approaches to risk management. Finally, the layered blending of three trust paradigms (institutional trust, platform trust, and architectural trust) offers a framework for choosing rationally between the two systems, one that moves beyond a simple binary.

These analytical tools form the starting point for the chapters that follow. The market for digital assets is not a simple narrative of decentralization replacing centralization; it is the largest natural experiment in financial history, using trillions of dollars of real capital to test, in real time, which elements of two centuries of financial institutions are essential, which can be replaced by technology, and which will reveal their indispensability only in the next crisis.

1.1 The trust architecture of traditional finance

The depth of the digital-asset transformation can be measured only against a full grasp of the institutional logic of traditional finance. This section therefore follows a single trade from start to finish, through the hands of at least nine independent institutions, and abstracts from that journey the five layers of trust on which traditional finance rests. It then unifies the institutional costs these functions generate under the analytical concept of the trust tax and argues that those costs, though high, buy real services.

1.1.1 The complete journey of a trade

The complexity of the traditional financial system is not the product of redundant design; it is the institutional response, evolved over two centuries, to a fundamental problem: how to transfer value safely among strangers who do not trust one another. This apparatus is best understood not from an abstract architecture diagram but by tracing a single trade from order submission to asset delivery, identifying at each step why a given institutional node exists and what it costs.

Alice decides to buy 100 shares of Microsoft stock. Through her broker's trading software she enters a limit buy order: 100 shares of MSFT at $450 or below. Behind this action, the broker, the first layer of intermediation between investor and market in traditional finance, performs three functions. Account management requires Alice to complete identity verification and anti-money-laundering (AML) screening when she opens the account; this is the first threshold of market access, ensuring that a participant's identity is traceable and accountable within the legal framework. Risk management and credit provision allow Alice to trade on leverage in a margin account, for which the broker monitors her margin ratio in real time and retains the right to liquidate her positions when risk exceeds limits. Smart order routing searches across more than 10 public exchanges and dozens of alternative trading systems for the best execution price. The SEC's Regulation NMS requires brokers to meet a best-execution obligation for their clients, one of the regulator's core requirements [4].

Between Alice and the market, the traditional financial system forms a clear pyramid. At the top are regulators, such as the SEC, the CFTC, and national central banks, which set rules, conduct oversight, and define the operating boundaries of the entire market. In the middle are infrastructure institutions, such as exchanges, central counterparties, and central securities depositories, which perform the core functions of price discovery, risk management, and asset delivery. At the bottom are client-facing intermediaries, such as brokers, market makers, and custodian banks, the channels through which investors reach the market. Trust flows from top to bottom: the broker at the bottom is trustworthy because the exchanges and clearing institutions above bind it to their rules, and those middle-tier institutions carry authority because the regulators at the top vouch for them. This hierarchical dependency means that a failure at any layer can unsettle the trust foundation of every layer above and below it. Yet the hierarchy also provides firewall isolation: each layer's institutional design can arrest contagion independently. The central counterparty (CCP) default waterfall, for instance, is itself an institutional firewall at Layer 3, designed precisely to contain an individual member's default losses within that layer. This firewall stands in sharp contrast to the flattened contagion paths of on-chain composability, which Section 1.3.3 examines.

Alice's limit buy order is routed to the Nasdaq exchange, where it enters the central limit order book. Harris (2003) [4] notes that the central limit order book is the core mechanism through which modern securities markets achieve effective price discovery, whereas on-chain order books face different constraints in latency and front-running risk, a subject taken up in later chapters. All unfilled limit orders are ranked by price priority and then time priority: higher-priced buy orders and lower-priced sell orders take precedence, and orders at the same price are ranked by arrival time. When a new market order arrives, it fills against the best available counterparty price and works down the book. This process is anonymous, automatic, and completed within microseconds. Through it the exchange aggregates countless dispersed, anonymous trading intentions into a public, continuous, and widely accepted transaction price, which is the essence of price discovery. On the book sits a limit sell order placed by a trader, Bob: 200 shares of MSFT at $450. Alice's 100-share buy order matches Bob's sell order, and the matching engine executes the trade within microseconds. Alice's 100 shares are filled at $450, and the remaining 100 shares of Bob's order stay on the book.

The exchange sends trade confirmations to both brokers.

Yet making a trade is not the same as completing it. At this moment the two parties have merely reached a legally binding contract, a promise to exchange assets and funds at some future point; no transfer of the assets or funds themselves has yet occurred. From here the trade enters the more complex phase of post-trade processing. The gap between execution and asset delivery is the very source of all the institutional complexity of traditional finance: it creates a window of counterparty-risk exposure, and the entire post-trade apparatus exists to manage the defaults, errors, and systemic risks that may arise within that window.

The post-trade phase brings in the single most important institutional innovation of traditional finance: the central counterparty. In the U.S. equity market, post-trade processing follows a T+1 settlement cycle, and before May 2024 it ran as long as T+2. The National Securities Clearing Corporation (NSCC), a subsidiary of DTCC, interposes itself in every trade through a legal process known as novation, becoming the seller to every buyer and the buyer to every seller. This novation is completed during a batch process on the evening of the trade date; until then, the two parties still bear the credit risk of their original counterparties. The original contract between Alice and Bob is replaced by two independent contracts, Alice with the NSCC and the NSCC with Bob, so that neither party owes a delivery obligation directly to the other. As long as the NSCC does not default, the trade's final completion carries an institutional guarantee. This arrangement lets market participants trade with anonymous counterparties in confidence, and it deepens market liquidity substantially. To manage the concentrated risk it assumes, the NSCC builds a multi-layered defense, the loss-sharing waterfall, whose layer-by-layer absorption sequence is detailed in Figure 1-1 of Section 1.1.2. Through strict membership admission, initial and variation margin posted against unsettled positions, and a default fund contributed jointly by all clearing members, it ensures that the clearing system keeps operating even if a member defaults. At the same time, the NSCC performs multilateral netting across the thousands of trades in a single day, offsetting a given clearing member's buys and sells in the same asset to compute a final net position. This reduces the volume that actually requires settlement by an order of magnitude and markedly lowers operating costs and liquidity needs.

Final asset delivery is carried out by the central securities depository, the DTC, also a DTCC subsidiary. The DTC is the centralized ledger that holds and records ownership of U.S. securities. Settlement is completed through delivery versus payment (DVP): the delivery of securities and the payment of funds are mutually conditional and occur simultaneously, eliminating principal risk, a principle realized more completely on-chain in the form of atomic settlement. Only after the final book-entry update on T+1 are 100 shares of MSFT formally credited to Alice's securities account, and only then does Bob receive the corresponding funds.

A seemingly simple purchase of 100 shares involves, in the back office, the sequential participation of at least nine independent institutions (broker, exchange, matching engine, clearing corporation, central counterparty, depository, payment system, regulator, and various data-service providers); it takes more than 24 hours, freezes substantial capital during settlement, and generates friction at every step. According to the post-

transition assessment jointly published by SIFMA, ICI, and DTCC, the NSCC clearing fund fell by an average of about $3 billion after the move to T+1, a decline of 23% [5]. Even under a T+1 cycle, then, settlement delay still levies a considerable cost in time. The internal logic of traditional finance becomes clear here: settlement delay buys the time that risk management requires, and centralized institutions supply the credit guarantees that make anonymous trading possible. These intermediation steps, compliance reviews, and waiting periods constitute the institutional cost society pays to keep trading safe in the absence of technological trust.

1.1.2 The five layers of trust

Alice's trading journey yields the five-layer functional architecture of trust in traditional finance. These layers are not the product of a single design; they are the accumulated sediment of two centuries of repeated failure, repair, further failure, and further repair. Each layer carries a core promise, and each has dedicated institutions to secure it. The primary lens for understanding these layers is not their flaws but the real protection they afford participants, protection that makes possible the worldwide allocation of hundreds of trillions of dollars.

The asset-custody layer ensures that client assets are strictly segregated from the intermediary's own assets. In the United States, SIPC insures each brokerage account up to $500,000 (of which $250,000 covers cash), and the FDIC insures bank deposits up to $250,000. These institutionalized safety nets allow ordinary savers and investors to participate in the market with confidence, without needing to understand the workings of the financial system. Alice need not assess her broker's solvency; the institution has made that judgment for her. The trade-execution layer ensures that orders are handled fairly, through independent exchanges and the best-execution obligation. The SEC's Regulation NMS requires brokers to seek the best price for clients across multiple venues rather than routing orders to whichever venue most benefits the broker itself [4]. At the clearing-and-settlement layer, the CCP mechanism expresses the protective power of traditional institutional design in its most concentrated form. Figure 1-1 presents the multi-layered defensive structure of the CCP loss-sharing waterfall: losses are absorbed starting with the defaulter's own resources and progressing to industry-wide mutualization, on the core principle that whoever creates the risk bears the loss first.

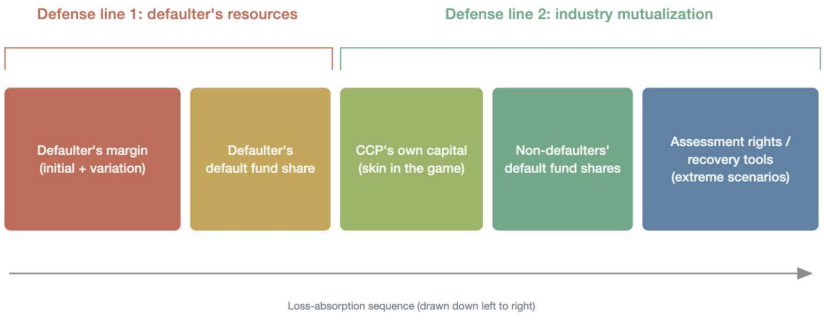


Figure 1-1. The CCP loss-sharing waterfall: when a clearing member defaults, losses are absorbed layer by layer in a preset order, extending from the first line of defense (the defaulter's own resources) to the second (industry mutualization), protecting the entire financial system's continued operation (the tiers indicate the sequence of loss absorption, and their widths do not represent proportional amounts; the default-management auction is a disposal step preceding the financial waterfall, and the tail includes assessment rights and recovery tools)

Most major CCPs contribute a portion of their own capital to the default waterfall, positioned ahead of the default fund contributions of non-defaulting members (the exact proportion and arrangement vary by institution and jurisdiction). This skin in the game ensures that the CCP has ample economic incentive to manage risk prudently. The waterfall defense withstood an extreme test in the 2008 financial crisis: when Lehman Brothers failed, tens of thousands of unsettled derivatives contracts were involved, yet the CCP system ultimately completed all clearing, and not a single CCP itself defaulted [6]. The data-verification layer ensures the authenticity of market prices through independent audits and regulated data vendors, providing trustworthy inputs for asset pricing and risk management. The rule-governance layer ensures the legitimacy and predictability of rule changes through the legislative process and the judicial system, giving market participants a stable set of expectations; a complete system of legal remedies, from arbitration to litigation to class action, offers investors a clear path of recourse.

The cumulative effect of these five layers of protection is a kind of trust free of understanding, the institutionalized expression of what Luhmann (1979) [7] defined as system trust. Most people need not understand the CCP default waterfall to entrust their savings to the stock market with confidence; this default trust is precious social capital accumulated over two centuries of institutional track record. But system trust exhibits a threshold effect: it accumulates gradually and is lost suddenly. Each layer of protection, however, has also suffered a defining failure that revealed the boundary of institutional safeguards. Table 1-1 sets out the five trust functions layer by layer, listing for each its core promise, the specific institutional arrangements that deliver it, and its most consequential historical failure. From the $1.6 billion scandal of MF Global's misappropriation of client funds at Layer 1 (asset custody) to the LIBOR manipulation affecting the pricing of $350 trillion in financial products at Layer 4 (data verification), each failure precisely exposed the boundary of the corresponding layer's safeguard.

Layer	Trust function	Core promise	Institutional implementation	Defining failure
Layer 1	Asset custody	Assets will not be misappropriated	CSD/DTC; SIPC insurance ($500,000 per account); FDIC deposit insurance ($250,000)	MF Global (2011): the futures broker behind the eighth-largest bankruptcy in U.S. history misappropriated $1.6 billion of segregated client funds to speculate on European sovereign debt [8]
Layer 2	Trade execution	Orders are executed fairly	Independent exchanges; best-execution obligation; market-surveillance systems	Dark pool scandals (2014–2016): dark pools operated by major banks such as Barclays systematically allowed high-frequency trading firms to front-run client orders; the SEC imposed $70 million in penalties [9]
Layer 3	Clearing and settlement	Trades will be delivered	CCP default waterfall; multilateral netting	Lehman Brothers (2008): the bankruptcy of the world's fourth-largest investment bank required the emergency handling of tens of thousands of unsettled OTC derivatives contracts, exposing the fatal weakness of OTC derivatives that lacked central clearing [6]
Layer 4	Data verification	Prices and data are authentic	Regulated data vendors; independent audits; benchmark-rate administrators	LIBOR manipulation (2012): the world's most widely referenced benchmark rate was systematically manipulated by more than 10 banks for years, affecting the pricing of over $350 trillion in financial products, with cumulative fines exceeding $9 billion [10]
Layer 5	Rule governance	Rule changes are lawful and transparent	SEC/CFTC; legislative process; judicial system	Post-2008: the financial crisis exposed systemic failures of the regulatory apparatus (the SEC failed to detect Madoff's 17-year Ponzi scheme), prompting the sweeping institutional repair of the Dodd-Frank Act [6]

Table 1-1. The five-layer trust functions of traditional finance and their defining failures (Data source: public records of the respective regulators [8][9][6][10]. Case-selection criterion: each layer's case is a landmark event with a completed regulatory investigation and an impact large enough to drive

institutional reform)

These failures should not be read simply as institutional incompetence. On the contrary, each prompted stricter institutional repair, and the system's capacity for self-correction is itself part of institutional trust. After MF Global, the CFTC strengthened the requirements for segregating client funds; after the LIBOR scandal, benchmark-rate setting shifted from banks' self-reported quotes to calculation from real transaction data; after the 2008 crisis, the Dodd-Frank Act brought previously unregulated OTC derivatives into central clearing and raised CCPs' capital and risk-control standards across the board. This cycle of failure, repair, further failure, and further repair yields three conclusions: trust is an institutional product that must be produced continuously rather than a natural endowment; the cost of failure is real and quantifiable; and these five functions give the later analysis a precise target as digital assets attempt to replace institutions with technology. Architectural trust must prove that it delivers at least an equivalent level of protection at each layer, or honestly acknowledge which layers it cannot yet replace. This five-layer structure is developed further in the verifiability ladder of Chapter 5 and the trust-stack design of Chapter 28.

1.1.3 The trust tax: a unified view of institutional cost

Alice's journey and the five layers of trust distill into a single unifying concept: the trust tax, the systemic cost society pays to maintain orderly trading in the absence of technological trust. The concept overlaps with, but is sharply distinct from, the transaction-cost economics of Williamson (1985) [11]. Williamson attributes transaction costs to two behavioral assumptions, bounded rationality and opportunism, and they span a broad range of costs such as search, bargaining, and monitoring. The trust tax is a specific subset of transaction costs: it concerns only the institutional costs that arise because the parties to a trade cannot establish trust at the technological level. Search and bargaining costs are not part of the trust tax, because they persist even under complete trust. The analytical value of this distinction is that it identifies precisely which costs digital assets can compress: architectural trust can replace part of what institutional trust does, but it cannot eliminate trust-independent frictions such as search and bargaining costs. The many seemingly unrelated trust costs of the traditional financial system all trace to the same root, and they are levied across six dimensions.

> The trust tax: the institutional cost society bears because it cannot establish trust at the technological level, the specific subset of transaction costs that architectural trust can compress.

These six dimensions are not an arbitrary list; they are extracted systematically from Alice's trading journey in Section 1.1.1, each corresponding to a class of observable institutional friction. Liquidity fragmentation (the splitting of liquidity across exchanges) and standardization costs (incompatible protocols across markets) are also real market frictions, but their root cause is coordination rather than an absence of trust, and they are therefore excluded from the trust-tax framework.

The settlement-delay tax manifests as the capital lockup and counterparty-risk window

created by the T+1 settlement cycle. Alice's funds are locked from T to T+1, unavailable for other trades, with counterparty-risk exposure exceeding 24 hours. As noted in Section 1.1.1, even though T+1 reduced the clearing fund by about $3 billion [5], the delay tax itself remains considerable. Figure 1-2 presents the historical evolution of the U.S. securities settlement cycle from T+5 to T+1 over more than three decades, each shortening demanding an industry coordination cycle on the order of a decade.

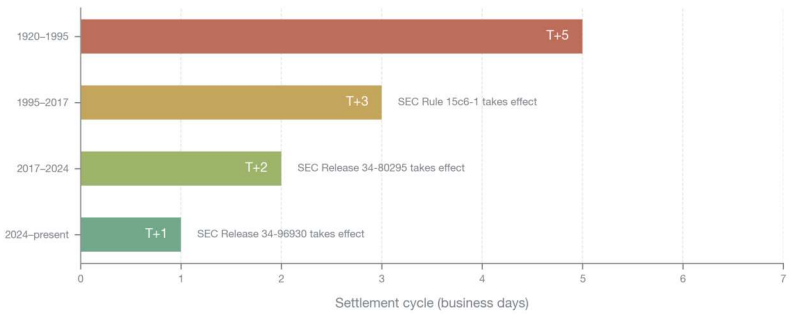


Figure 1-2. The historical evolution of the U.S. securities settlement cycle (Data source: SEC Rule 15c6-1, SEC Release 34-80295, SEC Release 34-96930 [5]; as of June 2026, T+1 remains the settlement cycle in force)

Each shortening released frozen capital, but the path itself reveals how entrenched settlement delay is: even when the technology had long permitted faster settlement, institutional inertia and coordination costs still held reform to a roughly decade-long cadence.

The trading-hours-fragmentation tax arises because restricted trading hours prevent information from being priced continuously. The New York Stock Exchange's regular session runs from 9:30 to 16:00 Eastern time, and the London Stock Exchange's from 8:00 to 16:30 Greenwich Mean Time; both close on weekends and public holidays. The U.S. stock market has roughly 252 trading days a year, meaning that on about 31% of days it does not trade at all. Information accumulates continuously while the market is closed but cannot be reflected in prices until it reopens, producing opening gaps and weekend risk exposure. If Alice holds positions in both New York and London, she also faces hedging difficulties created by the incomplete overlap of the two sessions.

The intermediary tax comes from the operating costs of brokers, clearinghouses, custodian banks, and similar institutions, costs ultimately borne by investors. Alice's trade passes through at least nine intermediaries, each with operating costs, compliance costs, and profit objectives. According to industry data cited by DTCC in advancing the T+1 reform, clearing and settlement is the main component of post-trade operating expenditure in traditional finance, accounting for roughly 20% to 25% [5]. The entry tax manifests as the licenses, capital requirements, and compliance thresholds that exclude the vast majority of potential participants. Alice must pass identity verification and AML screening before she can open an account; a market maker needs millions of dollars in capital and a months-long approval process; and the threshold for CCP clearing membership is higher still. Each entry barrier narrows the pool of participants and

limits competition.

The opacity tax arises from the information asymmetry that barriers to data access open up between institutions and retail investors. Professional-grade market-data terminals cost tens of thousands to hundreds of thousands of dollars a year [12], and Alice cannot obtain market data of the quality available to institutional investors. The LIBOR scandal [10] further demonstrated that even with institutional safeguards, key market data can still be manipulated through collusion among a handful of insiders; the question of who watches the watchers has never been fully resolved in traditional finance. The jurisdictional tax comes from the legal, time-zone, and currency frictions of cross-border transactions. According to World Bank data, the average cost of a global cross-border remittance is about 6.35% of the amount sent (based on a $200 remittance benchmark, first-quarter 2024 data; 6.4% in the fourth quarter of 2023) and takes three to five business days [13]. This figure reflects the cost structure of small retail remittances; institutional cross-border transfers are completed through the SWIFT/CLS system, and the percentage cost of large transfers is markedly lower. The trust tax is levied in structurally different proportions across participant tiers. If Alice wants to buy shares listed in London, she faces time-zone differences, exchange-rate risk, and multiple regulatory-compliance requirements all at once.

The analytical value of the trust-tax concept is that it unifies dispersed institutional costs into a single causal account: every cost traces to the same root. Parties cannot trust one another at the technological level, so institutions must supply that trust; and because institutions are costly, the cost is passed on. The five-layer function defines what trust must do; the trust tax measures what it costs. This framework gives precise coordinates for the analysis that follows: each of the six distinctive properties will be shown to compress some dimension of the trust tax, and the value of any financial innovation can be measured by how much trust tax it compresses against how much new cost it introduces.

1.1.4 The rationale for the trust tax

The trust tax is not useless institutional redundancy; it corresponds to four classes of real service. Systemic stability is the first. In the 2008 financial crisis, the CCP default waterfall proved to be the key institutional safeguard for the continued operation of the global financial system: as noted in Section 1.1.2, not a single CCP itself defaulted in the Lehman crisis [6]. The cost of this stability (the margin system, default funds, operating costs) is precisely the trust tax. Investor protection is the second: the institutionalized SIPC/FDIC safety net described in Section 1.1.2 lets the vast majority of people participate in the market with confidence, without understanding how the financial system works, and in doing so mobilizes trillions in capital. FTX's loss of more than $8 billion in customer assets [3] shows exactly what the absence of such protection costs. The FTX disaster was in essence a failure of the platform-trust model, not a defect of on-chain architectural trust: a centralized platform reproduced, in a new form, the misappropriation risk of traditional intermediaries, a distinction systematized in Section 1.4.3. Legal recourse is the third service. When a trading dispute arises, traditional finance provides a complete system of legal remedies from arbitra-

tion to litigation to class action, and investors have clear legal standing and a path of recourse. In on-chain finance, once assets are transferred through a code vulnerability or an attack, there is almost no legal mechanism for recovery; irreversibility is both a feature and a cost. The cumulative effect of institutional reputation is the fourth service. Two centuries of track record have created a default trust, the trust free of understanding discussed in Section 1.1.2. Although the DYOR culture of on-chain finance is healthier in principle, in practice it places the entire burden of risk management on each user. Behavioral-finance research shows that, faced with information overload, individual investors typically do not conduct genuinely independent research but rely on social proof and narrative shortcuts, which in crypto communities take the form of following KOLs, imitating smart-money addresses, and FOMO-driven decisions. The gap between the principle and the practice of DYOR is systemic, and self-directed diligence is unrealistic for most ordinary people.

The right question, then, is not how to abolish the trust tax but whether its current level far exceeds the value of the services it provides, and whether technology can deliver an equivalent level of service more efficiently. The existing evidence suggests that it can, but with one important qualification. Some dimensions of the trust tax can be compressed completely by technology: atomic settlement eliminates the T+1 delay, and on-chain transparency removes information barriers. Others can currently be compressed only in part; here the governance layer's trust tax stands out. The question of who has the right to modify the code is one the code itself cannot answer, and discretion in extreme cases may still require human judgment. At the governance layer, this projects the power and limits of *code as law* that Lessig (2000) [14] described (elaborated in Section 1.2.5). Whether the trust tax has an incompressible floor, whether some trust functions cannot in principle be replaced by code, is a question that runs through the entire book and is examined formally in the final chapter. Acknowledging the value of the trust tax while arguing that it can be compressed is the analytical stance that raises the book's framework above the one-dimensional narrative in which traditional finance is inefficient and digital assets are omnipotent.

1.2 The six distinctive properties and the compression of the trust tax

The six distinctive properties of digital assets are not a flat inventory of technical features; each property is a direct response to one dimension of the trust tax. This section adopts a single analytical framework throughout: the target of the trust tax, the technical response of digital assets, and the cost and limits of that response.

1.2.1 Permissionlessness: compressing the entry tax

The entry tax is the most perceptible dimension of the trust-tax system in traditional finance. From identity verification and anti-money-laundering screening for individual investors, to the capital and technical requirements imposed on market makers, to the top-tier banking credentials required of the clearing members of a central counterparty, traditional finance is essentially a system of nested permissions in which each layer

narrows the range of participants. Across the five layers of trust, the entry threshold at each layer is a direct expression of the entry tax.

Public blockchains fundamentally restructure this logic. Access rules are defined by code rather than administrative approval, and anyone who complies with the protocol's rules can participate. On Uniswap, for example, any user can interact directly with the smart contract to trade tokens, or become a de facto market maker by providing liquidity, without approval from any institution. The marginal cost of deploying a smart contract is minimal and the exit cost is zero; trust derives not from institutional endorsement but from the open-source nature of the code and the mathematical determinism of the protocol. Users retain full control of their assets by holding their private keys—a fundamental departure from the identity-based, permissioned system of traditional finance. Baumol's (1982) [15] theory of contestable markets provides an economic framework for understanding this shift: the persistent threat of potential competition forces incumbents to hold prices near marginal cost, systematically compressing the scope for rent-seeking. On-chain activity is not entirely free of sunk costs—gas fees, impermanent loss, and smart-contract deployment costs all constitute de facto entry barriers—but these costs are several orders of magnitude lower than the millions of dollars in capital and months-long approval cycles required to become a market maker in traditional finance. Market-making thus shifts from the privilege of a few institutions to open competition in which anyone can participate, and thousands of retail participants pool idle assets into liquidity pools whose total value locked reaches tens of billions of dollars. In execution quality (slippage and the capacity to absorb large orders) and dynamic price adjustment, however, automated market makers (AMMs) still lag well behind professional market makers, and liquidity providers bear the risk of impermanent loss. Even so, this permissionless model of liquidity provision has no precedent in traditional finance. On order-book decentralized exchanges (DEXs) such as Hyperliquid, permissionlessness likewise means that anyone can submit orders, yet effective market-making still requires sophisticated pricing algorithms and low-latency infrastructure; a lower entry threshold does not eliminate competitive advantage. This order-book model more accurately reflects the market structure that is the central subject of this book.

Permissionlessness compresses not only the entry tax but also the jurisdictional tax, in three directions at once. By removing geographic entry barriers, protocols on public blockchains treat users worldwide identically, without cross-border account opening or multiple layers of regulatory approval. Around-the-clock trading eliminates time-zone fragmentation. Atomic settlement eliminates cross-border settlement delay. A traditional cross-border remittance takes 3 to 5 business days and costs about 6.35% of the amount remitted (for the basis of this figure, see Section 1.1.3) [13], whereas an on-chain transfer completes in seconds to minutes and typically costs less than $1 on Layer 2 networks or low-gas-fee chains, though on Ethereum mainnet the cost can reach tens of dollars during periods of congestion. Together, these three shifts systematically compress the jurisdictional tax, moving financial services from the constraints of geographic jurisdiction toward global uniformity at the protocol level.

Permissionless access also lowers the cost of malicious behavior. It sits, at the same time, in growing tension with legal compliance obligations: the Financial Action Task

Force (FATF) Travel Rule [16] and the European Union's Markets in Crypto-Assets (MiCA) regulation are beginning to constrain technical permissionlessness from the legal side, and technical accessibility is not the same as legal accessibility. A lower entry threshold also lowers the cost of fraud: according to CertiK, crypto-asset losses from security incidents totaled approximately $3.35 billion over the course of 2025 [17]. Permissionlessness has also given rise to new strategic behaviors such as MEV bots and sandwich attacks, which preserve price efficiency while introducing new fairness problems. The net social benefit of permissionlessness depends on the maturity of the accompanying safeguards—disclosure regimes, reputation systems, and investor-protection tools.

1.2.2 Atomic divisibility: compressing the granularity barrier

Beyond licensing and compliance barriers, the entry tax also operates along an easily overlooked dimension, granularity—the institutional constraint on an asset's minimum tradable unit. When a single Class A share of Berkshire Hathaway trades above $600,000, the minimum investment unit is itself an entry barrier that excludes the vast majority of investors. The divisibility of traditional assets is tightly constrained by physical, legal, and institutional factors: a stock split requires a complex corporate action, trading distinguishes between round lots and odd lots, and price movements are limited by a minimum tick of $0.01 [4].

Digital assets remove these constraints at the protocol level. A single bitcoin can be divided into 100 million satoshis, and ether is divisible into as many as 10^{18} base units. On chain, transferring 0.000001 BTC is no different at the protocol level from transferring 1,000 BTC. In microstructural terms, this atomic divisibility means that prices can be distributed across a nearly continuous spectrum, giving the order book a price granularity far finer than that of traditional markets and thereby, in principle, improving the precision of price discovery. Combined with low transaction costs, atomic divisibility makes micropayments economically viable and provides the technical foundation for entirely new business models such as streaming payments and per-second billing. Together with the trend toward asset tokenization, atomic divisibility becomes the technical precondition for any asset to be split into arbitrarily small fractions and circulated worldwide—whether a tokenized U.S. Treasury fund or the tokenized real estate that may emerge in the future [18].

This property is not without cost. Excessively fine price granularity increases the complexity of the order book: when prices can be specified to 18 decimal places, liquidity is dispersed across an enormous number of price levels, the order book becomes *wide and shallow*, and effective liquidity may actually decline. Divisibility precision at the asset level must, however, be distinguished from the tick-size setting at the trading level. On real on-chain order-book DEXs, the tick size is set explicitly by protocol parameters—for a BTC/USD perpetual futures contract, for instance, the tick size is typically $0.1 to $1—far coarser than the smallest divisible unit of the underlying token; the risk of a wide, shallow book therefore arises more from the continuous price curve of AMMs than from the order book. This is also one of the sound reasons traditional markets impose a minimum tick. Price precision exhibits diminishing marginal returns: once

precision exceeds a trader's decision resolution, additional precision no longer creates economic value. An extremely low minimum investment threshold democratizes access, but it may also attract risk-unaware small investors who bear disproportionate losses in highly volatile markets.

1.2.3 Settlement finality: compressing the time tax

The time tax is one of the costliest dimensions of the trust-tax system. Alice's funds are locked between day T and day T+1, and this time gap has given rise to an entire institutional apparatus—central counterparties, margin systems, and loss-sharing waterfalls. Layer 3 of the five layers of trust is precisely the institutional response to this time gap.

Digital assets compress trade and settlement cryptographically into a single atomic operation: to trade is to settle. Once a transaction receives a sufficient number of network confirmations, the probability that it will be reversed falls exponentially with each additional confirmation. Different consensus mechanisms provide different types of finality guarantee. Bitcoin provides probabilistic finality; industry convention treats a transaction as effectively irreversible after 6 block confirmations (about 1 hour). Ethereum proof-of-stake (PoS) reaches economic finality—backed by validators' economic stake—after approximately 12.8 minutes. Purpose-built application chains typically reach finality on subsecond-to-second timescales. Whatever the type, this finality is secured by the computational resources or economic stake that the network commits, not by the promise of any centralized institution. In its 2025 annual report, the Bank for International Settlements (BIS) emphasizes that atomic settlement is a core feature of the next generation of monetary and financial systems: it allows assets to be exchanged simultaneously and thereby, in principle, eliminates principal risk [19]. Capital efficiency improves fundamentally as a result: market makers can recycle inventory more quickly, and operational friction across the financial system falls markedly. Atomic settlement, however, comes at a cost: while it eliminates the net-settlement time window, it also means that every transaction requires full, immediate funding, correspondingly raising the system's demand for instantaneous liquidity. Whether an equivalent batch-settlement mechanism (such as frequent batch auctions) exists on chain is an open design question.

Atomic settlement eliminates the settlement cascade—the chain of defaults triggered when one party fails to deliver on time, of which the domino effect set off by the collapse of Lehman Brothers is the classic case—but it cannot eliminate the liquidation cascade: the positive-feedback loop in which falling prices trigger forced liquidations, selling pressure, further declines, and still more liquidations. Atomic settlement not only fails to halt the liquidation cascade; by executing efficiently it may even accelerate the cascade's spread, compressing a liquidation process that takes days to weeks in traditional finance into minutes to hours. The factors that accelerate a liquidation cascade include the maintenance-margin ratio, the smoothing of the mark price, and the size of the insurance fund; Chapter 11 analyzes these design choices systematically, in its treatment of the endogenous reflexivity of leverage and liquidation. The focus of risk thus shifts from whether the counterparty will perform to three new dimensions:

whether the protocol code is correct, whether there is sufficient liquidity amid extreme volatility, and whether external data inputs are reliable. In perpetual futures, the mark price depends on oracle price feeds, and an oracle failure or manipulation can directly trigger an erroneous liquidation—neither a code vulnerability nor a liquidity problem, but a reliability risk in external data. This is the substitution of risk, not its elimination.

> The settlement cascade and the liquidation cascade are two fundamentally different mechanisms of risk transmission. Atomic settlement eliminates the former but may accelerate the latter—which explains why on-chain finance, having eliminated counterparty risk, still experiences violent shocks.

1.2.4 Around-the-clock trading: compressing temporal fragmentation

The fragmentation of trading hours makes it impossible to price information continuously. The regular trading session of the New York Stock Exchange covers only 6.5 hours of each business day, and information that arises during the remaining hours cannot be reflected in prices immediately. Blockchain networks have no concept of a market open or close: any piece of information can be priced immediately, turning the hard constraint of being unable to trade into the soft constraint of trading costs that vary across the hours of the day. Trading activity worldwide follows a cyclical rotation of liquidity over each 24-hour period. Figure 1-3 depicts this global liquidity cycle: trading activity across the three major time zones overlaps in successive waves, with liquidity concentrating in the overlap windows and falling markedly during the late-night gaps.

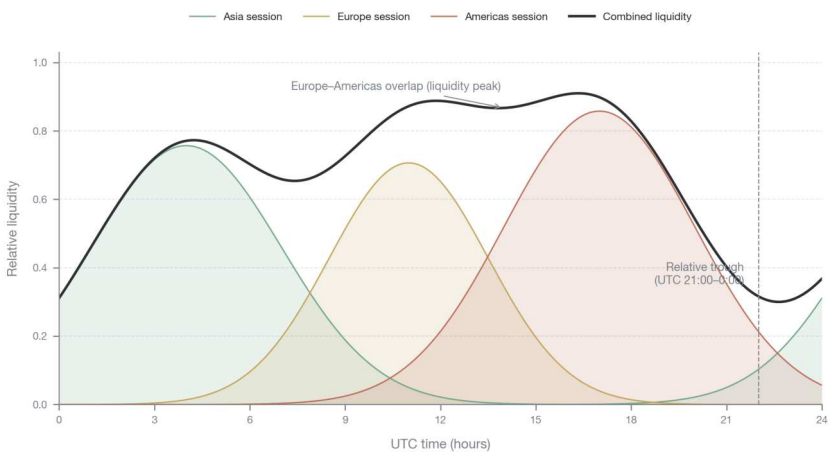


Figure 1-3. The 24-hour global liquidity cycle of digital-asset markets—the overlapping liquidity of the three major time zones: Asia, Europe, and the Americas (schematic modeling, not measured volume; because digital assets trade continuously 24/7, their troughs are shallower than those of foreign-exchange markets, and the timing of the overlap peaks drifts slightly with daylight saving time)

The activity peak of the Asian session typically falls between 1:00 and 7:00 UTC, and the European session between 8:00 and 14:00 UTC; when the European and American

sessions overlap (roughly 13:00 to 15:00 UTC), the market reaches its daily liquidity peak, while liquidity falls to a trough between 21:00 UTC and 0:00 the following day. During this low-liquidity window, the bid-ask spread can widen to several times its peak-session level, order-book depth drops noticeably, and the market-impact cost of large trades rises accordingly; this day-night variation in execution quality is a hidden cost specific to around-the-clock markets. Such cyclical rotation of global liquidity requires market makers and high-frequency trading firms to build global teams and infrastructure that continuously cover every session. The weekend effect—rising volatility accompanied by falling liquidity—already has systematic empirical support in crypto markets: Baur et al. (2019) [20], analyzing the intraday and cross-day patterns of Bitcoin trading volume and returns, confirm the statistical significance of this pattern.

Around-the-clock trading eliminates the systemic gap risk caused by market closures, but it does not eliminate price jumps caused by discontinuous liquidity or extreme events: when order-book depth thins abruptly during a low-liquidity period, a large order or a sudden news item can still produce a substantial price discontinuity. At the same time, around-the-clock trading removes three implicit protective functions that market closures once provided. After the close, market makers have time to reassess inventory, adjust models, and clear risk exposures; in an around-the-clock market this window for risk reassessment no longer exists, and market makers must manage all risk in real time under continuous operation. The suspension of trading during a closure dampens the immediate transmission of panic selling; in an around-the-clock market a liquidity crisis can spread without restraint in the thin-liquidity environment of the small hours. During a closure, an exchange can update its risk-control models and repair technical vulnerabilities; in an around-the-clock market such maintenance must be performed while the system is running. Exchanges in traditional markets have begun to explore extending trading hours: the Depository Trust & Clearing Corporation (DTCC) and Ernst & Young have discussed the possibility of a transition toward around-the-clock settlement [21], and BlackRock, in market research from the same period, likewise argues that 24-hour trading has structural implications for capital markets [22]. This reflects, in part, the competitive pressure that the around-the-clock trading model of digital assets exerts on traditional market structure; but it also demonstrates that around-the-clock operation is not without cost—otherwise traditional markets would have adopted it long ago.

1.2.5 Smart-contract execution: compressing the intermediary tax

The intermediary tax arises from the costs of interpreting and enforcing legal contracts, the inefficiency of manual processes, and the operating overhead of multiple layers of intermediaries. In traditional finance, each of the five layers of trust requires a specialized institution to operate it, and these operating costs constitute the core of the intermediary tax.

Smart contracts encode the terms of a financial agreement as a program that executes automatically on the blockchain, replacing ambiguous natural language with deterministic computation and human-intensive performance with automated execution. Lessig (2000) [14] presciently observed that code would become a regulatory force of deci-

sive influence in cyberspace, its power deriving both from its determinism and from its unavoidability. Smart contracts systematically replace the core functions of traditional intermediaries and have given rise to entirely new market mechanisms such as AMMs, decentralized lending, and algorithmic stablecoins. MakerDAO—a decentralized issuance system built from a cluster of smart contracts—shows that certain financial functions once borne mainly by sovereign institutions can be implemented in code [23]. This functional substitution, however, faces a challenge of legal characterization: MakerDAO's 2024 restructuring and rebranding to Sky was driven in part by regulatory pressure; DAI faces the risk of being classified as electronic money or a payment instrument in several jurisdictions; and MiCA imposes strict reserve and compliance requirements on stablecoin issuance. Once deployed, an AMM-based decentralized exchange executes its core matching function automatically through smart contracts, at very low marginal operating cost and without the back-office operations teams, physical trading infrastructure, and tiered compliance departments that a traditional exchange requires. Order-book DEXs such as Hyperliquid and dYdX v4, however, run on purpose-built application chains and require the continuous operation of a validator or sequencer network; although their total operating cost is far below that of a traditional exchange, it is nowhere near zero.

The principle of *code as rule* carries an inherent tension. Code cannot exercise discretion, cannot handle extreme situations its designers did not foresee, and, when code is law, the vulnerabilities in that code are executed automatically as well. According to Immunefi, losses from hacks and fraud of all kinds in 2023 alone amounted to approximately $1.8 billion (of which hacks accounted for about $1.6 billion, mostly related to DeFi smart contracts); this figure encompasses many distinct attack types—smart-contract logic flaws, cross-chain bridge attacks, and oracle manipulation—each with a different risk profile and defensive strategy [24]. Code must be upgraded, and the question of who has the authority to modify the code is itself a governance problem that only an institution can answer. Code as rule does not escape human governance entirely; it merely shifts the focus of governance from the enforcement of rules to their modification. Replacing legal compliance with code audits is not costless: formal verification and security audits themselves constitute a new trust tax, typically costing tens of thousands to hundreds of thousands of dollars depending on the complexity of the contract—though this remains far below the cost of traditional legal compliance.

1.2.6 On-chain transparency: the opacity tax and the transparency paradox

The opacity tax arises from information barriers: the information asymmetry between institutions and retail investors, and the monopolization of key market data by a handful of institutions. Layer 4 of the five layers of trust exists precisely to address the risk that data may not be truthful, yet this verification is itself opaque—the LIBOR scandal [10] demonstrated that data can still be manipulated even under institutional safeguards.

On a public blockchain, every transaction record and state change is publicly inspectable, and anyone can audit it independently. The actual degree of transparency,

however, depends on the underlying architecture: there is a marked difference between the full openness of Ethereum mainnet and the partial centralized operation of application chains (Hyperliquid's sequencer, for example, is operated by the team), where the accessibility and auditability of data depend on the transparency built into the sequencer's design. On-chain data-analytics platforms such as Dune Analytics, Nansen, and Arkham Intelligence parse, aggregate, and visualize raw data, greatly lowering the barrier to data analysis. For the first time, retail investors have the opportunity to obtain raw data of almost the same scope as institutions; yet democratizing access to data is not the same as democratizing decision quality. Behavioral-finance research suggests that greater access to information may increase overconfidence and trading frequency rather than improve investment decisions. The source of trust shifts, in part, from institutional brands and regulatory licenses to open-source code and verifiable data.

While it compresses the opacity tax, on-chain transparency creates a paradox of its own—the transparency paradox. In traditional finance, information asymmetry originates in differential rights of access to data, whereby some participants can reach information that others cannot. In on-chain finance, all participants can observe the same raw data, yet information asymmetry does not thereby disappear; it shifts from an asymmetry of access rights to an asymmetry of processing capability. Participants with stronger computational infrastructure and algorithmic capability can extract tradable signals more quickly from the same public data. Greater transparency therefore does not eliminate the information advantage but shifts the dimension of competition from data acquisition to data processing. MEV is the concentrated expression of this paradox: a public mempool makes all pending transactions visible to everyone, yet only searchers with the fastest infrastructure and the most sophisticated algorithms can convert that visibility into profit. The European Securities and Markets Authority (ESMA, 2025) [25] notes that MEV searchers typically pay more than 90% of their revenue to block proposers, making the right to order transactions a scarce resource; to address this, the Ethereum community has developed proposer-builder separation (PBS), which attempts to channel this competition from disorderly gaming into an orderly, market-based auction. Two forms of MEV must be distinguished, however: Ethereum-style public-mempool MEV (governed through the PBS architecture) and application-chain sequencer-style MEV, in which the sequencer determines transaction ordering—closer to a traditional exchange's last-look privilege. Later chapters analyze the MEV problems specific to order-book DEXs. Although on-chain markets are fully transparent, they are not necessarily fairer than traditional markets; the very definition of fairness needs to be re-examined—one of the central topics of the market-quality evaluation in Chapter 25.

> The transparency paradox: on-chain finance moves the competitive dimension of information asymmetry from the right to access data to the capability to process it. Transparency does not equal fairness.

Pseudonymity is also not the same as anonymity. The possibility of linking on-chain addresses to real-world identities through transaction-graph analysis has raised privacy concerns. The search for a balance between radical transparency and necessary privacy

is a key driver of frontier cryptographic techniques such as zero-knowledge proofs [26].

Each of these six distinctive properties compresses the trust tax along a different dimension: permissionlessness compresses the entry tax and the jurisdictional tax; atomic divisibility compresses the granularity dimension of the entry tax; embedded settlement finality compresses the time tax; around-the-clock trading compresses the trading-hours-fragmentation tax; smart contracts compress the intermediary tax; and on-chain transparency compresses the opacity tax. The mapping between the six properties and the six dimensions of the trust tax is not one-to-one but many-to-many: permissionlessness, for instance, compresses both the entry tax and the jurisdictional tax, while around-the-clock trading and atomic settlement jointly act on the time-related dimensions of the trust tax. This cross-mapping foreshadows the multiplicative interaction effects discussed in the next section. Thus far, however, each of these six analyses has proceeded in isolation, and the relationships among the dimensions have not yet been examined. Yet deep multiplicative interactions exist among the six properties, and it is these interactions that create entirely new financial primitives impossible in traditional finance.

1.3 The composability thesis

The previous section analyzed the six distinctive properties one by one, using the compression of the trust tax as its framework, but that property-by-property analysis obscured a deeper fact: the disruptive force of the six properties comes not from any single feature but from the multiplicative interactions among them. Schär (2021) [27], in the Federal Reserve Bank of St. Louis Review, was the first to analyze the economic implications of DeFi composability systematically, and Harvey, Ramachandran, and Santoro (2021) [28] extended this framework to institutional investment. Building on these analyses, this section shows both how the multiplicative effect of composability creates financial primitives impossible in traditional finance and how the same mechanism serves as a channel for risk contagion.

1.3.1 The interaction matrix of the six properties

Traditional finance can partially realize any one of the six distinctive properties—the extension of trading hours is being actively explored [21], and trials of T+0 settlement are under way—but it cannot realize all six at once; and it is precisely the simultaneous presence and interaction of all six that create the financial primitives impossible in traditional finance.

Table 1-2 presents, as a 6×6 symmetric matrix, the interaction effects produced by pairwise combinations of the six properties. Each cell in the matrix represents not the simple superposition of two features but the entirely new capability that emerges from their multiplicative interaction; several of these combinations give rise to financial primitives that are, in principle, impossible within the architecture of traditional finance.

	Permissionless	Divisible	Atomic settlement	24/7	Programmable	Transparent
Permissionless	—	Democratized micro-investing	Instant settlement with no entry barrier	Anyone, anywhere, anytime can participate	Anyone can deploy a protocol	Anyone can audit
Divisible		—	Atomic transfer of tiny amounts	Continuous streaming micropay-ments	Programmable precise fractions	Complete record of micro-transactions
Atomic settlement			—	Around-the-clock instant settlement	Flash loans	Fully verifiable settlement process
24/7				—	Continuous operation of the funding rate	Real-time monitoring
Programmable					—	MEV
Transparent						—

Table 1-2. The interaction matrix of the six distinctive properties (Data source: compiled by the author)

The 15 interaction cells above the diagonal display not the individual effect of each of the six features but the emergent results of 15 combinations. Three of these cells—flash loans, the continuous operation of the funding rate, and MEV—represent emergent products of paradigmatic significance: in traditional finance they are not merely unrealized but impossible to realize under that architecture.

1.3.2 Financial primitives impossible in traditional finance

The flash loan is the product of three interacting features: permissionlessness, atomic settlement, and programmability. Within a single atomic transaction, one can borrow, use, and repay any amount of an asset; if repayment fails, the entire transaction automatically reverts, achieving an instantaneous loan with zero collateral and zero risk. The Aave protocol pioneered this mechanism in 2020 [29]. In traditional finance, lending requires credit review (permissioned), incurs settlement delay (non-atomic), and involves manual approval (non-programmable); because none of these three preconditions is met, the flash loan is in principle impossible under that architecture. Flash loans change the economics of arbitrage: whereas arbitrage in traditional finance requires substantial upfront capital, flash loans make zero-collateral arbitrage possible (the real execution costs include gas fees, priority-fee payments in MEV competition, and infrastructure investment). They raise market efficiency while also lowering the cost barrier to attacks, and they are widely used for arbitrage, collateral swaps, and self-liquidation.

MEV arises from the interaction of three features: around-the-clock trading, transparency, and programmability. Because the market never closes, arbitrage opportuni-

ties persist continuously; because all pending transactions are visible to everyone in the public mempool, a program can analyze them within milliseconds and execute ahead of them. In traditional finance, comparable front-running is subject to a triple constraint: strict legal prohibition, non-public information, and exchanges' self-regulation. MEV is the economic embodiment of the transparency paradox. It both guards market efficiency (arbitrage corrects mispricing) and erodes fairness (sandwich attacks harm ordinary users); this dual nature makes the governance of MEV a central topic to which the Ethereum ecosystem continually devotes research resources [25].

Protocol-based arbitrage, exemplified by Ethena, combines three features: around-the-clock trading, atomic settlement, and programmability. Because the market runs continuously, the funding rate can settle automatically every 8 hours, and settlement is atomic; this allows the strategy of shorting perpetual futures while buying spot to earn the funding rate to be packaged into a single smart contract that anyone can join with one click and that manages billions of dollars in assets [30]. In traditional finance, this strategy cannot be turned into a protocol, for three reasons. First, settlement is not atomic: in a traditional arbitrage trade, the spot purchase and the derivative short settle through different clearing channels with a lag of several days, so the two cannot be packaged into a single indivisible operation. Second, rate settlement is not automated: adjusting the rate on a traditional interest rate swap or futures contract requires manual processes and multi-party confirmation. Third, access is not permissionless: a traditional trade of this kind requires an International Swaps and Derivatives Association (ISDA) master agreement, a prime-brokerage relationship, and a minimum position of millions of dollars. In addition, the maturity structures do not match: traditional futures have an expiry date and a basis trade has a definite convergence endpoint, whereas the absence of any expiry date for perpetual futures exposes a funding-rate strategy to the structural difference of an indefinite horizon—which is at once a precondition for turning the strategy into a protocol and a distinctive source of its risk. The simultaneous absence of these preconditions means that turning the strategy into a protocol is, in principle, impossible under the architecture of traditional finance, not merely difficult. Ethena's architecture packages a strategy once exclusive to institutional hedge funds into open, permissionless, protocol-level financial infrastructure. Its risk dimensions cannot be ignored, however: Ethena's hedging positions are concentrated on a handful of centralized exchanges, creating significant counterparty-concentration risk; the funding rate can turn negative for extended periods, so the yield on USDe can reverse into a loss; and USDe faces the risk of being classified as a security. More fundamentally, the degree to which Ethena is a protocol has a clear boundary: its core hedging operations rely on the platform trust of centralized exchanges rather than on pure architectural trust, so in the classification of trust paradigms in Section 1.4.3 it straddles two modes.

1.3.3 The dark side of composability

The multiplicative effect of composability runs both ways: it amplifies innovation and risk alike. When the output of protocol A becomes the input to protocol B and then serves as collateral in protocol C, a single failure in A propagates instantly to B and C, at a speed approaching the block time of the blockchain. Werner et al. (2022) [31], in a systematic review of DeFi, characterize this inter-protocol dependence as the core

structural source of risk contagion in on-chain finance.

The collapse of Terra/LUNA in May 2022 was an extreme demonstration of composability acting as a channel for risk contagion [32]. After UST lost its peg, LUNA entered a negative-feedback loop; several lending protocols that relied on UST as collateral underwent cascading liquidations; the entire DeFi ecosystem was shaken; and roughly $40 billion in market value (the total value of the LUNA and UST ecosystem, of which UST accounted for about $18 billion) evaporated within days. Behavioral-finance factors also played a substantial role here: Anchor Protocol's 19.5% "fixed yield" constituted a classic yield illusion, drawing many risk-unaware retail investors into overconcentrated allocations; and during the collapse, panic narratives spread on social media no slower than liquidations propagated on chain. The dark side of composability is not only technical and structural but also behavioral. This case is the intersection of composability with the settlement-cascade/liquidation-cascade distinction: atomic settlement accelerated the contagion, and composability widened its reach. Especially dangerous here is wrong-way risk: the core problem for protocols that relied on UST as collateral was that the value of the collateral and the positions it secured collapsed in the same direction under stress, a hidden correlation that is systematically underestimated in normal times.

The flash-loan attack on Euler Finance in March 2023 illustrated another contagion pathway. The attacker exploited a logic flaw in the Euler protocol and, through a series of flash-loan transactions, amplified the scale of the attack to extract about $197 million (the funds were later returned by the attacker). Because several DeFi protocols had deposited assets in Euler or used it as a source of yield, the impact of the attack extended far beyond Euler itself: protocols that depended on Euler, including Balancer, Angle Protocol, and Idle Finance, suffered losses simultaneously. Inter-protocol dependence in DeFi exhibits a highly concentrated hub-and-spoke topology in which a few core protocols anchor many dependency chains, so that an attack on a hub has a disproportionate systemic impact. The case is another clear instance of composability risk: inter-protocol dependence spreads the impact of a single vulnerability along the dependency chain to the entire protocol network.

Composability drives the marginal cost of financial innovation toward zero, but it also greatly accelerates the contagion of financial risk: intra-chain contagion can complete within a single transaction, or even a single block—faster than one block interval (in the Euler attack, the attacker extracted about $197 million within a few blocks)—whereas cross-chain contagion, constrained by bridging and oracle latency, spreads comparatively slowly. Innovation and risk contagion use the same set of inter-protocol call mechanisms, and this constitutes the core structural contradiction of on-chain finance. Protocol design in on-chain finance should not pursue maximum composability; it must strike a balance between the efficiency gains of composability and the contagion of risk—a subject taken up in depth in Chapter 20, on the endogenous fragility of liquidity, and Chapter 29, on the design of the clearing and settlement architecture.

> The core contradiction of composability: innovation and risk contagion share the same set of inter-protocol call mechanisms. This contradiction cannot be eliminated, only managed.

1.4 Microstructure reconstruction and the transformation of the risk surface

The transformations set in motion by the six distinctive properties and their composability ultimately converge into a systematic reconstruction of market microstructure and a fundamental transformation of the statistical properties of risk. Using the classic framework of Harris (2003) [4], this section summarizes the reconstruction of three elements—participants, processes, and information—then argues why the risk surfaces of the two financial systems require fundamentally different risk-management methodologies, and finally proposes a layered, hybrid framework of three trust paradigms.

1.4.1 The reconstruction of the three microstructure elements

The reconstruction of these three elements—participants, processes, and information—by digital assets has recurred throughout the analysis of the six distinctive properties in Sections 1.2.1 through 1.2.6; here it is consolidated within the Harris framework.

The reconstruction at the level of participants is the broadest in scope. Roles change from fixed identities defined by licenses to fluid states attached to wallet addresses: a single address can act as trader, market maker, arbitrageur, and liquidator on the same day. Automated economic agents—MEV searchers and liquidation bots, for example—become significant new market participants, and the traditional microstructure dichotomy between informed and uninformed traders no longer fully applies. The competitive structure flattens from a tiered, franchised hierarchy into open competition, but this can also lead to rent dissipation: competitors expend substantial resources contesting a fixed pool of MEV rents, driving the net social benefit toward zero. The PBS architecture is essentially a way to reduce this dissipation by converting disorderly competition into an auction mechanism. Chapter 4 develops a detailed model of seven categories of participants.

At the level of processes, once atomic settlement (Section 1.2.3) eliminates the T+N window of risk exposure, the focus of risk management shifts from assessing counterparty credit to auditing code security [19]. Smart contracts (Section 1.2.5) encode intermediary functions, structurally reducing operating costs. The atomicity of settlement guarantees the composability of financial operations, and the mode of innovation changes from designing new products to combining existing protocols.

At the level of information, information asymmetry shifts, as described in Section 1.2.6, from the right of access to processing capability. Tracking smart-money addresses and imitating their trades has become a common strategy, but it may intensify herding. The basis of trust changes from institutional brands to open-source code and verifiable data. The concept of privacy is redefined, and techniques such as zero-knowledge proofs are seeking a balance between radical transparency and necessary privacy [26].

The reconstruction of these three elements is the convergent expression, at the microstructure level, of the six distinctive properties. Building on this, two further questions arise: how these changes have altered the statistical properties of risk, and how the sources of trust should be reorganized.

1.4.2 The transformation of the risk surface

The risks of traditional finance and of digital assets differ not only in type but also fundamentally in their statistical properties, and therefore cannot be addressed with the same risk-management methodology. In his seminal paper, Mandelbrot (1963) [33] first demonstrated that the distribution of financial returns does not follow a normal distribution but instead exhibits fat tails—that is, extreme events occur with a probability far higher than the normal assumption predicts. Taleb (2007) [34] further argued that humans systematically underestimate the probability and impact of extreme events, and that traditional risk-control frameworks, designed under the normality assumption, fail systematically in a fat-tailed world. The risk surface of digital assets is an extreme expression of the fat-tailed world that Mandelbrot and Taleb described.

Table 1-3 compares the risk surfaces of the two financial systems along five dimensions: frequency, shock distribution, visibility, recovery speed, and risk-control tools. Each of the five dimensions points to the same conclusion: the risks of the two systems differ not in degree but in kind, and no single risk-management methodology can address both.

Dimension	The risk surface of traditional finance	The risk surface of digital assets
Frequency	Low—major crises occur on average once every 10 to 15 years [35]	High—protocol vulnerabilities and attacks occur weekly
Shock distribution	High-impact—once it occurs, it is systemic	Variable impact—most events have limited effect, but the tail is extremely fat
Visibility	Often invisible before a crisis, attributed only afterward	Many risks are visible on chain, but visible does not mean manageable
Recovery speed	Slow—institutional repair is measured in years	Fast—protocol repair is measured in days to weeks, but it may also be irreversible
Risk-control tools	VaR, stress testing, the Basel framework—based on the low-frequency, high-impact assumption	Require a new methodology that handles high-frequency small events and rare catastrophic events at once

Table 1-3. Comparison of the risk surfaces of traditional finance and digital assets (Data source: compiled by the author based on [35])

Figure 1-4 contrasts this difference in statistical properties with two probability-distribution curves: the tail probability of a fat-tailed distribution decays far more slowly than that of the normal (thin-tailed) distribution, and extreme events several standard deviations from the mean occur far more frequently than the normal assumption predicts. This is precisely why traditional value at risk (VaR) models fail systematically in on-chain markets.

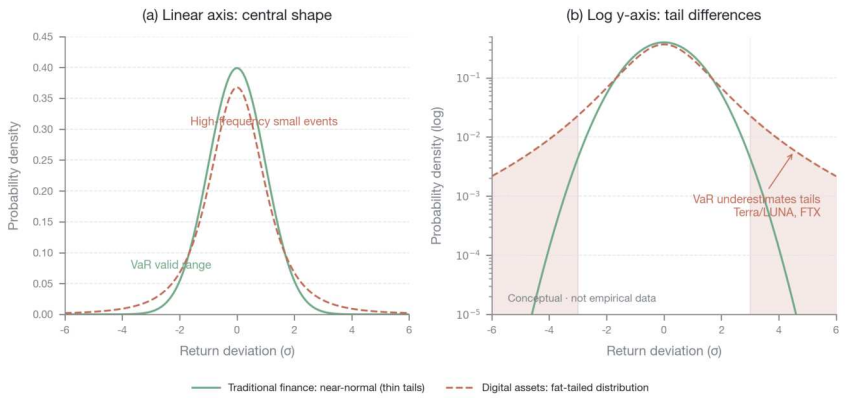


Figure 1-4. Comparison of the risk distributions of traditional finance and digital assets—the difference in tail probability between the normal (thin-tailed) distribution and the fat-tailed distribution (conceptual/theoretical illustration; the fat tail is modeled with a Student's t-distribution with 3 degrees of freedom; the Terra/LUNA and FTX labels in the figure are illustrative examples of tail events, not measured data points; the right panel uses a logarithmic vertical axis to highlight the tail difference; theoretical basis: Mandelbrot 1963 [33], Taleb 2007 [34])

To ask whether digital assets are more dangerous or safer is the wrong question. The right question is how the risk surfaces of the two systems differ and what different tools are needed to manage them. The risk surface of traditional finance is a distribution of low-frequency but systemic shocks. Reinhart and Rogoff's (2009) [35] systematic study of eight centuries of financial crises shows that banking crises recur on average every 10 to 15 years but are highly clustered and transmit across borders—and that traditional finance is not without high-frequency operational-risk events either (the 2010 U.S. stock-market flash crash, for instance, erased trillions of dollars in market value within minutes). The risk surface of digital assets is a distribution that is high-frequency and highly variable in the scale of its shocks: protocol vulnerabilities occur weekly, most with limited impact, but occasionally catastrophic. More striking still, the two risk surfaces are not independent: with the approval of Bitcoin exchange-traded funds (ETFs) and the entry of institutional investors, the correlation between digital assets and traditional finance rises markedly in periods of stress, and institutions that hold both asset classes become contagion channels when they deleverage, so the two are beginning to show signs of mutual contagion. This recognition is a prerequisite for understanding the risk analysis in Chapters 20 through 24.

The transformation of the risk surface also means that existing financial-regulatory frameworks such as Basel III—whose core parameter assumptions, such as the 10-day holding period in VaR and the 30-day stress period in the liquidity coverage ratio, are designed for the risk surface of traditional finance—may fail in two directions: overreacting to high-frequency small events (because the traditional framework treats any event as a potential systemic risk) while at the same time being underprepared for genuine tail risk (because the traditional framework's stress-test scenarios may not cover the extreme situations specific to on-chain markets, such as protocol-level vulnerabilities and the superposition of composability contagion and liquidation cascades). Investors accustomed to traditional finance may misread the high-frequency volatility of on-chain

markets as higher risk. Prospect theory shows that the psychological pain of frequent small losses can exceed their economic impact, prompting premature exit, while their underestimation of the probability of rare large losses leaves them underprepared for genuine tail risk. Systemic events on the scale of Terra/LUNA and FTX require a fundamentally different mental model and risk budget.

While compressing the traditional trust tax, digital assets introduce an entirely new cost structure. The security tax is the cost of smart-contract security audits and formal verification. As noted in Section 1.2.5, the one-time audit fee (tens of thousands to hundreds of thousands of dollars) is only the starting point: ongoing expenditures include bug-bounty programs (ranging from hundreds of thousands to millions of dollars), re-audits after contract upgrades, the operation of real-time monitoring systems, and the capital tied up in an insurance fund—and even after all of this, zero vulnerabilities cannot be guaranteed. The gas tax is the execution cost of on-chain transactions, which can spike tens of times over during network congestion and render small transactions economically infeasible. The private-key management tax is the cognitive burden and operational risk users bear in managing their own private keys: losing a private key means permanent loss of the assets, with no forgot-password recovery mechanism. Based on an analysis of dormant on-chain addresses, Chainalysis (in a 2020 report) estimated that about 3.7 million bitcoin—roughly 20% of the supply at the time—were permanently inaccessible because of lost private keys (this estimate rests on a dormancy heuristic and is somewhat contested, since it may conflate long-term holders with genuinely lost coins, though its order of magnitude is widely cited in the industry) [36]. The smart-contract vulnerability tax comprises the asset losses caused by exploited code flaws—irreversible and with no legal recourse [24]. The MEV-extraction tax is the hidden cost that competitive extraction of transaction-ordering rights imposes on ordinary users [25]. The governance-uncertainty tax shows up as low participation in DAO voting and the risk of governance attacks. Digital assets do not eliminate the trust tax; they replace an old trust tax with a new one. The standard for judging this paradigm shift is not whether the new trust tax is zero but whether its total is significantly lower than the old one and whether its distribution is fairer.

1.4.3 Three trust paradigms

Moving beyond the binary opposition of centralization and decentralization, trust can be organized into three paradigms. The concept of architectural trust builds on Werbach's (2018) [37] systematic analysis of blockchain trust mechanisms. Werbach distinguishes three traditional modes—peer-to-peer trust, Leviathan trust (state coercion), and intermediary trust—and argues that the blockchain creates a new trust architecture, one that establishes trust through technical design rather than interpersonal relationships or institutional authority. This chapter develops that analysis further into a threefold division of institutional trust, platform trust, and architectural trust; the introduction of platform trust reflects the key role that centralized platforms (such as FTX) play in practice in the digital-asset domain, a category not yet fully analyzed in Werbach's original framework.

Institutional trust draws on law, regulation, and specialized institutions; it offers a 200-

year track record, legal recourse, and consumer protection, but it is costly, opaque, slow, and difficult to enter, and the LIBOR scandal [10] proved that it can be manipulated by insiders. Platform trust draws on the operator of an exchange or platform; it offers high performance and a good user experience but depends entirely on a single entity, and the collapse of FTX [3] demonstrated the catastrophic consequences of that dependence—platform trust is essentially a repackaging of traditional centralized trust in the digital domain. Architectural trust draws on code, cryptography, and decentralized consensus; it is verifiable, globally accessible, and independent of any single entity, but it faces challenges such as code vulnerabilities, limited performance, and governance difficulties.

The three paradigms are not substitutes for one another; they can be mixed across layers. At the asset-custody layer, the self-custody approach of architectural trust eliminates, through private-key control, the risk of assets being misappropriated by a single centralized entity, but self-custody introduces a new spectrum of operational risks—key management, physical security, and disaster recovery. At the trade-execution layer, the centralized matching engine of platform trust offers an irreplaceable performance advantage in high-frequency settings, whereas the decentralized exchange of architectural trust is more competitive in transparency and censorship resistance; the optimal choice depends on the specific use case. Notably, the most successful on-chain order-book DEXs today (such as Hyperliquid and dYdX v4) adopt a hybrid architecture, in which centralization at the ordering and matching level coexists with on-chain settlement; this engineering compromise should not be classified simply as one paradigm or the other but as a pragmatic combination of the two. At the clearing and settlement layer, the atomic settlement of architectural trust eliminates the settlement cascade but cannot prevent the liquidation cascade, and extreme circumstances may still require the discretion that institutional trust provides. History already offers precedents: in the JELLY incident of March 2025, Hyperliquid's validator committee voted to delist the token and liquidate positions at a manually set price, and after The DAO incident of 2016 the Ethereum community reversed transactions through a hard fork—both instances in which architectural trust, under extreme stress, degenerated into discretionary decisions by a small number of people. At the data-verification layer, the on-chain transparency of architectural trust eliminates the possibility of LIBOR-style manipulation, but oracles (which bring off-chain data on chain) still require a degree of trust. At the rule-governance layer, the ultimate discretion and final arbitration that the legal and judicial systems of institutional trust provide in extreme cases remain, for now, irreplaceable; the question of who has the authority to modify the code still requires human judgment, and low participation in DAO governance and the risk of governance attacks together cast doubt on the reliability of pure architectural trust at this layer. In 2022, the Office of Foreign Assets Control (OFAC) sanctioned Tornado Cash, demonstrating that state power can intervene directly in protocol-level permissionlessness (a 2024 ruling of the Fifth Circuit later overturned the sanctions, and OFAC removed the entity from the Specially Designated Nationals (SDN) list in March 2025—a reminder that the exercise of state power is itself subject to judicial checks). The freezing orders that courts in several countries have developed in crypto-asset recovery cases, in turn, demonstrate that legal recourse is extending onto the chain.

The optimal financial infrastructure is neither fully decentralized nor fully centralized;

it is a layered architecture that makes the most reasonable choice of trust paradigm at each layer. This insight is systematized into concrete engineering choices in Chapter 5's spectrum of verifiability and Chapter 28's design of the trust stack. Combining the three paradigms with the trust-tax framework closes the loop: institutional trust carries the highest trust tax but the most comprehensive service; platform trust carries a moderate trust tax but an extremely high concentration risk; and architectural trust carries the lowest trust tax but incomplete service coverage. The optimal combination selects, at each layer, the paradigm with the lowest trust tax that still provides an adequate level of service—an engineering optimization problem, not an ideological stance.

1.5 Chapter summary

This chapter has assembled a complete analytical toolkit for understanding digital assets—the largest-scale natural experiment in financial history. The five layers of trust reveal the internal structure of traditional finance's trust architecture; each layer has a reason to exist, and each has failed at some point. The trust tax unifies the entire institutional cost of traditional finance into a causally explicable whole whose root cause is the absence of trust; the architecture of digital assets compresses this tax through technical means, but compression is not free, and the trust tax has its rationale. The composability thesis shows how the multiplicative interactions among the six distinctive properties create financial primitives impossible in traditional finance—flash loans, MEV, and protocol-based arbitrage—while composability is at the same time a channel for risk contagion. The transparency paradox establishes that on-chain transparency does not equal fairness: the dimension of competition shifts from who possesses information to who can process it fastest. The transformation of the risk surface demonstrates that the statistical properties of risk in the two financial systems are fundamentally different—not a question of more dangerous or safer, but one that requires different risk-management methodologies. The three trust paradigms provide a framework for choosing among different sources of trust, mixed across different functional layers, and the optimal solution is a layered, hybrid engineering problem.

These six concepts form a logical chain—from root-cause analysis (the trust tax), to technical response (the six distinctive properties), to emergent effects (the composability thesis and the transparency paradox), to risk reconstruction (the transformation of the risk surface), and finally to the reorganization of trust (the three trust paradigms)—and this chain is the starting point for the analysis developed in the chapters that follow.

The most far-reaching product of this experiment is perpetual futures—with an annual notional trading volume of about $86.2 trillion (on a centralized-exchange basis; about $92.9 trillion including decentralized exchanges) and accounting for more than 90% of crypto-derivatives trading volume [38]. Crypto-derivatives volume data, however, suffer from wash trading, so actual volume may be lower than the reported figures, and notional volume is inflated by high leverage; traditional exchange-traded derivatives lack a uniform annual notional-principal measure (the Futures Industry Association (FIA) counts by number of contracts—about 180 billion worldwide in 2024), making the two difficult to compare directly. Perpetual futures are the subject of the next chapter. The emergence of perpetual futures was no accident: it is the concentrated expression, in the

derivatives domain, of the composability of the six distinctive properties. In a market that trades continuously around the clock and settles in real time, the fixed-expiry design of traditional futures comes into structural conflict with the market's continuity. The theoretical conception of perpetual futures was proposed by Shiller (1993) [39], and BitMEX first engineered it in crypto markets in 2016. By abolishing the expiry date and introducing the funding-rate mechanism—which functionally replaces the price-convergence role of the expiry date—perpetual futures became a derivative form highly consonant with the continuous structure of this market, and they are the central object of the analysis that follows in this book.

References

[1] CoinShares. (2025). *Institutional report: 13F Bitcoin ETF data Q3 2025*. CoinShares. https://coinshares.com/us/insights/research-data/13f-filings-of-bitcoin-etfs-q3-2025-institutional-report/

[2] Alternative Investment Management Association & PwC. (2025). *7th annual global crypto hedge fund report*. AIMA. https://www.aima.org/article/press-release-crypto-friendly-regulatory-changes-accelerate-institutional-investment.html

[3] U.S. Securities and Exchange Commission. (2022). *SEC charges Samuel Bankman-Fried with defrauding investors in crypto asset trading platform FTX* [Press release]. https://www.sec.gov/news/press-release/2022-219

[4] Harris, L. (2003). *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press.

[5] Depository Trust & Clearing Corporation. (2024). *Accelerating the U.S. securities settlement cycle to T+1*. DTCC. https://www.dtcc.com/ust1

[6] Financial Crisis Inquiry Commission. (2011). *The financial crisis inquiry report: Final report of the National Commission on the Causes of the Financial and Economic Crisis in the United States*. U.S. Government Publishing Office.

[7] Luhmann, N. (1979). *Trust and Power*. Wiley.

[8] Commodity Futures Trading Commission. (2013). *CFTC orders MF Global Inc. to pay more than $1 billion penalty for unlawful misuse of customer funds* [Press release]. https://www.cftc.gov/PressRoom/PressReleases/pr6776-13

[9] U.S. Securities and Exchange Commission. (2016). *Barclays Capital Inc. charged with violations in handling of dark pool* [Press release]. https://www.sec.gov/news/pressrelease/2016-16.html

[10] Hou, D., & Skeie, D. (2014). LIBOR: Origins, economics, crisis, scandal, and reform. *Federal Reserve Bank of New York Staff Reports*, No. 667. https://www.newyorkfed.org/research/staff_reports/sr667

[11] Williamson, O. E. (1985). *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. Free Press.

[12] Burton-Taylor International Consulting. (2024). *Global market data revenue report 2024*. Burton-Taylor.

[13] World Bank. (2024). *Remittance prices worldwide: An analysis of trends in the cost of remittance services* (Issue 49). World Bank Group. https://remittanceprices.worldbank.org/

[14] Lessig, L. (2000). Code is law: On liberty in cyberspace. *Harvard Magazine*. https://www.harvardmagazine.com/2000/01/code-is-law-html

[15] Baumol, W. J. (1982). Contestable markets: An uprising in the theory of industry structure. *American Economic Review*, *72*(1), 1–15.

[16] Financial Action Task Force. (2021). *Updated Guidance for a Risk-Based Approach to Virtual Assets and Virtual Asset Service Providers*. FATF. https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Guidance-rba-virtual-assets-2021.html

[17] CertiK. (2025). *Hack3d: The Web3 security report 2025*. CertiK. https://www.certik.com/resources/blog/hack3d-the-web3-security-report-2025

[18] Boston Consulting Group & Aptos Labs. (2024). *Relevance of on-chain asset tokenization in crypto winter*. BCG. https://web-assets.bcg.com/1e/a2/5b5f2b7e42dfad2cb3113a291222/on-chain-asset-tokenization.pdf

[19] Bank for International Settlements. (2025). *Annual economic report 2025*. BIS. https://www.bis.org/publ/arpdf/ar2025e.htm

[20] Baur, D. G., Cahill, D., Godfrey, K., & Liu, Z. F. (2019). Bitcoin time-of-day, day-of-week and month-of-year effects in returns and trading volume. *Finance Research Letters*, *31*, 78–92. https://doi.org/10.1016/j.frl.2019.04.023

[21] Depository Trust & Clearing Corporation & Ernst & Young. (2025). *Moving toward 24/5 and beyond: Extending US securities settlement hours*. DTCC.

[22] Cohen, S., Brennan, M., De Jesus, H., Warr, J., Merwin, S., Nguyen, N., Cronin, R., & Shen, A. (2025). *Market spotlight: 24 hour trading*. BlackRock. https://www.blackrock.com/corporate/literature/whitepaper/blackrock-market-spotlight-24-hour-trading.pdf

[23] Brennecke, M., Guggenberger, T., Schellinger, B., & Urbach, N. (2022). The de-central bank in decentralized finance: A case study of MakerDAO. *Proceedings of the 55th Hawaii International Conference on System Sciences*, 6370–6379. https://doi.org/10.24251/hicss.2022.737

[24] Immunefi. (2024). *Crypto losses in 2023*. Immunefi. https://immunefi.com/research/

[25] European Securities and Markets Authority. (2025). *Maximal extractable value: Implications for crypto markets* (TRV Risk Analysis). ESMA. https://www.esma.europa.eu/sites/default/files/2025-07/ESMA50-481369926-29744_Maximal_Extractable_Value_Implications_for_crypto_markets.pdf

[26] Ethereum Foundation Privacy and Scaling Explorations. (2025). *End-to-end privacy roadmap for Ethereum*. https://pse.dev/

[27] Schär, F. (2021). Decentralized Finance: On Blockchain- and Smart Contract-Based Financial Markets. *Federal Reserve Bank of St. Louis Review*, *103*(2), 153–174. https://doi.org/10.20955/r.103.153-74

[28] Harvey, C. R., Ramachandran, A., & Santoro, J. (2021). *DeFi and the Future of Finance*. Wiley.

[29] Aave. (2020). *Aave Protocol V2: Flash loans*. Aave. https://docs.aave.com/developers/guides/flash-loans

[30] Ethena Labs. (2024). *Ethena: Internet native money*. Ethena. https://ethena-labs.gitbook.io/ethena-labs

[31] Werner, S. M., Perez, D., Gudgeon, L., Klages-Mundt, A., Harz, D., & Knottenbelt, W. J. (2022). SoK: Decentralized Finance (DeFi). In *Proceedings of the 4th ACM Conference on Advances in Financial Technologies* (pp. 1–16). ACM. https://doi.org/10.1145/3558535.3559780

[32] Liu, J., Makarov, I., & Schoar, A. (2023). Anatomy of a run: The Terra Luna crash. *NBER Working Paper*, No. 31160. https://www.nber.org/papers/w31160 https://doi.org/10.2139/ssrn.4416677

[33] Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business*, *36*(4), 394–419. https://doi.org/10.1086/294632

[34] Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. Random House.

[35] Reinhart, C. M., & Rogoff, K. S. (2009). *This Time Is Different: Eight Centuries of Financial Folly*. Princeton University Press.

[36] Chainalysis. (2020). *60% of Bitcoin is held long term as digital gold, while just 19% is used for trading*. Chainalysis. https://www.chainalysis.com/blog/60-of-bitcoin-is-held-long-term-as-digital-gold-while-just-19-is-used-for-trading/

[37] Werbach, K. (2018). *The Blockchain and the New Architecture of Trust*. MIT Press. https://doi.org/10.7551/mitpress/11449.001.0001

[38] CoinGecko. (2026). *2025 annual crypto industry report*. CoinGecko. https://www.coingecko.com/research/publications/2025-annual-crypto-report

[39] Shiller, R. J. (1993). *Macro Markets: Creating Institutions for Managing Society's Largest Economic Risks*. Oxford University Press.

Derivatives Markets and the Rise of Perpetual Futures

By CoinGecko's top-10 perpetual-exchange measure, crypto perpetual futures trading volume reached $92.9 trillion in 2025, of which centralized exchanges accounted for roughly $86.2 trillion and decentralized exchanges (DEXs) for roughly $6.7 trillion. Decentralized perpetual volume grew by roughly 3.5 times over the prior year's roughly $1.5 trillion, reaching approximately 4.5 times that level (about +347%) [1] [2]. A product that BitMEX did not formally launch until May 13, 2016 grew from nothing, in less than a decade, into one of the most important pricing and leverage infrastructures in the digital-asset market. Early BitMEX data show that derivatives markets lead BTC price discovery [3], and later research indicates that the regulated CME bitcoin futures market performs an important price-discovery function [4]. Together, this evidence shows that digital-asset price discovery has migrated substantially toward highly liquid derivatives markets, but it does not reduce to the claim that all spot prices mechanically track all perpetual futures (see Section 2.4.1).

This raises two questions: what problem does the design of perpetual futures solve, and what new cost does it introduce? Chapter 1 argued that the institutional costs of traditional finance can be unified as a trust tax, and that the six distinctive properties of digital assets compress this tax through multiplicative interaction. It closed by identifying perpetual futures as the composability product of these six properties in the domain of derivatives: the combination of 24/7 continuous trading, programmable rules, and atomic settlement creates a structural conflict between the fixed-expiry design of traditional futures and the continuous nature of digital assets. By abolishing the expiry date and introducing a funding rate mechanism, perpetual futures became the most natural form of derivative in this new environment. But being the *most natural* form does not make it *costless*. This chapter's central position is that perpetual futures are not an upgraded version of traditional futures but a costly design trade-off; only by understanding what they gain and what they lose can one understand this market.

This chapter retells the 400-year history of derivatives as three leaps in trust technology: personalized trust, institutionalized trust, and algorithmic trust. On that basis, it identifies three structural conflicts between traditional futures and digital assets—temporal mismatch, liquidity fragmentation, and trust mismatch—and shows how the two core design innovations of perpetual futures, together with their market outcome, address each conflict directly. It then analyzes the structural costs that accompany

these innovations and uses a leverage-cycle model to reveal the market's endogenous fragility. These frameworks are prerequisites for the book's later analysis: the multiple economic natures of the funding rate, the endogenous logic of the liquidation cascade, and the migration of the trust paradigm in decentralized perpetual futures.

2.1 The trust evolution of derivatives

The 400-year evolution of derivatives can be retold as three leaps in trust technology. Forward contracts rely on bilateral, personalized trust; futures contracts achieve institutionalized trust through standardization and the central counterparty (CCP); and perpetual futures achieve algorithmic trust through predefined rules and economic incentives. This section examines the trust logic and limitations of forwards, futures, and options in turn, distills a unified functional framework, and shows how perpetual futures complete the third leap in derivatives design.

2.1.1 Personalized trust and forward contracts

The forward contract is the logical starting point for all financial derivatives. Its definition is straightforward: a customized over-the-counter (OTC) agreement that binds two parties to buy or sell an underlying asset on a specified future date at a price agreed today. The wheat-forward trades between farmers and flour mills in the nineteenth-century American Midwest illustrate the basic form of intertemporal risk management: the farmer locks in the autumn selling price, the mill locks in its raw-material cost, and each sheds its exposure to future price uncertainty. Hull (2022) [5] notes that this risk-transfer function is the economic foundation of all derivatives.

The core dilemma of the forward contract is counterparty credit risk. If the autumn market price of wheat is far above the agreed contract price, the farmer has a strong incentive to default; if it is far below, the same is true of the mill. Without a neutral and powerful guarantee of performance, the forward market's size is limited by its radius of trust: it can operate only within a circle of acquaintances bound by social ties [5]. Illiquidity compounds the problem. Every forward contract is customized and can barely be transferred to a third party; once a position is established, the only exits are performance and default. The forward market is essentially a network of countless isolated bilateral credit relationships rather than a single, liquid market.

Viewed through the trust-layer framework of Chapter 1, the fundamental problem of the forward contract is that it has no independent trust-function layer: no independent custody, no independent clearing, and no independent data verification. All trust functions are folded into a single bilateral relationship, so if one party's credit collapses, the entire contract collapses. The forward contract therefore pays a very high tax for trust, in the form of a constrained market size and counterparty risk; every potential transaction beyond the radius of trust goes unconsummated because the cost of trust is too high.

2.1.2 Institutionalized trust and the futures revolution

Futures contracts answered the trust bottleneck of the forward through two core institutional innovations. Standardization converted the customization of the forward into uniform specifications—contract size, grade of the underlying, delivery date, and delivery location were all fixed—creating fungibility: any contract for a given month is identical to every other contract for that month. This fungibility lays the foundation for free transfer and markedly improves liquidity. The founding of the Chicago Board of Trade in 1848 marked the institutional starting point of this model [5]. The CCP interposes itself in every trade through novation, becoming the seller to every buyer and the buyer to every seller; it consolidates countless dispersed bilateral credit relationships into concentrated trust in a single central institution, fundamentally resolving counterparty risk.

The CCP's risk-management architecture centers on a loss-sharing waterfall: strict membership admission ensures that only well-capitalized institutions can participate; the margin system requires clearing members to post initial and maintenance margin and to submit to daily mark-to-market; and the default waterfall draws in turn on the defaulter's margin, the CCP's own capital, and the guarantee fund contributed by non-defaulting members. This architecture converts the problem of *trusting a person* into the problem of *trusting an institution*. The CCP's credit rating far exceeds that of any individual trader, but the CCP itself becomes a systemic node. *Concentrating risk in order to manage risk* is the inherent contradiction of the CCP model—a point Chapter 1 developed in its analysis of the chain reaction triggered by Lehman Brothers' collapse.

Viewed through Chapter 1's five-layer trust-function lens, futures contracts realize, through specialized institutions, the trust functions that forward contracts lacked: the exchange provides fair execution, the CCP provides clearing, custodian banks hold margin, and regulators impose rule-based constraints. Institutionalizing trust extended the market's reach from a circle of acquaintances to strangers worldwide. This extension raises the trust tax, because each independent trust-function layer needs a specialized institution to run it, with its own operating costs and compliance requirements; yet the liquidity and risk-management efficiency it releases far exceed that incremental tax. The inherent contradiction of the CCP model—whether technological means can inherit the CCP's risk-management function while avoiding its centralization risk—sets up the third leap.

2.1.3 Options and nonlinear payoffs

Options moved beyond the linear payoff of futures, granting the holder a right rather than an obligation and creating an asymmetric risk structure. The buyer of a call profits when the underlying rises and loses only the premium when it falls, combining limited downside with unlimited upside. The buyer of a put, conversely, gains protection when the underlying falls, retaining upside while hedging downside risk.

Black and Scholes (1973) [6] solved the central problem of option pricing, showing that an option's price is determined by five factors: the price of the underlying, the strike price, the risk-free rate, the time to expiry, and volatility. Their core insight is that the

higher the volatility, the greater the option's value. In the highly volatile digital-asset market, this relationship implies that options should, in theory, command a relatively high premium. The empirical outcome is the opposite: options have not become the dominant derivative in the digital-asset market, and crypto option trading volume in 2025 was far below that of perpetual futures [1]. This *thing that did not happen* demands explanation just as much as the pronounced dominance of perpetual futures.

From the standpoint of trust technology, the innovation of options lies in the payoff structure, not in the trust mechanism. Options rest on the same institutionalized-trust foundation as futures: exchange matching, CCP clearing, and margin management. The options clearing corporation—epitomized by the Options Clearing Corporation (OCC) in the United States—is structurally isomorphic to the futures CCP. Options therefore extend derivatives within the payoff space rather than mark a new leap in the trust paradigm. This distinction helps explain why options have failed to displace perpetual futures in the crypto market: options solve a payoff-structure problem, whereas perpetual futures solve a trust-paradigm problem, and the latter is the more fundamental challenge facing the digital-asset market.

Preliminary clues point in three directions. Although the nonlinear payoff of options is theoretically superior, for most crypto-market participants, who engage primarily in directional speculation, the linear payoff of futures is more intuitive and easier to grasp. Behavioral finance offers a deeper explanation: mental accounting leads traders to perceive an option premium as a *certain loss*, while leveraged margin is filed as a *temporary loss that may yet be recovered*, and this framing systematically favors futures over options. Market-making in options is also far more complex than in linear contracts, because market makers must manage five risk dimensions at once (delta, gamma, theta, vega, and rho); under the crypto market's limited liquidity, this complexity widens quoted spreads and thins depth. Most fundamentally, one core input to option pricing—the time to expiry—runs into structural difficulty in a product designed without an expiry date, a point analyzed in Section 2.2.4.

2.1.4 A unified functional framework

All derivatives, whether the structurally simple forward or the intricately designed option, serve four core economic functions. Risk transfer moves risk from those unwilling to bear it to those willing to bear it, allocating risk more efficiently across society. Price discovery, through traders' buying and selling, incorporates information and expectations into derivative prices, giving the whole economy forward-looking signals. The pioneering study by Garbade and Silber (1983) [7] provided a classic empirical framework for price discovery in futures markets, and its methodology has strongly shaped later research on price discovery in digital-asset markets. Capital efficiency, through the margin system, lets traders control positions of relatively large value with little capital. Liquidity provision, through standardization and centralized trading, attracts many traders and market makers, so participants can enter and exit at any time and at relatively low cost.

Through institutionalized trust, traditional derivatives have served global capital markets for more than a century, with exchanges such as CME, ICE, and Eurex processing

tens of trillions of dollars in contract volume each year. But this system rests on a set of basic assumptions about trading hours, jurisdiction, and trusted intermediaries. When those assumptions collide with the native characteristics of digital assets, structural conflicts become unavoidable. Section 2.2 develops these conflicts systematically.

2.1.5 Algorithmic trust and perpetual futures

The path from theory to practice passed through three milestones. Shiller (1993) [8] proposed a perpetual, delivery-free futures contract that would adjust the deviation between the contract price and the underlying index through an overnight financing rate. In traditional finance, the idea sat unused, overlooked by academics for more than two decades for lack of an application. In 2011, developers implemented a preliminary inverse bitcoin futures contract on the ICBIT exchange, demonstrating technical feasibility; limited in scale and influence, it nonetheless carried the idea from theory to a working prototype. On May 13, 2016, BitMEX formally launched the XBTUSD perpetual futures contract [9], turning the concept into market reality. Two core design innovations—abolishing the expiry-and-delivery date and anchoring the price through a funding rate—together with their market outcome, the concentration of liquidity in a single contract, established BitMEX as the paradigm for digital-asset derivatives.

The trust mechanism of perpetual futures no longer rests on the personalized trust of a circle of acquaintances, and it also reduces reliance on the institutionalized trust of the traditional CCP; instead, it hands part of risk control to predefined algorithmic rules and economic incentives. Through internal transfer payments among traders, the funding rate achieves price anchoring through a market-based negative feedback loop. The two settings differ. On a centralized exchange (CEX), anchoring still relies on platform matching, the mark price, the margin system, the insurance fund, and auto-deleveraging (ADL) rules; only in the DEX setting does it approach an architectural trust jointly constituted by code, oracles, sequencing, and governance. How these two innovations and their market outcome address the structural conflicts between traditional futures and digital assets is argued in Sections 2.2 and 2.3.

2.2 The three structural conflicts

The institutional design of traditional futures embeds three assumptions: discrete trading hours, geographically defined jurisdiction, and centralized trusted intermediaries. The native characteristics of digital assets (continuous global trading 24/7, the absence of jurisdictional boundaries, and code as the basis of trust) conflict with these assumptions not as technical friction but as a fundamental incompatibility between two market paradigms. Transplanting traditional futures directly into the digital-asset world therefore entails structural friction, and the market needs a derivative that fits digital assets more closely. Perpetual futures are one of the principal responses to this need; they are not an incremental improvement on traditional futures but a fundamental redesign built for the native character of digital assets—round-the-clock, globalized flow. The argument proceeds along three dimensions: temporal mismatch, liquidity fragmentation, and trust mismatch.

2.2.1 Temporal mismatch

One of the core assumptions of traditional futures is a fixed, discrete expiry date: the contract has a defined life cycle and is delivered or rolled over before expiry. The digital-asset market runs around the clock and year-round, with no open and no close, no weekends, and no holidays, and information must be priced continuously. This conflict generates friction along three dimensions.

The first friction concerns the expiry date. Long-term holders must roll over repeatedly—closing the contract about to expire and opening a new, more distant one—and so bear double the transaction costs and basis risk. In the highly volatile crypto market, basis fluctuations during the roll window can wipe out weeks of trading profit. The second concerns trading hours. Even when traditional futures extend trading through electronic sessions, they remain constrained by weekends, holidays, daily settlement, and the contract life cycle, whereas the digital-asset market generates information continuously, day and night. When information accumulates during a market closure or a low-liquidity window but cannot be fully priced, the result can be price gaps and interruptions in risk management. The third concerns informational continuity. Traditional futures struggle to provide fully continuous price discovery for digital assets, which generate information around the clock.

The essence of this conflict is that the temporal structure of traditional futures was designed for traditional assets. The information flow and trading activity of traditional assets follow a natural daily and weekly cycle, and digital assets break that cycle. Forcing a continuous information flow into a discrete contract structure necessarily produces friction. Roll costs deserve careful economic accounting here. The concentration of trading volume in the front-month contract is a general phenomenon in futures markets—whether for crude oil or Treasury futures, the front-month contract typically accounts for the vast majority of volume—so the concentration of volume in CME bitcoin front-month futures [10] is not itself specific to digital assets. The more direct mechanism is this: perpetual futures remove the choice of maturity from the product structure, so traders need not choose among front-month, quarterly, and back-month contracts, nor roll over repeatedly to hold a continuous position. This is why a single contract that never has to be rolled naturally attracts more concentrated liquidity.

2.2.2 Liquidity fragmentation

Traditional futures meet demand across different maturities through multiple monthly contracts (front-month, back-month, quarterly), each month forming an independent liquidity pool. A digital asset is globally unified: BTC, unlike oil, has no delivery location, and, unlike Treasuries, no differing maturities; in essence it calls for a single, maximally deep liquidity pool. The multi-month design scatters across several pools the liquidity that ought to be concentrated in one, leaving each pool too shallow and raising the impact cost of large trades. In the digital-asset market, where liquidity is already thin relative to traditional markets, this fragmentation is especially costly.

Liquidity fragmentation is closely tied to price discovery. Precisely because multi-month futures cannot provide enough depth, price discovery for digital assets cannot proceed effectively within traditional futures. The market needs a product that gathers all

liquidity in one place. By abolishing the expiry date, perpetual futures concentrate all trading depth in a single contract, achieving a high degree of liquidity aggregation. This is not a marginal improvement but a paradigm shift from *liquidity dispersed across multiple temporal dimensions* to *liquidity concentrated in a single contract with no temporal dimension*.

Yet while perpetual futures eliminate fragmentation in time, they do not eliminate it in space. Perpetual futures on the same underlying are spread across dozens of exchanges—Binance, OKX, Bybit, dYdX, and Hyperliquid each run independent order books—and their depth is not shared, so fragmentation shifts from the temporal to the spatial. Under normal conditions, cross-exchange arbitrageurs effectively close this spatial gap, hedging across platforms whenever a price gap appears and keeping prices consistent among exchanges. In extreme markets, however, arbitrageurs face capital constraints (margin is locked on multiple exchanges and cannot be moved between them) and execution constraints (an exchange may restrict new position-opening or suspend withdrawals), so spatial fragmentation reemerges precisely when unified liquidity is most needed. Section 2.3.1 develops the analysis of arbitrageurs' fragility in extreme environments.

2.2.3 Trust mismatch

The trust foundation of traditional futures is the CCP—a centralized, regulated, well-capitalized institution. Participants trust the CCP, and the CCP trusts its members, forming layer upon layer of nested institutional trust. The core value of the digital-asset community is verifiability: it seeks permissionless trust guaranteed by code. This philosophical incompatibility generates conflict at three levels.

The CCP model requires high entry barriers: membership, capital, and compliance review are prerequisites for participation, at odds with the permissionless access of digital assets. The CCP's operations are opaque to outside participants—its core processes, such as risk-management models, margin calculations, and shortfall-handling procedures, are not public—conflicting with the radical transparency of on-chain systems. And the CCP is a single point of failure whose collapse can trigger a systemic crisis, contrary to the distributed resilience that digital assets pursue. This conflict is not a technical obstacle but a fundamental tension between two trust paradigms; Chapter 1's analysis of the three trust paradigms maps here onto the derivatives market. Perpetual futures provide, through the funding rate mechanism, an algorithmic-trust solution that does not depend on a CCP; but the solution is imperfect, and its costs are developed in Sections 2.3 and 2.4.

2.2.4 Ruling out the alternatives

If abolishing the expiry date and adapting to a round-the-clock market are the core requirements, at least two other products could serve similar functions: the perpetual option and the contract for difference (CFD). Understanding why perpetual futures won out helps pinpoint their distinctive value.

The perpetual option faces a structural market obstacle. One core input to option pricing is the time to expiry, and time-value decay is a fundamental feature of options.

Abolishing the expiry date sends the time to expiry to infinity, and the standard Black-Scholes framework for European options no longer applies. In theory, a perpetual American option does have a closed-form solution—Merton (1973) [11] derived the analytical pricing formula for the perpetual American put, and Hull (2022) [5] sets it out systematically—and academics and practitioners have proposed innovations such as everlasting options. The two must be kept distinct. Merton's perpetual American option is a mathematical object whose pricing depends on solving for the optimal exercise boundary and assumes the holder can exercise at any moment; everlasting options, by contrast, are a product design that rolls a series of short-dated options into long-term exposure through a funding-rate mechanism similar to that of perpetual futures, so the holder makes no exercise decision but pays a continuous financing cost. The core obstacle to the perpetual option is therefore not theoretical feasibility but the practical constraints of liquidity and complexity. As Section 2.1.3 showed, the nonlinear payoff of options and the five-dimensional complexity of market-making make it hard to build competitive market-making depth under crypto-market liquidity conditions. Options thus have distinctive value in the crypto market, particularly for institutional tail-risk hedging, but for these structural reasons they cannot become the dominant derivative.

The CFD has existed in traditional finance for decades. It, too, is a synthetic derivative (no expiry date, high leverage, no need to hold the underlying) and looks very much like a perpetual future. Yet the two differ fundamentally in structure. Under the typical B-book model, the counterparty to a CFD is the broker or platform: the user bets against the platform, the platform profits from the user's losses, and a structural conflict of interest exists. Some CFD brokers adopt the A-book model, hedging client order flow to external liquidity providers; the broker then does not profit directly from client losses, but the A-book model is not the norm in the retail CFD market. Perpetual futures, by contrast, match a trader's order against other traders' orders through an order book, so the market discovers the price, the platform earns fees, and users' gains and losses are not directly tied to the platform. A CFD's price-anchoring mechanism varies by model: under the B-book model, anchoring depends on the broker's reputation and regulatory constraints and is essentially an unverifiable promise; under the A-book model, the price tracks the underlying market's liquidity but has no explicit algorithmic anchor. Either way, the anchor is unverifiable: the user cannot independently audit whether the price is fair. The funding rate of a perpetual future supplies a mathematically verifiable algorithmic anchor instead. The distinctive position of perpetual futures is thus clear: they combine the convenience of a CFD (no expiry date, synthetic exposure, high leverage) with the market-based price discovery of futures (order-book matching eliminates the platform-user conflict of interest), and, through the algorithmic anchor of the funding rate, they strike an effective balance among degree of innovation, simplicity, and market-makability. In 2021, the United Kingdom's Financial Conduct Authority (FCA) banned the sale of crypto derivatives to retail consumers, and the European Securities and Markets Authority (ESMA) has held that crypto assets should be classified by reference to the boundary of financial instruments [12] [13]. This suggests that, as a matter of legal classification, perpetual futures may be brought within the CFD or other derivatives frameworks and subjected

to the associated leverage limits, negative-balance protection, and retail-investor protection rules—a risk discussed further in Chapter 26 (for regulatory evolution, see Section 2.4.5).

In sum, the two core design innovations of perpetual futures map onto the three structural conflicts above: abolishing the expiry date addresses temporal mismatch, the funding rate addresses trust mismatch, and the concentration of liquidity in a single contract—the market outcome of these two innovations—addresses liquidity fragmentation. The innovations and the outcome are mutually dependent: abolishing the expiry date lets liquidity concentrate, but it also removes the hard anchor of expiry convergence, which is why the funding rate is needed as a substitute soft anchor. Section 2.3 turns to the internal logic of these mechanism designs.

Figure 2-1 places the four forms—forward, futures, option, and perpetual—on a timeline, corresponding to the three leaps of personalized, institutionalized, and algorithmic trust.

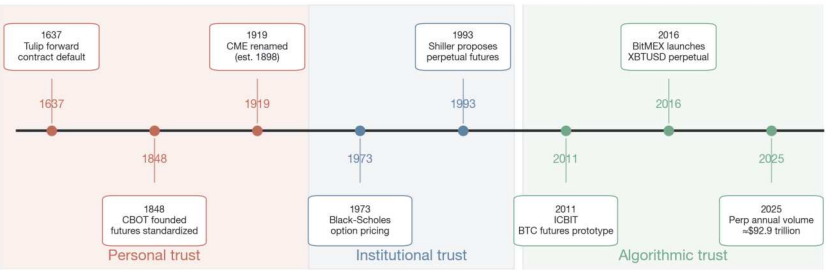


Figure 2-1. Timeline of derivatives evolution—the three leaps in trust mechanisms (Data source: Hull 2022 [5], Shiller 1993 [8], Hayes 2016 [9]; the roughly $92.9 trillion figure for 2025 annual perpetual futures volume shown in the figure follows [1][2])

The difference in payoff structure between futures and options is a key premise for understanding why perpetual futures rose to dominance. Figure 2-2 juxtaposes the two payoff curves: the futures payoff is linear and symmetric, while the option payoff is asymmetric (the downside is capped at the premium and the upside is theoretically unbounded). It is precisely this contrast between linear and nonlinear that drives the divergence in market-making complexity, cognitive barriers, and use cases described in Sections 2.1.3 and 2.2.4.

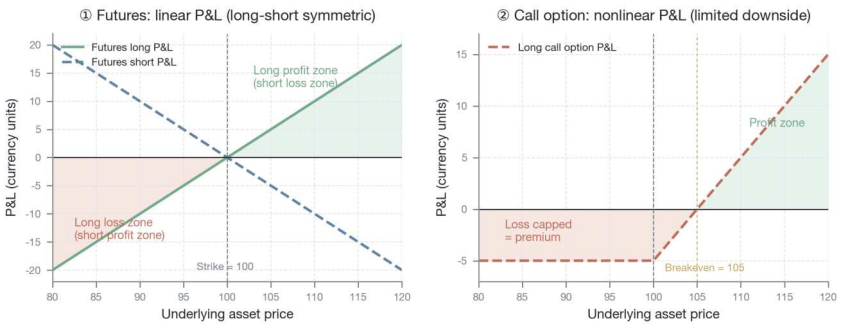


Figure 2-2. Comparison of the payoff structures of futures and options—linear versus nonlinear payoffs (Data source: Hull 2022 [5], Black & Scholes 1973 [6])

2.3 The core mechanisms of perpetual futures

This section examines the internal logic behind the design of perpetual futures. Their operation relies on two core mechanisms: the funding rate, which achieves price anchoring, and the mark price and liquidation mechanism, which achieve risk control. It also turns on two key design dimensions: the convexity difference between inverse and linear contracts, and the structural costs of abolishing the expiry date.

2.3.1 The funding rate

The funding rate is the core design innovation of perpetual futures. It is a fee settled directly between holders of long and short positions at regular intervals—commonly every 8 hours on mainstream exchanges, and every hour or continuously on some—whose direction and magnitude depend on the deviation between the perpetual price and the spot index price. The funding rate is not a fee charged by the exchange but an internal transfer payment among traders. Its common calculation framework is $\text{rate}_{8h} = \text{interest component}_{8h} + \text{premium component}_{8h}$; that is, the two components are summed over a single funding interval (such as 8 hours). Some exchanges set the interest component as a fixed term close to 0.01% per 8 hours, while others use different interest rates, caps and floors, and dynamic-adjustment mechanisms; the premium component is typically computed from a time-weighted average of the perpetual price's deviation from the spot index and is the main source of variation in the rate. Exchanges differ significantly in sampling frequency, weight-decay method, and the setting of caps and floors, and Chapter 10 develops a detailed comparison of the formulas.

Its core logic is a negative feedback loop: when the perpetual price is above spot, the funding rate is positive, longs pay shorts, the cost of being long rises, and longs are encouraged to close while new shorts enter, which pushes the price back down; the reverse holds when the perpetual price is below spot. Spot-futures arbitrageurs are the key force behind the anchoring mechanism: when the rate is positive, they short the perpetual while buying spot to earn the rate, placing selling pressure on the perpetual and buying pressure on spot and so accelerating the close of the gap. In extreme markets, however, arbitrageurs face multiple constraints, so the anchoring mechanism

is most likely to fail precisely when it is most needed. The margin on a short perpetual position is continuously eroded as the underlying price rises, and in extreme cases the arbitrageur is exposed to liquidation; cross-exchange arbitrage requires capital to be locked separately on the spot and derivatives exchanges, doubling the total capital committed with none of it usable across venues; and an exchange may restrict new position-opening or suspend withdrawals, directly blocking the arbitrage. These constraints mean that arbitrageurs—the force that holds the anchor—may be squeezed out or retreat voluntarily in extreme environments, which forms the microfoundation of the soft anchor's fragility and echoes the analysis in Section 2.3.4.

Ackerer, Hugonnier, and Jermann (2025) [14] derived an explicit pricing formula for perpetual futures under a framework of no frictions, continuous trading, and no arbitrage, proving mathematically that the funding rate mechanism can guarantee consistency between the futures price and the spot price and giving this negative feedback loop a rigorous theoretical basis. The model's core assumptions—especially frictionlessness and continuous trading—do not necessarily hold under real crypto-market conditions, which is itself a starting point for understanding how the rate deviates in extreme markets.

Figure 2-3 presents the internal logic of this negative feedback loop: when the perpetual price deviates from spot, the funding rate adjusts automatically with the deviation, guiding the price back by changing the cost of holding long and short positions. The figure also shows the two paths of premium and discount and the role of spot-futures arbitrageurs in accelerating convergence.

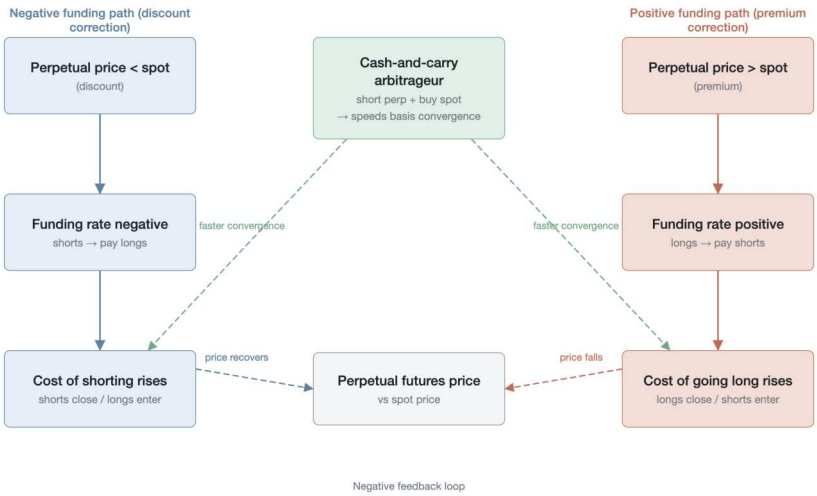


Figure 2-3. The negative feedback loop of the funding rate price-anchoring mechanism—the funding rate is positive when the perpetual price is above spot (longs pay) and negative when it is below spot (shorts pay), so a deviation automatically produces an opposing economic incentive (Data source: Ackerer, Hugonnier & Jermann 2025 [14])

Under extreme market conditions, the funding rate behaves in ways far more complex than this description suggests: it can spike to thousands of percent on an annualized basis, becoming a punitive tax on holding a position and triggering large-scale position

adjustments and further volatility. Chapter 10 develops the analysis of the rate's game-theoretic structure and extreme behavior.

The economic nature of the funding rate extends well beyond that of an anchoring tool. Its first nature is that of the continuous form of the traditional cost of carry. Traditional futures pricing follows the cost-of-carry model, in which the futures price equals the spot price plus the cost of carry; the cost of carry is embedded in the basis, and the more distant the expiry, the larger the basis. Perpetual futures abolish the expiry date and eliminate the explicit term basis, but they cannot eliminate the cost of carry: it shifts from being embedded in the price to being paid explicitly and continuously through the funding rate. A positive funding rate equals the cost of being long and can also be read as the implied interest rate on borrowed risk exposure. The Bank for International Settlements (BIS) working paper by Schmeling, Schrimpf, and Todorov shows that crypto carry can reach very high levels in some periods and that its returns relate to trend-chasing demand, limited arbitrage capital, and crowded-trade risk [15]. This is not a free lunch but the market's equilibrium pricing of leverage demand, funding constraints, and risk-bearing. The funding rate is in essence a real-time, explicit, and fragmented version of the basis.

Its second nature is that of a real-time market-sentiment indicator. Traditional sentiment indicators—the volatility index (VIX), based on option-implied volatility, and the put-call ratio, based on option volume—are complex to compute, limited in update frequency, and dependent on the liquidity of the options market. The funding rate provides a more direct, higher-frequency signal: a rate that stays positive and rises indicates that leveraged longs are accumulating and systemic fragility is rising. A positive rate, however, need not signal pure speculative long accumulation, because a large volume of delta-neutral basis trades also pushes the rate positive, and these arbitrage positions do not feed a liquidation cascade, so systemic fragility is far lower than when speculative longs dominate. The rate must therefore be read together with the increment (rather than the stock) of open interest, the distribution of liquidation thresholds, and spot-market capital flows; on its own it is a necessary but insufficient sentiment indicator. A rate that turns negative abruptly signals spreading panic, with longs forced to close or flip short; a rate above 100% annualized signals an extreme market state, in which the cost of holding a position itself begins to force traders to adjust. The rate is thus not only an anchoring tool but a piece of financial-information infrastructure in its own right, supplying real-time signals for volatility forecasting and market-state assessment that do not exist in traditional finance. Chapter 23 uses the rate to construct an early-warning indicator for the *calm trap*, and Chapter 24 incorporates it into a volatility-forecasting framework.

Its third nature is that of a foundation for an on-chain-native carry asset. The funding rate creates an on-chain-native carry trade: shorting the perpetual while buying spot to earn the rate; its risks are a rate reversal and liquidation. Projects such as Ethena have turned this strategy into a protocol; USDe delta-hedges spot against a corresponding short perpetual or dated futures position, packaging the strategy into a composable on-chain financial product [16]. But when a large amount of capital concentrates in the same carry strategy, that crowding fundamentally changes the return profile: it compresses carry, and when the rate reverses, the simultaneous unwinding of large

positions becomes a *carry unwind* that is itself an additional source of market impact. The carry strategy is therefore procyclical: it attracts capital and suppresses volatility in calm periods, and it is forced to deleverage and amplifies volatility in turbulent ones. Chapter 1's composability thesis is confirmed here at the level of derivatives: round-the-clock operation lets the rate settle continuously, atomic settlement makes rate payments certain, and programmability lets the strategy be packaged into a smart contract. This innovation is possible only where perpetual futures exist, marking their upgrade from a trading tool into infrastructure for a new asset class.

2.3.2 Mark price and the liquidation mechanism

The mark price is a fair-price benchmark used to calculate unrealized profit and loss (unrealized P&L) and margin levels. It synthesizes the price indices of several mainstream spot exchanges (for example, the median or weighted average of Coinbase, Binance, and Kraken) rather than relying on the traded price of a single exchange. It is designed to provide a manipulation-resistant liquidation benchmark, keeping an instantaneous price anomaly on a single exchange—a sharp, short-lived price deviation that quickly reverts, colloquially called a *wick*—from triggering unjust, large-scale liquidations. In the early history of the crypto market, mass liquidations caused by this kind of price manipulation were common. The mark price is typically smoothed with an exponential moving average (EMA) or a time-weighted average price (TWAP) to filter short-term noise. Even so, its price sources can still be manipulated. The 2019 BitMEX event, in which a flash crash on the single exchange Bitstamp caused mass liquidations, is a typical case: a large sell order propagated through a heavily weighted price source and indirectly triggered more than $200 million in liquidations. The governance mechanism for selecting price sources and adjusting weights is itself a point of operational risk; its decisions are typically made unilaterally by the exchange and lack transparency. Chapter 11 analyzes in depth how the smoothing design creates an *accumulation–release* effect in fast markets.

When the margin ratio (calculated on the mark price) falls below the maintenance margin ratio, the forced-liquidation process is triggered. The simplified formula for the margin ratio is:

$$\text{margin ratio} = \frac{\text{margin balance} + \text{unrealized P\&L}}{\text{position notional value}}, \qquad \text{liquidation} \iff \text{margin ratio} < \text{maintenance margin ratio}$$

where both unrealized P&L and notional value are calculated on the mark price. This is a conceptually simplified formula; for the rounding direction, precision alignment, and the margin-aggregation basis for isolated versus cross margin in a fixed-point implementation, see Chapter 11. Under isolated margin, only the margin balance and profit and loss of a single position count; under cross margin, the margin levels of all positions in the account are combined, so a loss on one position can be partly offset by profits on others, at the cost of cross-asset contagion risk. This section focuses on how liquidation-mechanism design differs across exchanges (for the basic framework of the liquidation-execution process, see Section 1.4.3), because those differences directly affect the intensity of the liquidation cascade.

Exchanges' choices in liquidation-mechanism design form a continuous spectrum from

aggressive to conservative. Table 2-1 sets out three representative models in the current market—cliff-edge liquidation, progressive liquidation, and vault takeover—comparing them along four dimensions (mechanism logic, market impact, advantages, and disadvantages) and revealing the different trade-offs each makes between the certainty of risk removal and minimal market impact.

Model	Mechanism	Market impact	Advantages	Disadvantages	Representative
Cliff-edge liquidation	100% of the position is liquidated instantly once margin hits the threshold	Greatest	Simple and certain; clears risk thoroughly	The entire liquidation volume is dumped at once, accelerating the cascade	Early CEX designs
Progressive liquidation	The position is reduced in batches as margin approaches the warning line	Moderate	Slows the cascade and gives traders time to react	Complex to implement and may delay risk removal	Bybit, OKX
Vault takeover	Liquidated positions are absorbed by the protocol's market-making vault	Smallest	Greatly slows the cascade and protects market structure	The vault bears directional risk and may itself become a risk point	Hyperliquid HLP

Table 2-1. Three design models for the liquidation mechanism of perpetual futures (Data source: compiled by the author)

Each design strikes this trade-off differently. Cliff-edge liquidation pursues certainty, but at the cost of dumping the entire liquidation volume into the order book at once, accelerating the liquidation cascade. Progressive liquidation seeks a balance, reducing positions in batches to spread the impact, but it adds implementation complexity and may, in extreme conditions, delay risk removal if the batch pace is too slow. Vault takeover is the newest direction, represented by Hyperliquid's HLP market-making vault: liquidated positions are absorbed within the vault rather than hitting the public order book, so the impact on the market price is minimal, but the vault itself bears concentrated directional risk. When a major directional move persists, the vault may accumulate a large opposing position and become a systemic risk point. Hyperliquid's JELLY incident of March 2025 offers a case study of this risk: public post-mortems show that the abnormal price and position pressure of the low-liquidity token JELLY forced the platform to protect the HLP vault through discretionary intervention at the validator level [17]. This does not mean that every vault-takeover model will fail, but it exposes a key tension: whether, in a crisis, a protocol retains enough governance and validator discretion to overturn market outcomes. Chapter 11 systematically compares these designs, and Chapter 29 designs an optimal solution from an engineering perspective.

2.3.3 Inverse contracts and linear contracts

BitMEX's XBTUSD perpetual futures contract is the classic inverse-contract design: it uses BTC as both margin and settlement currency and is quoted in USD. This design has an important economic feature: it superimposes coin-denominated settlement on same-direction movement of the collateral price. For an inverse contract with a fixed U.S.-dollar notional principal, the BTC-denominated profit and loss can typically be written as a reciprocal price difference of the form $\mathrm{PnL}_{\mathrm{coin}} \propto Q\left(\frac{1}{P_{\mathrm{entry}}} - \frac{1}{P_{\mathrm{exit}}}\right)$ (where Q is the contract quantity, or notional principal); nonlinearity therefore first appears in the coin-denominated settlement outcome rather than in a simple dollar-P&L formula. The trader in fact bears a dual exposure: the position's direction is exposed to the BTC price, and so is the margin asset itself. When the price falls, the position's loss widens while the dollar value of the margin shrinks, the margin buffer thins nonlinearly, and liquidation triggers faster. An inverse contract thus fuses the directional bet and the collateral exposure into one, creating a nonlinear risk structure rarely seen in traditional dollar-margined linear contracts.

The convexity effect of the inverse contract shows up on the payoff curve. Figure 2-4 juxtaposes the long payoff curves of inverse and linear contracts: the linear contract, quoted in a stablecoin, is approximately linear and proportional, whereas the inverse contract superimposes the coin-denominated payoff on the value of the BTC collateral, so the margin buffer is consumed faster in a falling market—the geometric reason it triggers liquidation faster.

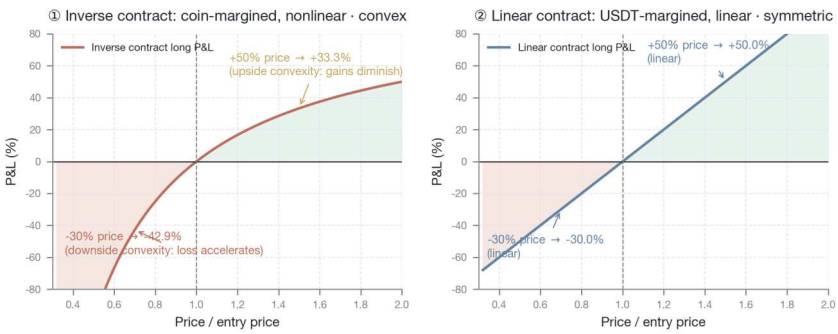


Figure 2-4. Comparison of the payoff curves of inverse and linear contracts—the convexity effect (Data source: derived by the author)

As stablecoins such as USDT and USDC became widespread, the market shifted toward linear contracts margined in stablecoins. A linear contract's profit and loss is linear, more intuitive and easier to compute for risk, but it loses the convexity of the inverse contract. In a rising market, a linear contract's leverage does not amplify automatically as an inverse contract's does. Linear contracts now dominate, because the spread of stablecoins has made dollar-denominated linear payoffs the default mental model for users. Inverse contracts retain their market, particularly where stablecoins are not used and in strategies that require the convexity effect. The difference between inverse

and linear contracts bears directly on the later analysis of liquidation mechanisms and market anomalies: Chapter 11 formalizes the nonlinear liquidation complexity of the inverse contract, and Chapter 18 analyzes the pricing differences between the two.

2.3.4 The cost of abolishing the expiry date

Abolishing the expiry date is the fundamental design decision that sets perpetual futures apart from traditional futures, but it is not free. In traditional futures, the expiry date performs three often-overlooked economic functions, and their disappearance imposes a threefold cost that perpetual futures must confront.

The first cost is the disappearance of a mechanism for the forced disclosure of information. The expiry-and-delivery of traditional futures guarantees that the contract price must converge to the spot price on the expiry date—a hard anchor backed by physical delivery or cash settlement. All market participants know this endpoint exists, so on every day before expiry, price discovery revolves around where the price will ultimately converge. Perpetual futures abolish this endpoint; the funding rate is a soft anchor that guides the price back through economic incentives but does not guarantee the return. The strength of the anchor depends on the activity and capital adequacy of arbitrageurs. Under normal conditions, arbitrageurs can usually compress the deviation between the perpetual and the spot index fairly quickly; but in extreme markets, capital constraints, exchange risk, and execution constraints weaken their capacity to correct the deviation, so it persists longer. As a result, the deviation between the perpetual price and the spot price can, in theory, be more persistent and more severe than in futures with an expiry date. One root of the persistent-premium or persistent-discount anomaly is precisely this structural weakening of the anchor once a hard anchor gives way to a soft one, an anomaly analyzed in depth in Chapter 18.

The second cost is the possibility of continuous leverage accumulation. Traditional futures have expiry, rolling, daily mark-to-market, and margin adjustment—mechanisms that together impose periodic repricing and risk checks; a position can be rolled over, so systemic exposure is not automatically reset to zero on an expiry date, but the roll window forces traders to reconfirm their funding cost, the basis, and margin constraints. Perpetual futures have no such window. Leverage can accumulate more continuously during calm periods until, at some trigger point, it sets off a liquidation cascade that forces destructive deleveraging. The leverage-cycle dynamics of Section 2.4.3 formalize this process.

The third cost is the loss of a termination date for market-making risk. In futures with an expiry date, market makers know that their inventory risk has a definite end date and can price and manage it precisely on that basis; in the worst case, everything is settled on the expiry date. In perpetual futures, a market maker's risk exposure is in theory indefinite and terminates only when the position is actively closed or liquidated. This changes their behavior: in a long-lasting trend especially, they may lean toward reducing quote depth or widening spreads, because they cannot be sure when the trend will end. Chapter 19 formalizes this effect in the market maker's profit equation. This cost echoes Chapter 1's analysis of round-the-clock trading at the level of derivatives: the expiry date is another risk-mitigation mechanism for market makers, and perpetual

futures abolish it.

Perpetual futures replace hard anchoring, periodic resets, and finite-horizon risk with soft anchoring, the possibility of unlimited accumulation, and indefinite-horizon risk. They are not an upgraded version of traditional futures but a different design trade-off, accepting a new set of structural costs in exchange for convenience. Understanding these costs is the prerequisite for analyzing the many distinctive phenomena in the perpetual futures market.

2.4 Market landscape and ecosystem evolution

Mechanism design determines what perpetual futures can do; the market itself reveals the impact they have had. By CoinGecko's top-10 perpetual-exchange measure, perpetual futures grew in less than a decade from nothing into a market with annual trading volume exceeding $92 trillion [1]. Their influence reaches far beyond a trading tool: they have reshaped the price-discovery hierarchy of digital assets, popularized high-leverage trading, given rise to an endogenous Minsky-style fragility, driven the rise of decentralized alternatives, and prompted a reconstruction of the global regulatory framework.

2.4.1 Market size and price discovery

In 2025, in the CoinGecko top-10 sample, the annual perpetual futures volume of centralized exchanges reached $86.2 trillion, up 47.4% year over year, while the annual volume of decentralized perpetual futures reached $6.7 trillion, up 346% year over year [1] [2]. By combining high liquidity, low transaction costs, and a broad base of participants, perpetual futures naturally became an important venue for expressing market information.

The evidence on price discovery must be read separately by market type. Alexander et al. (2020) [3] found that early BitMEX derivatives contributed a significant information share to BTC price discovery (this study uses data from 2016 to 2018, after which market structure has changed significantly: competition among exchanges, the degree of institutional participation, and the distribution of market liquidity are all now different). The cross-market analysis of Robertson and Zhang (2025) [4], by contrast, indicates that the regulated CME bitcoin futures market played a leading role in price formation within their sample. Together, the two support a more careful judgment: highly liquid derivatives markets tend to become important venues for expressing digital-asset information, but the conclusion for CME dated futures cannot be extended directly to all offshore or on-chain perpetual futures. This pattern, in which futures guide spot, has its root in the fragmentation of the digital-asset spot market (hundreds of exchanges worldwide, with no central market), which leaves leading derivatives exchanges, with their deeper liquidity pools, as the de facto centers of price coordination. When an important piece of information appears, the first to react may be the derivatives market, with its higher leverage efficiency and more concentrated liquidity; the reaction is then transmitted to the various spot exchanges through arbitrage.

This constitutes a conceptual inversion: the derivatization of the spot market. In

traditional finance, the spot market is usually primary and derivatives are built on top of it. In the digital-asset market, this hierarchy weakens, at least in some periods and for some trading pairs: derivatives markets such as perpetual futures and dated futures may become the faster layer of information expression, with spot prices then absorbing that information through arbitrage and market-making. This structural change means that understanding the dynamics of the derivatives market is just as important to understanding spot prices. A deeper systemic risk follows: when price discovery is shaped by leveraged market depth, the price signal may embed the systemic noise of the leverage cycle. This has profound effects on DeFi protocols that rely on spot prices (for collateral valuation and oracle price feeds, for example), because they typically assume the spot price reflects fundamental information rather than the accumulation or liquidation of leveraged positions. Just how strongly perpetual futures lead price discovery, and whether that leadership is stable across market states, calls for microstructure tools such as the Hasbrouck information share, which Chapters 13 through 15 analyze systematically.

2.4.2 The democratization of leverage

In traditional finance, high-leverage trading is usually subject to strict account, margin, and investor-suitability constraints. In the U.S. stock market, Regulation T limits initial margin to 50%, equivalent to a maximum of 2x leverage [18]. Perpetual futures sharply lowered the barrier to leverage, turning nominal 10x-to-125x leveraged trading into something anyone can access simply by downloading an app or connecting a wallet. Nominal leverage must be distinguished from effective leverage: 125x is usually available only for very small positions, the tiered-margin system makes the effective maximum for large positions far lower, and actual leverage is further constrained by platform risk controls, liquidity, and liquidation thresholds. This is a democratization of leverage that never appeared in traditional finance. It echoes the broader debate over the democratization of finance in fintech: technology has lowered the barrier to entry for financial services, but whether wider access comes with adequate risk education and investor protection remains a central, ongoing debate among academics and regulators. This book uses *democratization of leverage* as a descriptive term, with no normative connotation.

Understanding it starts with the behavioral drivers on the demand side. The forces that lead retail traders to choose 50x-to-125x leverage include overconfidence bias, gambling utility, and a returns-display bias on social media: profit screenshots are widely shared while losses are hidden, producing a socialized amplification of survivorship bias. The default setting of the leverage slider and one-tap execution also reduce perceived friction, making high leverage cognitively too easy to choose.

The democratization of leverage cuts both ways. On the positive side, it releases capital efficiency: traders with little capital can gain exposure to large positions on limited margin, hedgers can lock in risk with less capital, and overall capital efficiency improves markedly. On the negative side, it amplifies risk systemically: a large volume of dispersed leverage risk accumulates at the system level, and every high-leverage retail account is a potential liquidation trigger. When the market makes a directional move,

these dispersed liquidations fire at once, forming a liquidation cascade. Democratizing leverage thus concentrates risk. The concentration unfolds along three dimensions: the dense distribution of liquidation thresholds across the price space produces the cascade effect; tail losses concentrate on the insurance fund and ADL recipients; and systemic risk concentrates on a few leading exchange platforms. At the individual level, 50x-to-100x leverage means that a mere 1%-to-2% adverse move triggers liquidation, leaving the trader's risk-tolerance window extremely narrow. At the system level, when millions of such narrow windows are densely packed along the price axis, any directional shock sweeps across many liquidation thresholds, forming a wave of liquidations. The selling pressure from those liquidations pushes the price further, sweeping across still more thresholds, and a positive feedback loop forms—the microstructure mechanism of the liquidation cascade.

2.4.3 Leverage-cycle dynamics

The liquidation cascade is not an occasional event but an endogenous, cyclical dynamic of the perpetual futures market. This cycle maps closely onto Minsky's (1986) [19] financial instability hypothesis, but it runs far faster and lacks the institutional tools to manage the cycle.

Figure 2-5 formalizes this leverage cycle as a four-stage closed loop (accumulation, criticality, deleveraging, and recovery), in which the end of the recovery phase is the start of the next accumulation phase, so the cycle repeats. The figure marks the typical direction of change in each phase's core observables—volatility, open interest, the direction of the funding rate, and effective liquidity depth—providing a signal framework for later empirical identification.

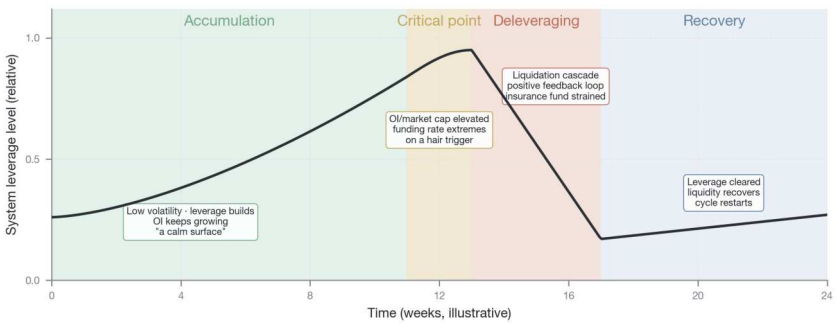


Figure 2-5. A four-stage model of the leverage cycle in the perpetual futures market—accumulation, criticality, deleveraging, and recovery (conceptual schematic; the leverage level is on a relative scale and the time axis is illustrative; theoretical source: Minsky 1986 [19])

In the accumulation phase, fragility builds beneath surface calm. Volatility is low, traders gradually add leverage, the rate is mildly positive, the carry trade attracts more capital, open interest grows steadily, and the market appears to run smoothly. The process shows a clear reflexivity loop: low volatility → added leverage → carry attracts capital → volatility suppressed further. Danielsson, Shin, and Zigrand's (2012)

[20] analysis of endogenous risk reveals the core paradox of this loop: the risk indicator is itself endogenously suppressed by trading behavior, so risk assessments based on historical volatility systematically underestimate true risk. Effective liquidity—order-book depth divided by the size of positions that can be liquidated—declines steadily. This is precisely the stability-breeds-instability that Minsky described: hedge financing gradually degrades into speculative financing and then into Ponzi financing, and it unfolds faster in the perpetual futures market. At the critical point, the system is extremely fragile. The ratio of open interest to market capitalization climbs to a high level, the rate exceeds 100% annualized, the cost of holding rises but long holders do not cut positions, and the market's tipping point falls ever lower. Here a medium-sized shock can, given the system's fragility, produce consequences far exceeding its own scale, and the system's nonlinear amplification peaks.

In the deleveraging phase, the liquidation cascade unfolds. The market event of October 10–11, 2025 provides a typical empirical case [21] [22]. An external macro shock—the spark, though months of prior leverage accumulation were the dry tinder—set off the first wave of liquidations; the selling pressure from those liquidations pushed the price down further and triggered still more, forming a positive feedback loop. Public data indicate that reported liquidations exceeded $19 billion within about one day [21]; Ali (2025), in an SSRN working paper, uses a measure of open interest erased to describe a contraction of open interest of roughly $19 billion within 36 hours [23]. The two figures are close but mean different things: the former measures reported forced liquidations, the latter the reduction in open interest, and they cannot be used interchangeably. Exchange liquidation data, moreover, generally underestimate the true scale.

Figure 2-6 uses market data from October 10–11, 2025 to depict this positive feedback loop: the phases in which the BTC price accelerated downward coincide closely on the time axis with pulse-like surges in hourly liquidation volume, confirming the positive-feedback amplification between forced-liquidation selling pressure and price decline.

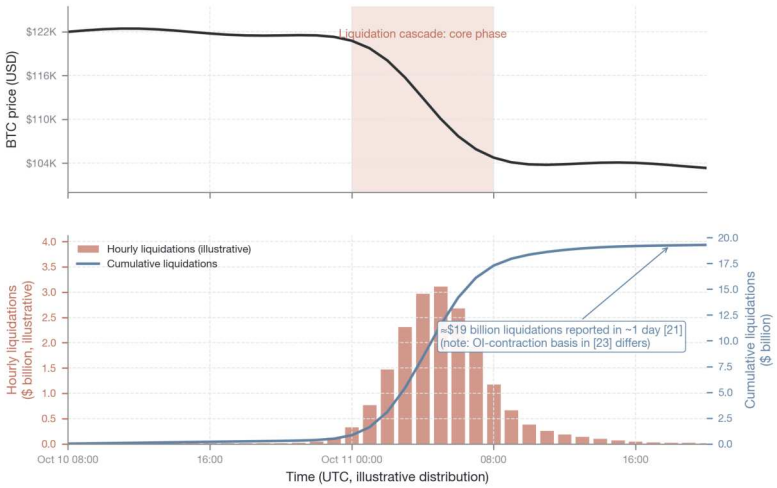


Figure 2-6. During the October 2025 liquidation-cascade event, the BTC price trajectory (top) and hourly and cumulative liquidation volume (bottom); the phases of accelerating price decline coincide closely in time with the peaks in liquidation volume, revealing forced liquidation as a positive-feedback amplification mechanism of the market crash (the data form is an illustrative model based on the event's public measures, not a tick-by-tick measurement; the cumulative-liquidation figure of "reported liquidations exceeding $19 billion within about one day" is a reported measure [21] and differs in meaning from the open-interest-contraction measure [23]; source: FTI Consulting 2025 [21], CoinGlass 2026 [22])

Ali (2025) [23] dissects three key risk amplifiers in this event through a market-microstructure lens. First, the procyclicality of unified margin: when multiple assets serve as unified margin, the contagion path unfolds clearly—initial shock → decline in the value of margin assets → cross-asset forced liquidation → decline in the price of the liquidated asset → further margin shrinkage—forming a cross-asset liquidation spiral, for which Brunnermeier and Pedersen's (2009) [24] framework of the margin spiral provides a rigorous analytical basis. Second, exchange infrastructure bottlenecks: during periods of sharp price movement, APIs and interfaces stall, and traders cannot top up margin or close positions manually in time. Third, insurance-fund strain and auto-deleveraging risk: the insurance fund is funded by the surplus when positions are liquidated at prices better than the bankruptcy price. In most deleveraging events the fund is merely strained, but in extreme cases it can be exhausted, triggering ADL. ADL forcibly cuts the positions of the winning side, stripping away their profits—an additional blow that further shakes market confidence. The second cost of abolishing the expiry date becomes concrete here: the expiry, rolling, daily mark-to-market, and margin system of traditional futures provide periodic repricing windows; perpetual futures have no expiry window, so leverage accumulates more easily and continuously and is ultimately released through a catastrophic deleveraging.

Similar leverage accumulation and catastrophic deleveraging have played out repeatedly in the digital-asset market. The chain-liquidation event of May 2021 followed a similar accumulation–criticality–deleveraging–recovery path, providing cross-period empirical support for the model's cyclicality hypothesis. The deleveraging triggered by the col-

lapse of FTX in November 2022 must be classified separately: it was counterparty-risk-type deleveraging (exchange failure → collapse of trust → forced deleveraging), fundamentally different in trigger from the market-endogenous deleveraging of May 2021 and October 2025 (leverage accumulation + price shock → liquidation cascade). This chapter's leverage-cycle model primarily explains market-endogenous deleveraging; counterparty-risk-type deleveraging involves the broader problem of institutional trust. The October 2025 event, larger in scale and faster in pace, further confirmed the systemic nature of this dynamic.

The recovery phase is asymmetric: the crash is measured in hours, recovery in days to weeks. Leverage is cleared out, open interest falls to a low level, market makers reenter, liquidity is restored, the market returns to the accumulation phase, and the cycle begins again.

The perpetual futures leverage cycle is highly isomorphic to Minsky's classic credit cycle in the dynamics of financial-leverage accumulation and collapse, but it lacks the dimensions of capital misallocation and overcapacity found in the real economy. Brunnermeier and Pedersen's (2009) [24] theory of the margin spiral may be the more precise benchmark, since it directly models the evaporation of liquidity and forced deleveraging under margin constraints and fits the microstructure of the perpetual futures market more closely. The two models differ significantly in speed, management tools, and information visibility. Table 2-2 compares them along four core dimensions, showing why the perpetual futures cycle is faster, more severe, and thinner in institutional buffers, yet at the same time more amenable to early warning because of the transparency of on-chain data.

Dimension	Traditional credit cycle	Perpetual futures leverage cycle
Speed	Measured in years; a full cycle lasts 5 to 10 years	Measured in weeks to months; a full cycle within a few months
Cycle-management tools	Central-bank rate hikes and cuts, macroprudential policy, lender of last resort	The funding rate passively reflects the market state; no active steering tool
Deleveraging mode	Policy guidance, orderly institutional bankruptcy, and central-bank liquidity injection	Liquidation cascade; market-crash-style automatic deleveraging; limited insurance-fund capacity
Information visibility	Leverage accumulation is usually invisible, attributed after the fact	Aggregate-level open-interest (OI) and rate data are public, but key risk-distribution information (the distribution of liquidation thresholds, whale concentration, cross-platform net exposure) remains opaque on CEXs; on-chain DEX data are more transparent but limited in share. Visible does not mean manageable

Table 2-2. Comparison of the traditional Minsky credit cycle and the perpetual futures leverage cycle (Data source: compiled by the author, drawing on Minsky (1986) [19] and Brunnermeier & Pedersen

(2009) [24])

Is the liquidation cascade an endogenous feature of perpetual futures, or can mechanism design mitigate it? Why does liquidity disappear precisely when it is most needed? Of the price crash produced by liquidations, how much is information and how much is noise? Can leverage accumulation during calm periods be monitored and flagged in real time? These questions are taken up, respectively, in Chapter 11 on liquidation-mechanism design, Chapters 19 through 21 on the economics of liquidity, Chapter 22 on mechanism-driven volatility, and Chapters 23 and 24 on volatility forecasting and Minsky-moment early warning.

2.4.4 Risks unique to DEXs

The structural risks of centralized exchanges have grown increasingly exposed (the collapse of FTX being the most extreme case), driving significant growth in the market share of decentralized perpetual futures exchanges. Extrapolating from full-year volume ($6.7 trillion / $86.2 trillion), the ratio of DEX perpetual to CEX perpetual volume in 2025 was roughly 7.8%, well above the roughly 2.1% level in early 2023 [1] [2]. Total DEX perpetual volume in 2025 was about $6.7 trillion, up roughly 3.5 times over the prior year (for the data basis, see the opening of this chapter). DEX nominal volume may be inflated by trade-mining incentives and wash trading, so real economic volume may be lower than the nominal figure. With its distinctive architectural choices—a purpose-built Layer-1 app-chain, an on-chain central limit order book, and the HLP market-making vault—Hyperliquid recorded about $2.9 trillion in DEX perpetual volume for the full year of 2025, roughly 43% of the CoinGecko top-10 DEX perpetual sample, and now represents the architectural direction of app-chains, on-chain order books, and protocol-run market-making vaults [1] [25]. Hyperliquid's own Layer-1 chain warrants careful assessment: the size and distribution of its validator set and the decentralization of its sequencer fall short of protocols that run on the Ethereum Layer 1, and the discretionary intervention at the validator level during the JELLY incident (discussed in Section 2.3.2) is precisely a symptom of this tension [17]. Chapter 9 analyzes in depth the gains and losses of its architectural choices.

The difference between centralized and decentralized perpetual futures exchanges is not merely one of technical architecture but a fundamental divide in trust paradigm. Table 2-3 compares the two models along five dimensions—asset custody, transparency, core risk, trading experience, and trust paradigm—providing a structured reference for the analysis of DEX-specific risks that follows.

Feature	Centralized exchange	Decentralized exchange
Asset custody	The exchange's centralized wallet; users must trust the platform	Smart contract or self-custody; non-custodial
Transparency	Low; relies on audits and proof of reserves	Very high; all trades and funds are verifiable on-chain
Core risk	Operational opacity, asset misappropriation, and a single point of failure	Smart-contract vulnerabilities, oracle risk, and maximal extractable value (MEV) attacks
Trading experience	Smooth, low-latency, feature-rich	Rapidly improving, but limited by the performance of the underlying chain
Trust paradigm	Platform trust	Architectural trust

Table 2-3. Core differences between centralized and decentralized perpetual futures exchanges (Data source: compiled by the author)

While achieving non-custodial operation and on-chain transparency, DEXs introduce several risks that do not exist, or are milder, on CEXs. Oracle risk arises because DEXs rely on external price oracles (such as Chainlink or Pyth) to obtain the spot prices used to calculate the mark price and trigger liquidations. The BIS analysis of the DeFi oracle problem notes that an on-chain financial system's reliance on external price sources reintroduces issues of data quality, price-source manipulation, and governance trust at the protocol layer [26]. In low-liquidity token markets, an attacker can influence the oracle's quote by manipulating a few price sources, triggering a liquidation in their own favor. Oracle price updates carry an inherent delay (typically on the order of seconds), which in extremely volatile markets can open a significant gap between the mark price and the true market price, throwing off the timing of liquidations. An interruption in oracle service can pause a DEX's liquidation system, so that accumulated risk is released all at once when the oracle recovers.

MEV-based attacks are the second class of risk unique to DEXs. On a public blockchain, pending transactions are visible to everyone in the mempool, and block builders can extract value by reordering transactions. A sandwich attack—in which an MEV searcher, on detecting a user's large trade, inserts its own transactions before and after it to profit from the price slippage—worsens the user's execution price. Liquidation opportunities are visible to all, so multiple liquidators compete to execute, producing gas competition, and part of the liquidation profit is converted into validator income. The sandwich attack targets mainly DEXs based on the automated market maker (AMM) model; for DEXs that use a central limit order book (such as Hyperliquid), MEV risk shows up more as front-running and information-advantage problems at the sequencer level, tied directly to how decentralized the sequencer is. On a CEX, the exchange controls trade ordering and there is no public-mempool MEV problem, but a DEX exposes trade ordering to all block producers.

Smart-contract vulnerabilities are the third class of risk unique to DEXs. All of a DEX's logic is coded in smart contracts, and once a contract is deployed its logic is public to all, so attackers can audit it in search of exploitable vulnerabilities. Security

reports from firms such as Immunefi show that DeFi and on-chain trading protocols have long suffered fund losses from smart-contract vulnerabilities, private-key leakage, and economic attacks [27]; the DeFi policy recommendations of the International Organization of Securities Commissions (IOSCO) likewise list governance, code, oracles, and cross-chain dependencies as important risk surfaces [28]. Although auditing and formal verification keep improving, zero vulnerabilities remains an unattainable goal, especially in a perpetual futures system where modules such as liquidation logic, margin calculation, and multi-collateral management interact in highly complex ways. Upgradeable smart contracts fix vulnerabilities but raise the trust question of who holds the authority to upgrade; if that authority rests with a few, the promise of decentralization weakens. Non-custodial operation eliminates the risk of platform misappropriation but introduces contract-vulnerability and admin-key risk, and the admin authority of an upgradeable contract is essentially a form of *code custody*. Chapter 28 systematically discusses this tension in the context of trust-stack design.

Regulatory-compliance risk likewise cannot be ignored. In 2023, the enforcement action by the Commodity Futures Trading Commission (CFTC) against several DeFi protocols, including Opyn, ZeroEx, and Deridex, showed that decentralization is not grounds for regulatory exemption [29]. Satisfying anti-money-laundering and know-your-customer (AML/KYC) requirements in a permissionless environment is a dilemma both technical and philosophical: there is a fundamental contradiction between mandatory identity verification and permissionless access. Protocol developers and governance participants may face personal legal liability, a risk that deeply constrains the long-term development of the DEX ecosystem.

DEXs also face two further risks. Governance-concentration risk arises when a few token holders or a team control key protocol parameters (such as the margin ratio, liquidation thresholds, and rate parameters), which may amount to de facto centralization. Cross-chain-bridging risk arises when assets are bridged to an app-chain: once a bridge contract is breached, user assets face systemic loss.

Together, these risks make up the DEX's trust-tax substitution: the CEX's trust tax is reliance on a centralized operator, while the DEX's is reliance on the correctness of code, the reliability of oracles, and the fairness of block producers. Chapter 1's risk-surface-transformation framework is confirmed once again: decentralization does not eliminate risk but transforms it from one surface to another.

2.4.5 Regulatory evolution

The regulatory evolution of perpetual futures passed through four phases. In the vacuum period, from 2016 to 2020, regulators had little awareness of or attention to crypto derivatives, the market grew in an almost unregulated environment, and exchanges such as BitMEX and Binance offered high-leverage trading worldwide without identity verification or anti-money-laundering review. In the enforcement-based period, from 2020 to 2023, regulators began to state their positions through enforcement actions: in 2020 the CFTC filed suit against BitMEX [30], and the collapse of FTX directly accelerated the tightening of regulation. The hallmark of this period, however, was the absence of a clear rule framework, so participants did not know what was compliant, only what

would be prosecuted. In the framework-building period, from 2023 to 2025, major economies shifted from enforcement to legislation. The European Union's Markets in Crypto-Assets (MiCA) regulation applied in phases: the rules for asset-referenced tokens and e-money tokens from June 30, 2024, and the rules for crypto-asset service providers (CASPs) and other main provisions from December 30, 2024 [31]. MiCA primarily governs the issuance of, and service providers for, crypto assets not regulated as financial instruments under existing EU financial-services law; a crypto derivative, if it constitutes a financial instrument, is more likely to be governed under the MiFID II framework. ESMA's guidelines explain that the boundary between MiCA and MiFID II depends on the rights conferred by the asset, the contract structure, and the trading arrangement, rather than on the single factor of *whether it is on-chain* [12]. MiCA thus covers the spot side and the CASP framework, while MiFID II covers the financial-instrument and derivatives side. In the United States, the jurisdictional dispute between the U.S. Securities and Exchange Commission (SEC) and the CFTC is the core obstacle, and Congress is accelerating legislation to clarify which crypto assets are securities and which are commodities. Hong Kong implemented a licensing regime for virtual asset service providers (VASPs) in 2023, and Singapore applies a differentiated regime to crypto derivatives through its Monetary Authority. The compliance-transition period began in 2025, as market participants undertake compliance retrofits under a clear framework and compliance becomes a necessary condition for market access.

Yet a fundamental tension exists between the global nature of the perpetual futures market and the territoriality of regulation. When major economies tighten regulation, trading volume may shift to offshore jurisdictions with looser rules, and FTX's exploitation of the Bahamian regulatory environment is a cautionary precedent. The persistence of regulatory arbitrage may weaken the substantive effect of the compliance transition, and progress in international regulatory-coordination mechanisms—such as the cross-border coordination frameworks of the Financial Stability Board (FSB) and IOSCO—will be the decisive factor.

The dual effect of regulation must be confronted directly. In the short term, it constrains the market: it limits the leverage available to retail traders, raises exchanges' compliance costs, and suppresses some innovation—DeFi protocols in particular face the technical and philosophical dilemma of satisfying identity-verification requirements in a permissionless environment. In the long term, it enables the market: a more compliant, more transparent market with better risk controls is a prerequisite for attracting institutional investors at scale; regulatory clarity reduces uncertainty and lowers the compliance-risk premium, and consumer-protection measures reduce fraud and manipulation, strengthening the market's overall credibility and sustainability. Chapter 26 develops a systematic economic analysis of regulation.

2.5 Chapter summary

The 400-year evolution of derivatives is, in essence, three leaps in trust technology. Forward contracts rely on personalized trust; their radius of trust is limited to a circle of acquaintances, and market size is strictly constrained because the cost of trust is too

high. Futures contracts achieve institutionalized trust through standardization and the CCP, extending the radius of trust to strangers worldwide, but they depend on a centralized institution and introduce the inherent contradiction of concentrating risk in order to manage it. Perpetual futures introduce algorithmic anchoring through the funding rate and margin rules, reducing reliance on the traditional CCP; but in the CEX setting they still depend on platform rules and operation, while in the DEX setting they shift trust to code, oracles, sequencing, and governance. Each leap responds to the trust bottleneck left by the previous one, while also generating a new trust tax.

The three structural conflicts between traditional futures and digital assets—temporal mismatch, liquidity fragmentation, and trust mismatch—make it infeasible to transplant traditional futures directly into the digital-asset world. The two core design innovations of perpetual futures (abolishing the expiry date and the funding rate) and their market outcome (the concentration of liquidity in a single contract) address these three conflicts directly; the causal logic of their mutual dependence (a soft anchor replacing a hard anchor) is analyzed in Section 2.2.4.

The core theoretical contribution of this chapter is cost awareness. Abolishing the expiry date eliminates roll costs but weakens three economic functions: the hard anchor of expiry convergence becomes the soft anchor of the funding rate, the periodic repricing window gives way to continuous position-holding, and the termination date for market-making risk becomes indefinite. These three costs explain many of the distinctive phenomena in the perpetual futures market. The threefold economic nature of the funding rate—the continuous form of the traditional cost of carry, a real-time market-sentiment indicator, and the foundation for an on-chain-native carry asset—upgrades perpetual futures from a trading tool into financial-information infrastructure and the foundation for a new asset class. The leverage-cycle dynamics reveal that the liquidation cascade is an endogenous, Minsky-style cycle of the perpetual futures market, faster than the traditional credit cycle and lacking central-bank-style tools to manage it. The October 2025 event (roughly $19 billion by both the reported-liquidation and open-interest-contraction measures, detailed in Section 2.4.3 [21] [23]) provides direct evidence for this dynamic.

Perpetual futures are both a product of the trust tax—the trust mismatch of traditional futures gave rise to them—and a restructuring of it: CEX perpetuals concentrate trust in platform rules, risk control, and asset custody, while DEX perpetuals shift trust to code, oracles, sequencing, and governance. Compression is not free, and perpetual futures introduce their own trust tax: the threefold cost of abolishing the expiry date, the endogenous fragility of the leverage cycle, the new risk surfaces of smart contracts and oracles, and the transitional cost of regulatory uncertainty.

The next chapter closes out Part 1. Chapter 1 established the analytical framework of the trust tax, and this chapter set out the logic and cost of the birth of perpetual futures; Chapter 3 now argues that perpetual futures are not merely a crypto-native tool but represent a general direction in the evolution of financial infrastructure, pushing the narrative of trust-tax compression toward a more macroeconomic perspective. As the core object of analysis in this book, perpetual futures run through all subsequent chapters: from mechanism deconstruction to price discovery, from the arbitrage ecosystem

to the economics of liquidity, from volatility dynamics to market-quality assessment, and from decentralization engineering to the future in the age of AI.

References

[1] CoinGecko. (2026). *2025 annual crypto industry report*. https://www.coingecko.com/research/publications/2025-annual-crypto-report

[2] CoinGecko. (2026). *The rise of perpetuals and perp DEXs*. https://www.coingecko.com/learn/rise-of-perpetuals-and-perp-dexs

[3] Alexander, C., Choi, J., Park, H., & Sohn, S. (2020). BitMEX bitcoin derivatives: Price discovery, informational efficiency, and hedging effectiveness. *Journal of Futures Markets*, *40*(1), 23–43. https://doi.org/10.1002/fut.22050

[4] Robertson, K., & Zhang, R. (2025). Price discovery in bitcoin spot and futures markets. *Journal of International Money and Finance*, *159*, 103415. https://doi.org/10.1016/j.jimonfin.2025.103415

[5] Hull, J. C. (2022). *Options, futures, and other derivatives* (11th ed.). Pearson.

[6] Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654. https://doi.org/10.1086/260062

[7] Garbade, K. D., & Silber, W. L. (1983). Price movements and price discovery in futures and cash markets. *The Review of Economics and Statistics*, *65*(2), 289–297. https://doi.org/10.2307/1924495

[8] Shiller, R. J. (1993). Measuring asset values for cash settlement in derivative markets: Hedonic repeated measures indices and perpetual futures. *The Journal of Finance*, *48*(3), 911–931. https://doi.org/10.1111/j.1540-6261.1993.tb04024.x

[9] Hayes, A. (2016, May 13). *Announcing the launch of the perpetual XBTUSD leveraged swap*. BitMEX Blog. https://www.bitmex.com/blog/announcing-the-launch-of-the-perpetual-xbtusd-leveraged-swap

[10] CME Group. (2025). *Bitcoin futures and options on futures*. https://www.cmegroup.com/markets/cryptocurrencies/bitcoin/bitcoin.html

[11] Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science*, *4*(1), 141–183. https://doi.org/10.2307/3003143

[12] European Securities and Markets Authority. (2024). *Guidelines on the conditions and criteria for the qualification of crypto-assets as financial instruments*. https://www.esma.europa.eu/document/guidelines-conditions-and-criteria-qualification-crypto-assets-financial-instruments

[13] Financial Conduct Authority. (2020). *FCA bans the sale of crypto-derivatives to retail consumers*. https://www.fca.org.uk/news/press-releases/fca-bans-sale-crypto-derivatives-retail-consumers

[14] Ackerer, D., Hugonnier, J., & Jermann, U. (2025). Perpetual futures pricing. *Mathematical Finance*. Advance online publication. https://doi.org/10.1111/mafi.70018

[15] Schmeling, M., Schrimpf, A., & Todorov, K. (2023; revised 2025). *Crypto carry* (BIS Working Papers No. 1087). Bank for International Settlements. https://www.bis.org/publ/work1087.htm

[16] Ethena Labs. (2026). *Ethena documentation*. https://docs.ethena.fi/

[17] Hyperliquid Community Docs. (2025). *March 26 JELLY incident*. https://hyperliquid-co.gitbook.io/community-docs/introduction/roadmap/2025-26-03_incident

[18] Board of Governors of the Federal Reserve System. (2023). *Regulation T: Credit by brokers and dealers*. https://www.ecfr.gov/current/title-12/chapter-II/subchapter-A/part-220

[19] Minsky, H. P. (1986). *Stabilizing an unstable economy*. Yale University Press.

[20] Danielsson, J., Shin, H. S., & Zigrand, J.-P. (2012). Endogenous and systemic risk. In *Quantifying Systemic Risk* (pp. 73–94). University of Chicago Press. https://doi.org/10.7208/chicago/9780226921969.003.0004

[21] FTI Consulting. (2025, December 24). *Crypto crash Oct 2025: When leverage met liquidity*. https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity

[22] CoinGlass. (2026). *2025 crypto derivatives market annual report*. https://www.coinglass.com/learn/2025-annual-report-en

[23] Ali, Z. (2025, October 15). Anatomy of the Oct 10–11, 2025 crypto liquidation cascade: Macroeconomic triggers, market microstructure, and systemic risk lessons. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5611392

[24] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[25] Hyperliquid. (2025). *Hyperliquid documentation*. https://hyperliquid.gitbook.io/hyperliquid-docs

[26] Bank for International Settlements. (2023). *The oracle problem and the future of DeFi* (BIS Bulletin No. 76). https://www.bis.org/publ/bisbull76.pdf

[27] Immunefi. (2024). *Crypto losses in 2023*. https://immunefi.com/research/

[28] International Organization of Securities Commissions. (2023). *Policy recommendations for decentralized finance*. https://www.iosco.org/library/pubdocs/pdf/ioscopd754.pdf

[29] Commodity Futures Trading Commission. (2023). *CFTC orders operators of three DeFi protocols to pay penalties for illegal derivatives trading* [Press release]. https://www.cftc.gov/PressRoom/PressReleases/8774-23

[30] Commodity Futures Trading Commission. (2020). *CFTC charges BitMEX owners with illegally operating a cryptocurrency derivatives trading platform* [Press release]. https://www.cftc.gov/PressRoom/PressReleases/8270-20

[31] European Parliament & Council of the European Union. (2023). Regulation (EU) 2023/1114 on markets in crypto-assets (MiCA). *Official Journal of the European Union*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023R1114

The Distinctive Value of Digital Assets and the Future of Finance

In March 2024, BlackRock launched BUIDL, a tokenized money market fund, on Ethereum; within roughly a year its assets under management (AUM) surpassed $1 billion, and by the end of 2025 it held approximately $2.5 billion to $2.9 billion [1]. Over the same period, the Depository Trust & Clearing Corporation (DTCC) explored blockchain-based T+0 atomic settlement [2], the European Investment Bank issued a €100 million digital bond on Ethereum in 2021 [3], Siemens issued a €60 million digital bond on a public blockchain in 2023 [4], Visa ran stablecoin settlement tests on Solana [5], and the Monetary Authority of Singapore's Project Guardian pursued institutional-grade DeFi [6]. These initiatives come from core institutions of the global financial system—asset managers overseeing more than $10 trillion in assets, settlement-infrastructure operators processing more than $2 quadrillion in securities transactions each year, sovereign-level investors, and global payment networks—and they represent not the tentative adoption of a new technology but strategic deployment grounded in cost-benefit analysis. By the end of 2025, the total market capitalization of stablecoins worldwide stood at roughly $310 billion, with annualized on-chain settlement volume of approximately $33 trillion (Artemis/Bloomberg [7]; this figure includes internal DeFi flows and maximal extractable value (MEV) and is not netted, and payment-adjusted estimates diverge widely by method, ranging from roughly $0.4 trillion to $10 trillion [8]). Even at the adjusted figure, this volume remains within the same order of magnitude as Visa's annual net payment volume [9]. The functional boundary between on-chain finance and traditional finance continues to narrow.

Chapters 1 and 2 were, at bottom, still answering questions internal to the crypto world: what makes digital assets different, and how those differences gave rise to perpetual futures. This chapter shifts the vantage point—from inside the crypto world to the global financial system—to answer a more fundamental question: is the distinctiveness of digital assets merely a specialized tool serving a particular constituency, or does it represent the universal direction in which the entire financial infrastructure is evolving? The answer here rests on evidence and analytical frameworks, not on vision.

This chapter advances five named theoretical concepts to organize its argument. The institutional convergence thesis holds that traditional finance and on-chain finance are moving toward each other, driven by the competitive compression of the trust tax. The trust-tax compression curve shows that the difficulty of moving different financial func-

tions on-chain runs along a gradient from the easy-to-compress to the hard-to-compress, which explains why stablecoins moved on-chain before Treasuries and Treasuries before equities. The stablecoin settlement-layer migration thesis argues that the $33 trillion in annual settlement volume marks a structural migration of the settlement layer. The three-horizon framework divides the evolution of on-chain finance into three overlapping horizons with fundamentally different levels of maturity and standards of judgment. The impossibility boundaries identify three permanent structural constraints—the sovereignty boundary, the physical-world interface, and the governance paradox—that separate obstacles which effort can overcome from constraints that must be lived with, establishing the mature epistemological stance of *maximizing within constraints*. Together, these five concepts give the reader a complete analytical toolkit for judging the real prospects of on-chain finance, and they show how the economic logic behind perpetual futures (the most successful product of the first horizon) extends to broader financial functions.

3.1 An economic reexamination of the four distinctive properties

Four pain points pervade traditional financial markets: barriers to entry, the fragmentation of liquidity along the time dimension, persistent counterparty risk during settlement, and the high cost of manual execution. These do not exist in isolation; they are interwoven, mutually reinforcing systemic problems. Each of the four distinctive properties of digital assets responds to one of them. Figure 3-1 maps this full relationship, from pain point to distinctive property to value creation. The mapping is not a simple one-to-one correspondence, because the four properties reinforce one another across categories: programmability combined with atomic settlement gave rise to flash loans, and permissionless access combined with around-the-clock trading makes the global liquidity rotation of perpetual futures possible.

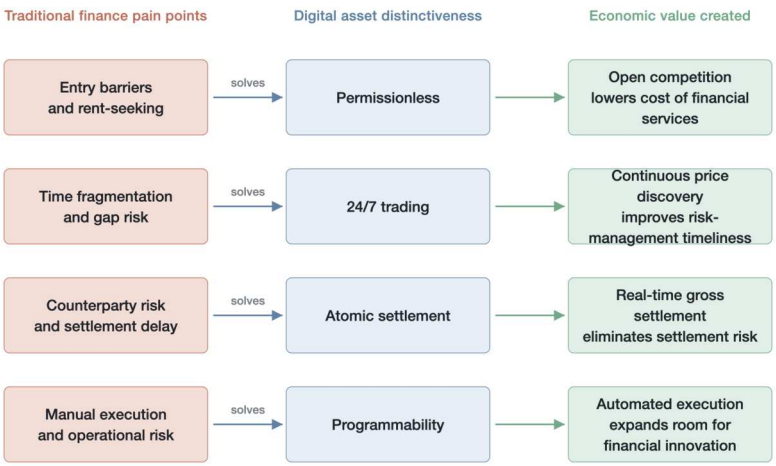


Figure 3-1. The correspondence between the four pain points of traditional finance and the four distinctive properties of digital assets

3.1.1 Permissionlessness and contestable markets

The traditional financial system is, at its core, a finely engineered permission-based structure. A company seeking to raise capital through a listing, an innovator offering financial services, and a trading firm hoping to become an exchange member must each pass through a complex admission process built from regulatory approval, capital requirements, industry relationships, and technical standards. This permission regime provides a foundational framework for financial stability and investor protection, but its high thresholds also, inevitably, breed muted competition and pervasive rent-seeking. The steep underwriting fees that investment banks charge during an initial public offering (IPO) stem in part from the limited number of banks able to win underwriting mandates; becoming a designated market maker on the New York Stock Exchange requires millions of dollars in capital and months of approval.

This phenomenon can be explained through the theory of contestable markets. The central proposition of Baumol, Panzar, and Willig (1982) [10] is that in a market with very low sunk costs, even a single incumbent behaves as it would under perfect competition; it is the persistent threat of potential competition, rather than the number of actual competitors, that forces the incumbent to keep prices near marginal cost. Traditional financial markets are a textbook case of non-contestable markets: becoming a regulated bank or exchange requires enormous outlays to obtain licenses, build infrastructure, and satisfy compliance requirements, and because most of these costs are sunk, they confer substantial market power on incumbents. Shepherd's (1984) [11] critique of contestable-market theory notes that the conditions for a purely contestable market are extremely difficult to satisfy in full, and that brand effects, user inertia, and hidden switching costs often form barriers to entry that Baumol's model does not adequately account for. This critique matters equally for assessing the contestability of on-chain finance: although decentralized exchanges (DEXs) eliminate the sunk costs of licensing and technical infrastructure, liquidity depth itself is a hidden barrier, and a

new protocol must overcome the cold-start problem of liquidity before it can compete effectively. On-chain financial markets are therefore more contestable than traditional finance, but they are not frictionless, perfectly contestable markets.

This cost differential can be quantified. According to industry estimates from firms such as Autonomous Research and Oliver Wyman, a traditional licensed exchange faces regulatory compliance costs of roughly $50 million, technical infrastructure costs of roughly $20 million, and product-listing costs of roughly $5 million, for total sunk costs exceeding tens of millions of dollars [12]; by contrast, deploying a simple automated market maker (AMM) protocol on a public blockchain (such as a Uniswap V2 fork)—including development, audit, and deployment gas fees—typically costs from a few thousand to a few tens of thousands of dollars, with an exit cost close to zero. A fully featured order-book DEX (with a matching engine, risk-control system, liquidation engine, and its own consensus layer, such as Hyperliquid or dYdX v4) costs millions to tens of millions of dollars to develop, markedly less than a traditional exchange, though the gap narrows to one or two orders of magnitude.

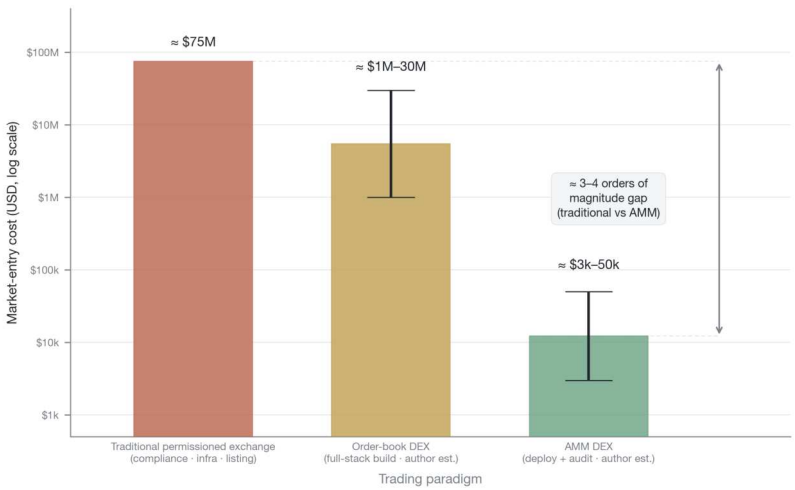


Figure 3-2. A comparison of the market-entry (sunk) cost magnitudes of three trading paradigms (logarithmic scale; a traditional licensed exchange runs to roughly $75 million, of which regulatory compliance is about $50 million, technical infrastructure about $20 million, and product listing about $5 million, based on industry estimates from Autonomous Research, *The Cost of Running a Regulated Exchange*, 2023, and Oliver Wyman, 2023 [12]; the order-book DEX figure of roughly $1 million–$30 million and the AMM DEX figure of roughly $3,000–$50,000 are the author's order-of-magnitude estimates based on public data; the gap between traditional and AMM exchanges is roughly 3–4 orders of magnitude; this figure is an order-of-magnitude illustration rather than precise accounting, mixing empirical data with the author's estimates)

Figure 3-2 reveals a fact easily underestimated on a linear scale: for AMM-type DEXs, the cost difference along the two dimensions of regulatory compliance and technical infrastructure is not a marginal gap of a few tens of percent but a structural chasm of 3–4 orders of magnitude; for high-performance order-book DEXs, the gap narrows to 1–2 orders of magnitude but remains structural. It is this order-of-magnitude difference that fundamentally alters the contestability of the market, expanding the pool of potential

competitors from a handful of licensed institutions to a far broader set of technical teams worldwide.

Permissionless finance built on public blockchains satisfies, in theory, the conditions for a contestable market almost perfectly. The direct cost of deploying a smart contract is minimal and involves almost no sunk cost; open-source code sharply lowers the technical barrier; and the exit cost is likewise zero. This low-sunk-cost property means that any protocol attempting to charge excessive fees risks being forked, as competitors can copy its open-source code and deploy a functionally identical version at lower fees. The effectiveness of fork competition, however, is constrained by brand trust, developer community, and liquidity inertia. The "vampire attack" that SushiSwap launched against Uniswap in 2020 drew substantial liquidity in the short term, but after Uniswap counterattacked by issuing the UNI governance token, it retained its market dominance over the long run. Even in a zero-sunk-cost environment, then, sustained innovation and brand trust still form a substantive moat. Hyperliquid's rapid growth is a direct case study in this economic logic. This perpetual-futures order-book DEX, built on its own application-specific chain, lets anyone list new perpetual-futures trading pairs through governance proposals at very low cost; its cumulative trading volume surpassed $3 trillion by the end of 2025, its user base grew to more than 1.4 million, and the number of listed perpetual-futures pairs exceeded 200 [13][14]. Listing a new futures product on a traditional exchange often takes months or even years of internal review and regulatory dialogue, whereas the permissionless mechanism lets the market respond quickly to new assets and new trading demand. Hyperliquid's permissionlessness appears at the user-access level (anyone can register and trade), but its infrastructure layer remains markedly centralized: its validator set consists of roughly 16–20 nodes, most operated by the team and its affiliates; order matching runs on a centralized sequencer at millisecond latency; and geofencing is enforced through front-end IP detection, which a virtual private network (VPN) can bypass. This architecture—permissionless at the user layer, only partly decentralized at the infrastructure layer—is common to today's high-performance order-book DEXs, whose degree of decentralization is far lower than that of general-purpose public blockchains such as Ethereum.

While the permissionless property lowers the cost of innovation, it also lowers the cost of fraud. According to CertiK, crypto-asset losses from security incidents totaled roughly $3.35 billion over the course of 2025 [15], and losses from scams and fraud ran higher still, into the tens of billions of dollars. The net social benefit of permissionlessness ultimately depends on the maturity of supporting institutions—that is, on whether effective disclosure mechanisms, decentralized reputation systems, and investor-protection tools can be established.

3.1.2 Around-the-clock continuous trading and the transformation of constraints

Traditional financial markets run on clocks and calendars. When markets are closed, information keeps being produced but cannot be priced immediately; it accumulates until the next session opens, when it is released all at once as an opening gap. These gaps interrupt the management of risk exposure and leave capital idle while the market

is closed. The deeper economic value of around-the-clock, uninterrupted trading lies in continuous price discovery: it transforms what is a hard constraint in traditional markets, the inability to trade when the market is closed, into a soft constraint in which only trading costs vary across time periods. Trading by participants around the world produces a cyclical rotation of liquidity across the global day.

This transformation from hard constraint to soft constraint is not costless. Around-the-clock markets do not eliminate differences in liquidity; they convert those differences from a binary state (open versus closed) into a continuous but uneven distribution. The data show that liquidity in crypto markets peaks during the overlap of the European and U.S. morning sessions (roughly 14:00–18:00 UTC) and falls to a trough after the U.S. market closes, with order-book depth differing by as much as 40% between peak and trough [16]. The trading cost of large orders therefore varies systematically by time of day. For market makers, the absence of a close means they cannot recalibrate inventory risk within a protected window, so they must maintain around-the-clock operational capacity or rely on algorithmic risk-management systems that run continuously without human supervision.

Chapter 2 argued that the high leverage of perpetual futures makes positions extremely sensitive to price volatility. Were there closed sessions, traders could not manage their risk exposure while the market was shut, which could produce the extreme case of an *opening-gap liquidation*. Around-the-clock trading is thus a necessary precondition for the perpetual-futures funding-rate mechanism to work: if the market did not run continuously, the funding-rate calculation would be interrupted and its anchoring effect would weaken markedly. More concretely, a 10x-leveraged position held through an overnight close would face unmanageable gap risk, and the funding rate's continuous anchoring would break, producing persistent basis deviation. This is the fundamental reason perpetual futures could not have emerged in a market with trading-hour restrictions. The three implicit protective functions that market closure provides in traditional markets, analyzed in Chapter 1—a risk-repricing window for market makers, a cooling-off period for panic, and a maintenance window for the system—all disappear under around-the-clock trading.

Blonien and Ober (2024) [17] point out that for smaller markets, concentrated trading sessions may be more efficient than around-the-clock trading because they pool liquidity. The value of around-the-clock trading thus depends on market size and liquidity depth; it is not a universally superior solution. Traditional markets are exploring extended trading hours, and institutions such as DTCC and BlackRock are discussing a transition toward around-the-clock operation—itself one dimension of institutional convergence.

3.1.3 The scope of atomic settlement

In traditional financial markets, the execution and settlement of a trade are two separate processes, typically requiring two business days to complete. This time window builds in a structural weakness: counterparty risk. Any party's default within the settlement window can trigger a chain reaction, producing a systemic settlement cascade. The collapse of Lehman Brothers during the 2008 financial crisis was the starkest embodiment of this risk. The atomic-settlement property of digital assets uses cryp-

tography to guarantee that the delivery of an asset and the payment of funds occur simultaneously within a single, indivisible on-chain transaction, fundamentally eliminating counterparty risk during settlement.

Treating atomic settlement as a fundamental solution to systemic risk is an oversimplification. To understand its true effect, one must draw a distinction that shapes the conclusions of any risk assessment.

A settlement cascade is triggered by the settlement failure of a trading counterparty and propagates downstream as a chain of defaults through the breaking of a trust chain; a liquidation cascade, by contrast, is triggered by a decline in asset prices and forms a negative feedback loop of price decline–liquidation–further decline through the forced closure of leveraged positions. Atomic settlement effectively eliminates the former, but for the latter it cannot prevent the cascade and may even accelerate its propagation through efficient execution. Table 3-1 compares the two risks along four dimensions: triggering mechanism, propagation path, the effect of atomic settlement, and time scale.

Risk type	Trigger	Propagation mechanism	Effect of atomic settlement	Time scale
Settlement cascade	Settlement failure of a trading counterparty	A broken trust chain triggers a chain of defaults	Effectively eliminated: simultaneous delivery and payment removes the possibility of default	Traditional finance: days to weeks
Liquidation cascade	A decline in asset prices triggers forced closures	Forced liquidation of leveraged positions intensifies selling pressure in a negative feedback loop	Cannot be prevented and may even be accelerated: efficient liquidation execution compresses propagation from days to minutes	Instant mark-price liquidation architecture: seconds to minutes; auction-based liquidation architecture: minutes to hours; insurance funds and auto-deleveraging (ADL) mechanisms can act as propagation dampers

Table 3-1. A comparison of settlement cascades and liquidation cascades—the scope of atomic settlement (Data source: compiled by the author)

Figure 3-3 contrasts the divergent paths of the settlement cascade (linear propagation) and the liquidation cascade (circular propagation), along with the scope of atomic settlement.

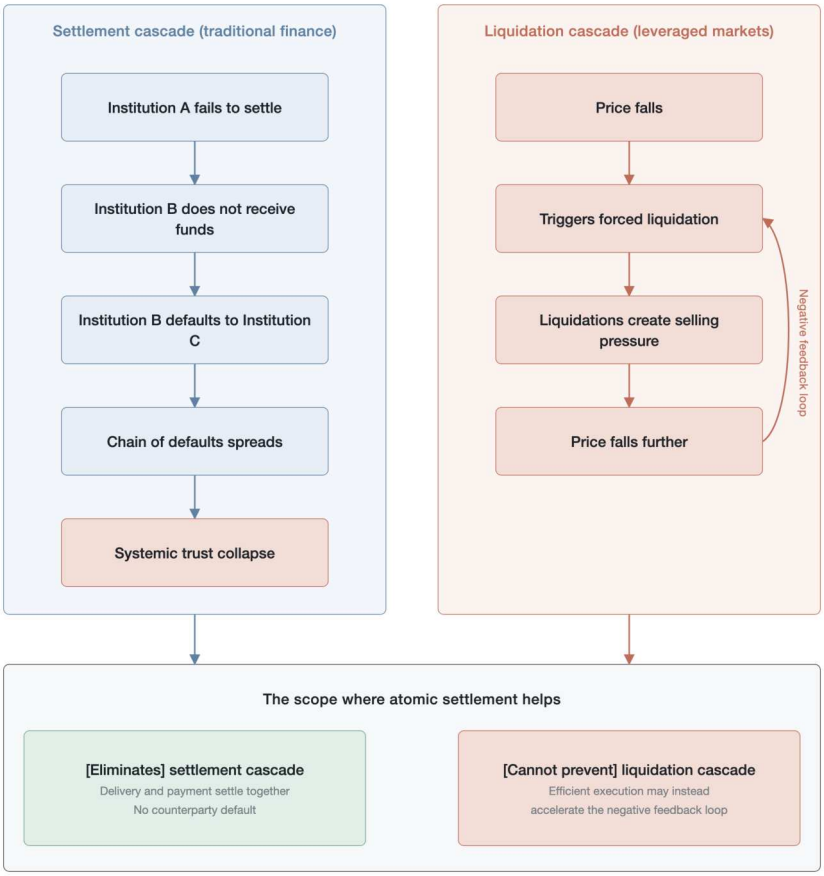


Figure 3-3. The propagation paths of settlement cascades and liquidation cascades and the scope of atomic settlement

In its 2024 report, the Financial Stability Board noted that tokenization can reduce counterparty risk through atomic settlement but may increase market participants' demand for liquidity [18]. This is a textbook risk-surface transformation—converting settlement risk into liquidity risk. Atomic settlement shifts the focus of risk from whether the counterparty will default to whether the market can withstand a rapid liquidation cascade and whether the protocol code is correct—a concrete mapping of the risk-surface transformation framework introduced in Chapter 1, namely that technological change does not eliminate the total quantity of risk but converts one set of risks into another of a different nature.

The crypto-market crash of March 12, 2020, provides an empirical case of a liquidation cascade. BitMEX, a centralized exchange, went down repeatedly during the event (the exchange attributed this to a distributed denial-of-service (DDoS) attack that overwhelmed its API and database layers, though it initially and mistakenly described it as a hardware failure), showing that even in a traditional, non-atomic settlement architecture, infrastructure failure during a surge in liquidations is already a systemic threat. In the same event, MakerDAO suffered zero-bid liquidations of its Vaults (triggered when Ethereum network congestion delayed oracle price updates) for a loss of roughly

$8 million, illustrating how underlying network congestion in an on-chain environment can intensify the propagation of a liquidation cascade. The collapse of FTX, in turn, revealed another layer of problems with the "atomic settlement" of centralized exchanges: the "instant" fills that users saw were merely digital changes in a private database, and when such trust is abused, so-called atomic settlement loses any substantive meaning. By contrast, in an on-chain order-book DEX the settlement of each trade is guaranteed by publicly verifiable cryptographic consensus, a stronger trust guarantee than a centralized exchange can offer. Chapter 1 defined this trust model—based on cryptography and protocol design rather than institutional reputation—as *architectural trust*.

3.1.4 The power and cost of programmability

The operation of traditional finance depends heavily on legal contracts and manual execution; the pricing, execution, and settlement of complex derivatives require large back-office staffs and multiple layers of intermediaries, a model that is both costly and rife with operational risk. Smart contracts—which Szabo (1997) [19] described as a set of promises, specified in digital form, including the protocols within which the parties perform on these promises—encode the terms and execution logic of a financial agreement directly into computer programs that run on a blockchain, replacing informal natural language with deterministic computational logic and labor-intensive fulfillment with automated execution. The perpetual-futures funding-rate mechanism analyzed in Chapter 2 is a textbook success of programmable finance, using automated, code-based payment flows to keep the derivative's price tightly anchored to the spot price. Programmability expands the design space of financial innovation; AMMs and decentralized lending are successful products of programmability interacting with the other distinctive properties, whereas algorithmic stablecoins (in particular the May 2022 collapse of UST/LUNA, which wiped out roughly $40 billion in total ecosystem value, of which the UST stablecoin component accounted for roughly $18 billion) warn that programmability can just as readily create mechanisms with endogenous fragility. The design space that programmability opens up contains both fertile and dangerous regions; the composability thesis of Chapter 1 finds continuing validation here, though it must now be qualified by careful design evaluation.

The value of programmability lies less in automation than in creating entirely new financial mechanisms with no counterpart in traditional finance. An AMM replaces the centralized order-book matching engine with a mathematical pricing function, giving any token pair a liquidity market without any market maker actively quoting prices. The flash loan permits uncollateralized borrowing within a single atomic transaction—an instrument that is logically impossible in any non-atomic settlement system. These innovations are not the result of adapting traditional financial products to a blockchain; they are native products of the programmable environment and cannot exist apart from it.

The cost of programmability manifests at three levels. First, determinism means rigidity. The incomplete-contracts theory of Hart and Moore (1999) [20] holds that no contract can exhaust all contingent states, and that residual control rights are therefore needed to address unforeseen situations. Smart contracts push this theoretical difficulty to

an extreme: the deterministic execution of code precludes any possibility of discretion, and situations the designer failed to foresee cannot be handled flexibly. In March 2023, Euler Finance suffered a flash-loan attack for a loss of $197 million; the attack was executed entirely within the bounds the code permitted, and the vulnerability was not a program defect in the traditional sense but a logical edge case the designers failed to anticipate. Attack losses over the course of 2025 reached $3.35 billion [15], of which the single Bybit incident accounted for $1.45 billion. These losses stem not only from code vulnerabilities but also from economic attacks that turn the deterministic nature of smart contracts against the designers' intent. Second, the cost of security audits constitutes a new form of trust tax; replacing legal compliance with code audits is not costless.

The third and more fundamental problem is one of governance: who has the right to modify the code itself requires an institutional answer. The ideal of "code is law" cannot entirely escape human governance. Contracts form a spectrum, from fully immutable (no one can modify them) to the upgradeable-proxy model (an administrator can change the logic), a continuous space of choices. This spectrum embodies a fundamental tension: immutability gives users confidence that the rules will not change, but upgradeability is a necessary condition for fixing vulnerabilities and adapting to new conditions. Between full immutability and full upgradeability, every protocol must choose a position, and that choice is essentially a governance decision rather than a technical one. Chapter 30 discusses design approaches for embedded compliance in detail.

3.2 The institutional convergence thesis

Institutional convergence is a classic research topic in comparative political economy. The institutional-isomorphism theory of DiMaggio and Powell (1983) [21] identifies three mechanisms that drive organizational convergence: coercive isomorphism (regulatory pressure), mimetic isomorphism (imitative behavior under uncertainty), and normative isomorphism (the diffusion of professional standards). This section applies the framework of institutional convergence to the relationship between traditional finance and on-chain finance and identifies a fourth mechanism not covered by DiMaggio and Powell: the competitive compression of the trust tax. Traditional finance and on-chain finance are moving from mutual opposition toward functional convergence, a convergence driven not only by regulatory pressure (coercive isomorphism) and industry imitation (mimetic isomorphism) but, more fundamentally, by the competitive compression of the trust tax—that is, as the two compete for the same user base, each compresses its own trust tax from a different direction. Traditional finance adopts on-chain technology to compress its trust tax, while on-chain finance introduces institutional elements to raise its level of trust. The BUIDL fund is the landmark case of convergence proceeding in both directions at once.

3.2.1 The on-chain deployment of traditional finance

The move on-chain by traditional financial institutions has moved from the 2020–2022 proofs of concept into a phase of strategic deployment in 2024–2026. This is not incre-

mental exploration but a systematic strategic turn.

Table 3-2 surveys the on-chain strategic actions of six representative institutions. These institutions span six domains—asset management, securities settlement, sovereign-level issuance, industrial-enterprise financing, global payments, and sovereign regulation—and the significance of their collective turn lies in this: the adoption of on-chain finance is no longer confined to crypto-native institutions but has extended to core participants in the traditional financial system.

Institution	Action	Signaling significance
BlackRock	BUIDL fund—a tokenized money market fund on Ethereum, with AUM near $3 billion	The world's largest asset manager chose a public blockchain rather than a private one
DTCC	Exploring T+0 atomic settlement, the Smart NAV project [2]	Core U.S. financial infrastructure exploring on-chain settlement
European Investment Bank	Issued a €100 million digital bond on Ethereum [3]	A sovereign-level institution chose a public blockchain for issuance
Siemens	Issued a €60 million digital bond on a public blockchain [4]	A European blue-chip industrial enterprise adopted on-chain issuance
Visa and Mastercard	Stablecoin settlement integration; Visa tested USDC settlement on Solana [5]	Global payment networks view on-chain settlement as an efficiency gain
Monetary Authority of Singapore	Project Guardian—institutional-grade DeFi	A regulator proactively driving rather than reactively responding

Table 3-2. The strategic on-chain deployment of traditional financial institutions (Data source: public disclosures from each institution [1][2][3][4][5][6])

What drives these institutions is not technological idealism but rigorous cost-benefit analysis. Their core economic logic is this: when the compression of the trust tax exceeds the cost of migration, migration is rational. Each year, DTCC and its participants incur settlement costs measured in the billions of dollars (including reconciliation, dispute resolution, margin management, and the maintenance of multiple intermediary layers [22]); if atomic settlement can eliminate most of that, migration is not innovation but cost optimization. BlackRock's BUIDL fund turns money market fund subscriptions and redemptions from T+1 into instant, around-the-clock operations, and turns fund shares from a closed account system into globally tradable tokens. Most of the on-chain experiments of 2020–2022 took place on private or consortium chains, signaling that these institutions acknowledged the value of blockchain technology but did not trust public blockchains; the strategic deployment of 2024–2026 has shifted markedly toward public blockchains, where the signal has become that the security, transparency, and network effects of public blockchains now exceed the controllability advantages of private ones. Traditional finance is not merely adopting on-chain technology; it is accepting the architectural-trust paradigm of on-chain finance.

3.2.2 The institutionalization of on-chain finance

Convergence is bidirectional. On-chain finance is undergoing systematic institutionalization, moving from a state that lacked any regulatory framework toward compliant innovation within one. The European Union's Markets in Crypto-Assets (MiCA) regulation came fully into force in 2024, and as of April 2025, 17 crypto-asset service providers had been authorized by their respective national competent authorities (NCAs) and entered into the ESMA public register [23][24]. Hyperliquid has introduced partial compliance mechanisms, dYdX has set up geofencing to restrict users in specific jurisdictions, and Coinbase, as a public company, has obtained comprehensive regulatory licenses. The infrastructure layer of on-chain finance is upgrading from crypto-native to institutional-grade: custodians such as Fireblocks and Anchorage provide multi-signature wallets and insurance coverage, while on-chain analytics platforms such as Chainalysis and TRM Labs provide anti-money-laundering and transaction-monitoring services to financial institutions.

The governance of on-chain finance is evolving from the early ideal of "code is law, no human intervention" toward a more mature, multi-layer architecture: smart contracts define and automatically enforce core rules, a governance committee deliberates on parameter adjustments, an emergency multi-signature executes interventions of limited scope in extreme crises, and a time-lock provides a community-review window for all non-emergency changes. The March 2025 JELLY incident at Hyperliquid, in which validators voted to force the delisting of a token, is a practice of exactly this hybrid governance; although it sparked wide debate over the degree of decentralization, it showed that institutionalized human intervention is a necessary condition for the system's survival [25]. What drives on-chain finance toward institutions is not regulatory pressure but the demands of scale: to expand from serving millions of crypto-native users to serving billions of traditional-finance users, it must provide legal recourse, investor protection, compliance frameworks, and professional-grade security guarantees.

3.2.3 The competitive compression of the trust tax

The previous two sections described each one-directional movement separately: traditional finance toward the chain, and on-chain finance toward institutions. Recasting the trust-tax concept from Chapter 1 (all the direct and indirect costs paid in a financial transaction to establish and maintain trust: intermediary fees, compliance costs, the opportunity cost of settlement delay, and the information-asymmetry premium) from static to dynamic, competitive analysis unifies these two directions within a single economic framework. The logic here is deductive rather than inductive: when two financial-service systems offering overlapping functions compete for the same user base, the transaction-cost theory of Coase (1937) [26] and the governance-structure analysis of Williamson (1985) [27] predict that users will flow toward the system with lower transaction costs, and the resulting competitive pressure forces both systems to compress their own trust tax. Deriving convergence—rather than one side replacing the other, or a market split—from transaction-cost competition, however, requires an additional condition: that each system possess an irreplaceable comparative advantage along certain dimensions. The institutional trust of traditional finance (legal recourse,

regulatory protection, deposit insurance) and the architectural trust of on-chain finance (transparency, composability, around-the-clock operation) cannot at present be fully replicated by the other, which provides a structural basis for convergence rather than replacement. Traditional finance and on-chain finance are in the midst of just such a competitive compression of the trust tax, converging from different directions toward the same equilibrium range. Figure 3-4 plots time on the horizontal axis and trust tax or trust level on the vertical axis, showing the declining trust-tax curve of traditional finance and the rising trust-level curve of on-chain finance converging into an equilibrium range of functional convergence.

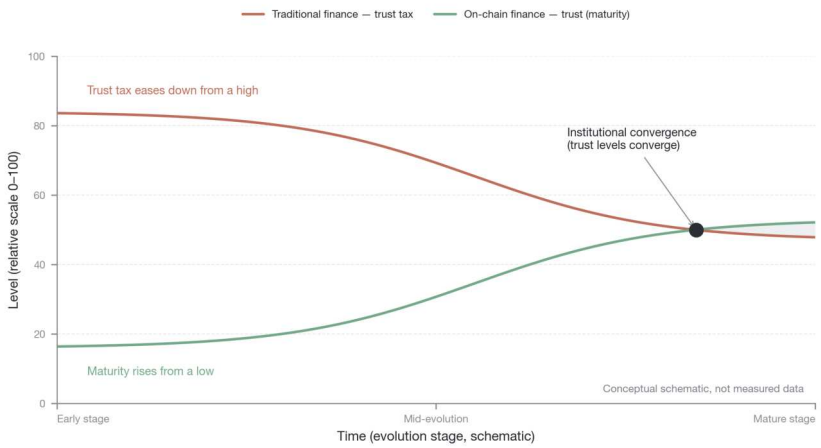


Figure 3-4. Institutional convergence—the converging trend of the declining trust tax of traditional finance and the rising trust level of on-chain finance (drawn by the author)

The two sides pursue different compression strategies. Traditional finance introduces on-chain technology to compress its own trust tax: DTCC explores atomic settlement to compress the time tax, BlackRock uses tokenization to compress the intermediary tax, SWIFT pilots blockchain to compress the jurisdiction tax, and traditional exchanges extend trading hours to compress the time-fragmentation tax. On-chain finance introduces institutional elements to raise its own level of trust: Hyperliquid adds a market-making vault as protocol-level insurance, Ethena obtains traditional-finance custody to strengthen architectural trust, and Circle obtains an electronic-money license that upgrades USDC from a crypto token to regulated electronic money. Both are moving toward the same destination—a lower trust tax and a higher level of trust—and that is convergence.

Once set in motion, this competitive process is hard to reverse. Competitive pressure makes retreat economically infeasible: once BlackRock has lowered its operating costs with a tokenized fund, its competitors must follow or lose market share on price. The on-chain infrastructure already built represents tens of billions of dollars in investment; this investment will not be scrapped but will continue to generate user inertia and network effects. Tens of thousands of engineers have mastered the on-chain finance technology stack, and thousands of institutions have built on-chain operational capacity; this human capital and organizational capability are likewise irreversible. Yet the

irreversibility of convergence is not absolute. At least three scenarios could slow or partly reverse it: a systemic crisis of trust triggered by a large-scale security incident (such as confirmation that a stablecoin issuer's reserves are insolvent), a coercive ban by a sovereign state (such as a major economy comprehensively prohibiting financial activity on public blockchains), and a collapse of institutional confidence triggered by a governance failure internal to on-chain finance. In 2022, the Australian Securities Exchange (ASX) abandoned its blockchain settlement project, writing off more than A$250 million—a reminder that institutional adoption follows a narrative cycle, and that enthusiasm for a blockchain strategy can contract sharply during a phase of disillusionment. In a more extreme scenario, the infrastructure already built would not physically vanish, but the user activity and capital allocation atop it could flow back to the traditional system on a large scale. A more prudent formulation is therefore this: under current institutional conditions and the current trajectory of technological development, convergence is highly path-dependent and self-reinforcing, and its reversal would be extremely costly—but not logically impossible. The end state of convergence is more likely to be the optimal combination of the three trust paradigms described in Chapter 1: selecting, at each layer, the trust paradigm best suited to that layer's needs.

3.2.4 The institutional-convergence significance of the BUIDL fund

The significance of BlackRock's BUIDL fund lies not in BlackRock's stature but in its simultaneous demonstration of convergence in both directions, which makes it the most complete single piece of evidence for the institutional convergence thesis. In the direction of traditional finance moving on-chain, BUIDL chose the Ethereum public blockchain rather than a private one, adopted the ERC-20 token standard so that fund shares can be used in any compatible Ethereum application, partnered with Circle to enable instant, around-the-clock subscription and redemption in USDC, and let BUIDL serve as collateral in DeFi protocols [1]. In the direction of on-chain finance moving toward institutions, BUIDL strictly complies with U.S. securities law, is registered with the SEC, is open only to accredited investors, has its assets held in custody by BNY Mellon, and uses Securitize as the tokenization platform providing compliant on-chain securities issuance and management.

BUIDL introduced a key resource that on-chain finance had previously lacked: high-quality, low-risk collateral that generates a stable yield. Before BUIDL, DeFi protocols relied mostly on highly volatile crypto assets as collateral, which built in enormous systemic risk: when the crypto market fell as a whole, collateral values dropped and triggered liquidations, those liquidations triggered further selling pressure, and the reflexive downward spiral analyzed in Chapter 1 took hold. BUIDL gave on-chain finance, for the first time, low-volatility, high-quality collateral anchored to U.S. Treasury yields. As an ERC-20 token, BUIDL is inherently composable and can serve as Aave collateral, perpetual-futures margin, and more (this composability derives from a shared settlement layer, the mechanism of which is discussed in Section 3.4.3).

BUIDL's composability is also subject to three practical limitations. The first is liquidity: the depth of BUIDL's on-chain secondary market may be far below what its

AUM implies; when BUIDL tokens must be sold quickly during a liquidation event, the on-chain market may be unable to absorb them, forcing a discounted sale—the classic traditional-finance problem of using an illiquid asset as collateral. The second is timing: BUIDL subscription and redemption are available only on business days, creating a structural mismatch with the margin needs of around-the-clock perpetual-futures trading, just as the USDC depegging event analyzed in Section 3.4.4 reveals the break between on-chain 24/7 operation and off-chain 5×8 banking hours. The deepest effect is systemic: once tokenized Treasuries are embedded at scale in DeFi, volatility in the U.S. Treasury market (such as the sharp decline in Treasury prices during the rapid rate-hike cycle of 2022) transmits directly to on-chain finance through the collateral-value channel, opening a cross-system risk-transmission path that did not previously exist and leaving on-chain finance no longer independent of the interest-rate risk of the traditional financial system.

Moreover, BUIDL's composability is *permissioned*, not the permissionless, fully open composability of native DeFi tokens. As an SEC-registered security, BUIDL uses Securitize's whitelist mechanism (a transfer restriction) to limit on-chain transfers to accredited investors, and an address not on the whitelist cannot hold BUIDL tokens. Regulators have not yet clearly tested the boundaries of this restriction within a composable DeFi environment: when BUIDL serves as collateral for a lending protocol, whether the protocol can ensure that all interacting parties meet the accredited-investor requirement—and the SEC's stance on the circulation of tokenized securities within DeFi—remains in flux. This compliance constraint is itself another instance of on-chain finance introducing institutional elements, exactly as the convergence thesis predicts: the technical potential of composability must be realized within an institutional framework.

3.3 The trust-tax compression curve and the three-horizon framework

Institutional convergence does not advance uniformly. The difficulty of moving different financial functions on-chain varies systematically, and this variation can be explained through the trust-tax compression curve. The three-horizon framework, in turn, divides the evolution of on-chain finance into three stages with different levels of maturity and standards of judgment.

3.3.1 The compression-difficulty gradient

Institutional convergence is under way, but it advances unevenly: stablecoin settlement has already reached the tens of trillions, Treasury tokenization is breaking through, and equity tokenization is still nascent. This unevenness is not random but reflects a gradient in the difficulty of compressing the trust tax. Different financial functions have different trust-tax structures. In some, the trust tax is primarily technical and can be compressed with code, because algorithms and cryptography can directly replace what users need to trust; in others, it is primarily institutional and must be maintained through law, regulation, and politics, because the trust involves domains that code

cannot replace—sovereign power, the physical world, and human discretion.

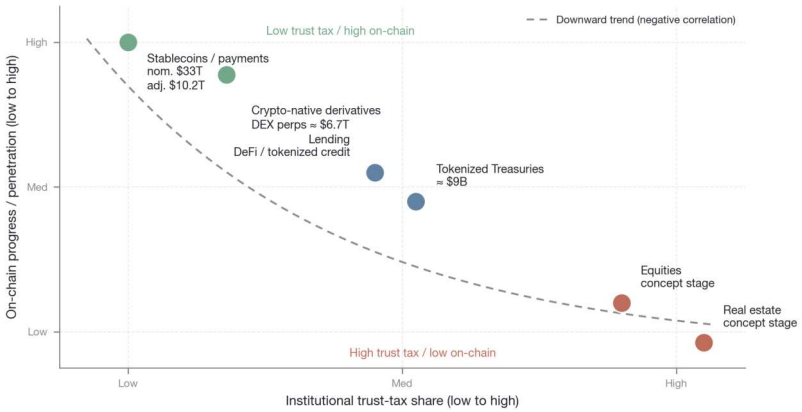


Figure 3-5. The trust-tax compression curve—the higher the share of the institutional trust tax in a financial function, the lower its on-chain progress (the scatter points of the various financial functions in the figure lie along a single negatively sloped trend line and represent the author's empirical judgment and estimates: the share of the institutional trust tax is the author's qualitative assessment of each financial function's trust structure, and the data anchors for on-chain progress are sourced by item—stablecoin nominal and adjusted transaction volumes are from Visa Onchain Analytics [38], data provided by Allium, raw/adjusted figures, accessed June 2026; DEX perpetual-futures trading volume is from CoinGecko [39]; and the size of tokenized Treasuries is from RWA.xyz [28])

In Figure 3-5, the horizontal axis is the share of the institutional trust tax (increasing from left to right) and the vertical axis is current on-chain progress (increasing from bottom to top); six categories of financial function appear as scatter points along a downward-sloping trend line, showing a clear negative correlation. Payments and settlement sit in the upper left, with the lowest share of institutional trust tax and the highest on-chain progress; real estate sits in the lower right, with the highest share of institutional trust tax and the lowest on-chain progress. This inverse relationship is not coincidental but causal: the higher the proportion of a financial function's trust needs that are technically replaceable, the lower the resistance to moving it on-chain and the faster its progress.

Table 3-3 arranges the six categories of financial function along the trust-tax compression-difficulty gradient, revealing a clear pattern: from payments and settlement to real estate, the share of the technical trust tax declines, the share of the institutional trust tax rises, and current on-chain progress correlates positively with the share of the technical trust tax.

Financial function	Share of technical trust tax	On-chain compression difficulty	Current progress	Key bottleneck
Payments and settlement	Very high	Low	$33 trillion in annual stablecoin settlement; escape velocity reached	Anti-money-laundering compliance
Fixed income	Medium-high	Medium	Roughly $9 billion in tokenized Treasuries, with BUIDL in the lead	Adapting the legal framework
Lending	Medium	Medium	Tens of billions of dollars in total value locked (TVL) across Aave, Compound, and others	Credit-assessment systems
Derivatives	Medium	Medium	$6.7 trillion in annual perpetual-futures DEX volume	Regulatory classification and extraterritorial jurisdiction (CFTC swap-jurisdiction claims, extraterritorial application of the Commodity Exchange Act, or CEA)
Equities and equity stakes	Low	High	Very early exploration	Corporate-law reform
Real estate	Very low	Very high	Conceptual stage	Cooperation of the entire legal system

Table 3-3. The trust-tax compression-difficulty gradient—from payments to real estate (Data sources: RWA.xyz [28], CoinGecko [9])

The current progress of on-chain finance unfolds along this curve from left to right—not because payments are simpler than equities, but because the trust tax of payments is primarily technical: correspondent banks, SWIFT nodes, and settlement delays can all be replaced by smart contracts and public blockchains, whereas the trust tax of equities is primarily institutional, and corporate governance, shareholder voting rights, dividend rights, securities registration, and annual audits require the cooperation of the entire legal framework.

This curve explains the past and also predicts the future: the next breakthrough is very likely to come in structured products and complex derivatives, whereas breakthroughs in equities and real estate require not technological progress but fundamental legal and institutional innovation. In 2022, the Boston Consulting Group and ADDX predicted

that tokenized-asset volume could reach $16 trillion by 2030 [29]; McKinsey, in 2024, was more cautious, projecting that tokenized assets (excluding stablecoins and deposits) could reach $2 trillion by 2030, and roughly $4 trillion in an optimistic scenario [30]. The order-of-magnitude difference between the two projections reflects differing assumptions about how fast institutional obstacles will be overcome. Actual progress will more likely depend on the speed of the transition along the curve from fixed income to equities, and the bottleneck of that transition lies in the structural constraints analyzed in Section 3.6: the sovereignty boundary and the physical-world interface.

3.3.2 H1: Stablecoin settlement and crypto-native derivatives

The first horizon comprises those on-chain financial functions that have already reached significant scale, whose trust tax is primarily technical and has been effectively compressed by on-chain infrastructure. The two pillars of H1 are stablecoin settlement and crypto-native derivatives. The total market capitalization of stablecoins is roughly $310 billion, and their annual settlement volume reaches $33 trillion (Section 3.4 analyzes the statistical basis and the economically effective settlement volume in detail); escape velocity has been reached [9]. On the crypto-native derivatives side, perpetual-futures DEXs have annual trading volume exceeding $6.7 trillion, and the entire crypto-derivatives market has annual trading volume of roughly $92.9 trillion, split between centralized exchanges (CEXs) at roughly $86.2 trillion and DEXs at roughly $6.7 trillion. Chapter 2 analyzed the mechanism design of perpetual futures in detail; they are the most successful product in H1 and the central object of analysis throughout the rest of this book.

The total nominal economic activity of H1 products—including the $33 trillion in stablecoin settlement (see Section 3.4.1 for the basis) and the roughly $92.9 trillion in nominal derivatives trading volume (CEXs roughly $86.2 trillion, DEXs roughly $6.7 trillion; nominal volume is amplified by leverage and may include non-organic components such as rebate trading)—is within a comparable order of magnitude to the corresponding products of the traditional derivatives market. Of this, DEX perpetual futures running entirely on on-chain architectural trust contributed roughly $6.7 trillion, and achieving this scale does not rely on traditional financial infrastructure: no clearinghouse, no membership requirement, no central counterparty. The absence of a central counterparty (CCP) eliminates the risk of a centralized single point of failure, but it also means forgoing the three key functions a CCP provides: multilateral netting (which lowers systemic margin requirements), a default fund (which absorbs extreme losses), and an institutionalized default-management process. On-chain DEXs use insurance funds and ADL as substitutes, but the risk-bearing capacity of these mechanisms has not yet been tested under extreme stress with the same systemic intensity as a CCP; the losses that Hyperliquid's insurance vault faced during the March 2025 JELLY incident exposed the limits of these substitutes. Their economic significance lies not in the absolute numbers but in demonstrating that financial infrastructure based on architectural trust can operate at significant scale, even as its risk-management architecture continues to evolve from minimum viable toward systemically robust.

The trust profile of H1 is architectural trust as the mainstay, plus a minimum of in-

stitutional trust: users primarily trust cryptography and smart-contract code, while institutional trust is embodied mainly in stablecoin issuers' reserve management and exchanges' operational security. Its risk profile differs qualitatively from that of traditional finance. Protocol risk (smart-contract vulnerabilities) replaces operational risk (back-office human error), depegging risk (a stablecoin losing its peg) replaces credit risk (a bank's insolvency), and liquidation-cascade risk replaces settlement-cascade risk. Each of these conversions replaces a familiar, well-studied risk with a newer one that is not yet fully understood—a concrete manifestation of the risk-surface transformation analyzed in Chapter 1.

The success of H1 provides the empirical foundation for all of this book's analysis. Perpetual futures, as the highest-volume product in H1, prove that the four distinctive properties can generate a financial instrument superior in specific dimensions to its traditional counterpart, the quarterly future: continuous trading eliminates gap risk, a continuous funding rate achieves sustained price anchoring, permissionless global access broadens the participant base, and atomic settlement eliminates counterparty risk. Chapters 4 through 33 analyze this product's microstructure, risk dynamics, and future evolution in depth.

3.3.3 H2: Tokenized securities and institutional DeFi

The second horizon comprises the on-chain financial functions moving from proof of concept toward large-scale application, whose trust tax is a mixture of the technical and the institutional. Tokenized real-world assets (RWAs, excluding stablecoins) stood at roughly $24 billion to $28 billion by the end of 2025 (depending on the statistical basis, per RWA.xyz) [28] (note: Figure 3-9 ranks by the size of each subsegment, in which private credit uses an active-outstanding basis, so the sum of the subsegments exceeds the distributed total given here). Private credit is roughly $18.9 billion, with platforms such as Apollo and Securitize tokenizing long-illiquid private-credit assets; Treasuries and money market funds are roughly $9 billion, the fastest-growing segment, with BUIDL and Ondo Finance's USDY as representative products. Private credit's dominant position within H2 is no accident: this asset class has historically been plagued by extreme illiquidity, with lock-up periods typically of 3–7 years and a secondary market that barely exists. Tokenization addresses this structural weakness by enabling fractional ownership and around-the-clock secondary trading, lowering the liquidity premium that investors demand and expanding the pool of potential investors from large institutions to a broader group of accredited investors. The collaboration between Apollo and Securitize is a textbook case of this logic: accredited investors can enter a tokenized private-credit fund at a markedly reduced minimum investment, while the fund benefits from more efficient capital formation.

Treasury tokenization occupies a strategically different position. Unlike private credit, U.S. Treasuries already have extremely high liquidity in traditional markets. The value of tokenizing Treasuries lies not in improving liquidity but in achieving DeFi composability: once a U.S. Treasury fund exists in the form of an ERC-20 token, it can serve as collateral in lending protocols, as margin in derivatives trading, and as yield-bearing reserves in stablecoin mechanisms. BUIDL's rapid growth in AUM to near $3 bil-

lion proves that market demand for this composability is substantive. The core thesis of Treasury tokenization is this: the most trusted asset class in traditional finance, sovereign debt, can become the collateral backbone of the entire DeFi ecosystem (the mechanism by which it replaces highly volatile crypto assets and lowers systemic risk is discussed in Section 3.2.4).

The breakthrough of H2 depends more on institutional conditions than on technical capability. The full implementation of MiCA, the U.S. shift from enforcement-driven to legislation-driven regulation, and the regulatory sandbox regimes of Singapore and Hong Kong together turn the compliance risk of institutional participation from unassessable into assessable and manageable [23]. The years 2024–2030 are a critical window: if regulatory certainty continues to improve and early H2 products show reliable operational performance, institutional capital may flow into tokenized assets at a nonlinearly accelerating pace.

3.3.4 H3: Programmable money and autonomous finance

The third horizon comprises domains still at the conceptual and early-exploration stage, whose trust tax is almost entirely institutional. Central bank digital currency is the most direct intersection of programmability and monetary sovereignty. As of 2025, more than 130 countries (the Atlantic Council tracker now lists roughly 146) are exploring central bank digital currency [31], and China's digital yuan had a cumulative transaction value of roughly ¥16.7 trillion by the end of 2025 (¥7 trillion was a mid-2024 milestone) [32]. Its transformative potential lies in programmable monetary policy: targeted stimulus spending can be restricted to specific consumption categories, interest rates can be set differentially by holder type, and taxes can be collected automatically at the transaction level. Yet this programmability also raises fundamental concerns: a currency that the state can program can equally be used for surveillance, censorship, and behavioral control. The design choices around the privacy architecture of a central bank digital currency will define the boundary between efficient governance and authoritarian overreach.

AI agents are evolving from tools that assist human traders into independent entities that autonomously evaluate opportunities, execute trades, and manage portfolios. In 2025, projects such as Virtuals Protocol and Spectral had already deployed AI agents with their own wallets and the ability to trade on-chain. The long-term implications go beyond the automation of trading itself: if AI agents become the primary participants in DeFi markets, the market-microstructure assumptions on which current protocol design rests (human reaction times, behavioral biases, information-processing limits) may need fundamental revision. Chapters 31 through 33 analyze in detail how AI agents interact with perpetual-futures markets.

A cross-sovereign global settlement layer (a single neutral infrastructure serving all value transfer) is the logical endpoint of the stablecoin settlement-layer migration analyzed in Section 3.4. Yet this vision conflicts directly with the sovereignty boundary analyzed in Section 3.6.1. A more realistic direction of evolution is federated global settlement: multiple sovereign jurisdictions each define their own rules—know-your-customer (KYC) requirements, asset restrictions, tax reporting—while sharing a common technical in-

frastructure layer. This model captures the efficiency gains of a unified settlement network while preserving sovereignty. The standards for judging H3 are inherently uncertain: no established metrics for the progress of programmable money or autonomous finance yet exist, and premature quantification would bring only misleading precision.

3.3.5 Differences in the trust needs of the three horizons

The trust needs of the three horizons are fundamentally different and should not be judged by a single standard. Using the logic of "stablecoins have already succeeded, so blockchain will change everything" to infer the prospects of H3, and using the logic of "real-estate tokenization has not yet succeeded, so blockchain is worthless" to deny the achievements of H1, commit the same methodological error: conflating the maturity levels and standards of judgment of different horizons.

Figure 3-6 uses a timeline to show the evolutionary path and differing trust needs of the three horizons (H1, H2, and H3).

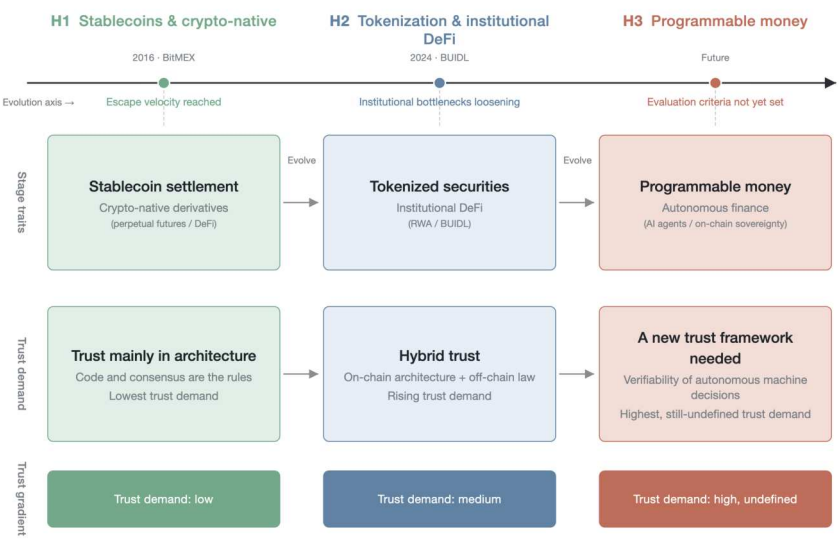


Figure 3-6. The three-horizon framework—the evolutionary path and differing trust needs of on-chain finance from H1 to H3 (drawn by the author)

The differences can be compared along five dimensions: trust paradigm, regulatory need, current scale, standard of judgment, and time scale. Table 3-4 summarizes the structural distinctions among the three horizons along these dimensions.

Dimension	H1: Stablecoins and crypto derivatives	H2: Tokenized securities and institutional DeFi	H3: Programmable money and autonomous finance
Primary trust paradigm	Architectural trust as the mainstay	A mix of architectural and institutional trust	Requires an entirely new trust framework
Regulatory need	Anti-money-laundering and reserve transparency	Securities-law compliance and investor protection	Requires an entirely new regulatory paradigm
Current scale	$33 trillion in annual settlement (roughly $0.4–10 trillion adjusted to a payment basis), roughly $92.9 trillion in nominal derivatives trading (CEX roughly $86.2 trillion, DEX roughly $6.7 trillion)	Roughly $25 billion in tokenized RWAs	Concept and early pilots
Standard of judgment	Trading volume, settlement efficiency, and stability	AUM growth and institutional adoption rate	No established standard yet
Time scale	Scale already reached	Breaking through; 2024–2030 the critical window	Long-term vision, the 2030s and beyond

Table 3-4. A systematic comparison of the trust needs of the three horizons (Data source: compiled by the author)

The most notable feature of Table 3-4 is the difference in the standard-of-judgment row. The success of H1 can be measured directly with hard metrics such as trading volume and settlement efficiency; the progress of H2 can be tracked through AUM growth and the institutional adoption rate; but H3 has not even a commonly accepted standard of judgment, which means any quantitative assessment of it should be treated as highly preliminary. This difference also explains why public discussion of the prospects of on-chain finance so often becomes polarized: optimists extrapolate from the hard data of H1, pessimists dismiss the field on the uncertainty of H3, and the two sides are in fact debating different questions on different horizons.

3.4 Stablecoins: a structural migration of the settlement layer

Stablecoins are the largest and most far-reaching product category in the first horizon of on-chain finance. The scale of their annual settlement volume marks not merely the success of an asset class but a structural migration under way in the settlement layer of the financial system.

3.4.1 The meaning of $33 trillion in settlement volume

As of the end of 2025, the total market capitalization of stablecoins worldwide was roughly $310 billion, with annualized on-chain settlement volume of approximately $33

trillion (Artemis/Bloomberg [7]), exceeding the combined annual processing volume of Visa (roughly $14 trillion; FY2025 payment volume, roughly $17 trillion including cash) and Mastercard (roughly $10.6 trillion) [9][33]. Figure 3-7 compares these magnitudes. The significance of the comparison lies not in the absolute numbers (differences in statistical basis are discussed below) but in the fact that an on-chain settlement infrastructure less than 10 years old already stands, in aggregate terms, within a comparable order of magnitude of the world's most mature payment networks.

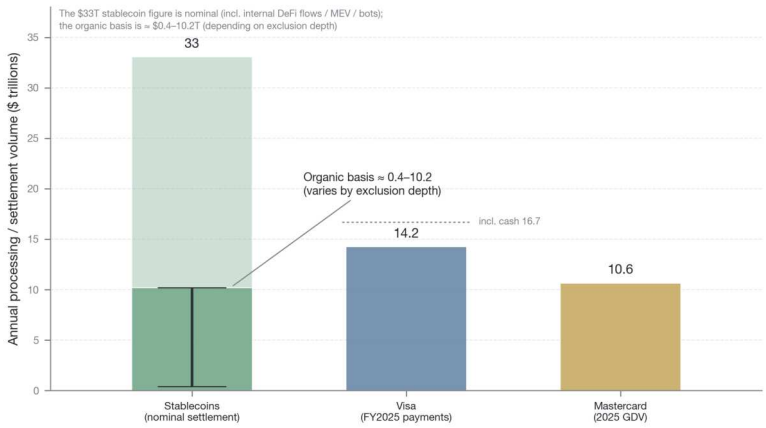


Figure 3-7. A comparison of the annual magnitudes of stablecoins and the card networks (nominal basis vs. organic basis; the nominal volume of stablecoin on-chain transfers for all of 2025 was roughly $33 trillion, up 72% year over year, including non-payment activity such as internal DeFi flows, MEV, arbitrage, and market-making bots, and is not directly comparable to card-network payment volume; the organic basis, after excluding non-payment activity, is roughly $0.4 trillion–$10.2 trillion, with the lower bound being roughly $0.39 trillion in true payments annualized and the upper bound roughly $10.2 trillion in adjusted transfer volume, the width of the range depending on the degree of exclusion, as marked with a range in the figure; Visa's payment volume of $14.2 trillion and Mastercard's gross dollar volume (GDV) of $10.6 trillion are shown for comparison, with Visa's cash-inclusive figure at $16.7 trillion; data sources: Artemis/Bloomberg [7]; Visa Onchain Analytics adjusted basis, trailing 12 months, accessed June 26, 2026; McKinsey × Artemis, annualized as of December 2025; Visa FY2025 10-K, as of September 30, 2025 [33]; Mastercard FY2025 results)

The statistical basis requires clarification: the on-chain settlement volume of stablecoins includes all on-chain transfers, a considerable share of which are internal DeFi flows, MEV arbitrage, and bot loop trades—not fully comparable to the consumer payments of Visa and Mastercard. Estimates after excluding these non-economic transactions diverge widely by method: Visa's on-chain analytics gives roughly $10.2 trillion (adjusted); Chainalysis gives roughly $28 trillion (true economic activity); McKinsey and Artemis estimate that only roughly $0.39 trillion is true real-world payments (roughly 1% of the total, as of January 2026); and the heuristic of Brevan Howard Digital (2024) [8] gives an organic share of roughly 10%. Even at the higher adjusted figure, this scale remains within the same order of magnitude as Visa's annual net payment volume, and its growth rate far exceeds Visa's.

The core economic advantage lies in extremely low marginal cost and global reach. Traditional cross-border payments must pass through multiple correspondent banks layer by layer, taking 3–5 business days at a cost of 5–10% of the transaction amount. On-chain stablecoin transfers, regardless of amount, complete within minutes, at a cost typically under $1 on Layer 2 networks (though the Ethereum mainnet can reach several to tens of dollars during congested periods). This efficiency difference stems from an architectural one; the specific comparison of serial versus parallel trust is detailed in Section 3.5.2.

3.4.2 The network effects and escape velocity of stablecoins

Stablecoins have reached escape velocity. This chapter operationalizes escape velocity as follows: a network has reached it when the marginal user-growth rate driven by network effects persistently exceeds the natural attrition rate, and when growth no longer depends on an exogenous impetus (such as price speculation or subsidy incentives) but has become endogenous and self-maintaining. The network-externality theory of Katz and Shapiro (1985) [34] provides a standard framework for this dynamic: when a user's utility increases as the network grows, a critical size threshold exists beyond which growth becomes self-accelerating. This judgment must withstand cyclical testing. During the 2022 bear market, the total market capitalization of stablecoins shrank from roughly $180 billion to roughly $120 billion (about 33%), showing that growth is not entirely endogenous. But the key distinction is this: the contraction came mainly from the ebbing of speculative demand (the collapse of the algorithmic stablecoin UST alone accounted for roughly $18 billion of it), whereas the cross-cycle trend in settlement volume and active-address count indicates that actual usage demand kept growing. A more stringent test will come after the next complete bear-market cycle: if stablecoin settlement volume and active-user count keep growing through the next downturn, the escape-velocity judgment will gain stronger empirical support. A threefold network effect gives stablecoin growth this self-reinforcing character.

The first is a liquidity network effect. The more traders use USDC and USDT, the deeper the market liquidity and the narrower the bid-ask spread, which in turn attracts more traders, forming a positive feedback loop. This effect explains the extreme concentration in which USDT and USDC hold more than 85% market share. A newly entering stablecoin faces a cold-start dilemma: without liquidity, traders will not adopt it, and without traders, liquidity cannot form. This dynamic gives the current duopoly a self-reinforcing character, much as the joint market position of Visa and Mastercard in card payments has held firm over decades despite persistent competitive entry.

The second is an integration network effect. The more DeFi protocols support USDC, the greater USDC's utility as an on-chain dollar, which in turn attracts more protocols to integrate it. USDC is already natively supported by hundreds of DeFi protocols, serving as the default quote currency in AMM pools, standard collateral in lending protocols, and the primary settlement token in perpetual-futures trading. Each additional integration raises the cost of switching to an alternative stablecoin, producing a standard lock-in effect.

The third is a settlement network effect. The more institutions and individuals settle

in stablecoins, the lower the friction of cross-border settlement, which in turn drives broader adoption. The moves by Visa and Mastercard to integrate stablecoin settlement amplify this effect, connecting on-chain settlement to the vast user base of traditional payment networks. Once escape velocity is reached, growth becomes self-maintaining, no longer dependent on new technological breakthroughs; the network effect itself is the engine of growth.

3.4.3 The upper-layer effects after the settlement layer is established

The financial system is layered, and settlement is the bottommost layer. Once the settlement layer is established on-chain, upper-layer applications naturally migrate to the same layer, because composability works within a single settlement infrastructure but cannot span different ones. When the BUIDL fund exists as an ERC-20 token on Ethereum and settles in USDC, it automatically gains composability with all other Ethereum DeFi protocols: it can serve as Aave collateral, trade on Uniswap, be incorporated into structured products, and serve as margin for perpetual futures. This cross-application interoperability, which would require months of legal negotiation and technical integration in traditional finance, is a natural property of shared settlement infrastructure on-chain.

The upper-layer effect produces a self-reinforcing dynamic. Perpetual-futures DEXs use stablecoins as margin, tokenized Treasuries subscribe and redeem in USDC, DeFi lending protocols denominate in stablecoins—all upper-layer applications already run atop the on-chain settlement layer. As more upper-layer applications accumulate, the settlement layer becomes ever more entrenched: migrating away from it would require migrating every dependent application at once, which grows less and less realistic as the ecosystem expands. This lock-in dynamic explains why the stablecoin settlement-layer migration, once past the escape-velocity threshold, becomes structurally irreversible. To see stablecoins merely as an on-chain dollar substitute is to underestimate their systemic significance: they are the foundation on which the entire on-chain financial technology stack is built.

3.4.4 The structural risks of stablecoins

Stablecoins also carry structural risks. The first is reserve-transparency risk. The controversy over the composition of Tether's reserves has persisted for years. For a financial instrument with a market capitalization exceeding $100 billion, incomplete or delayed reserve disclosure constitutes a systemic risk. Tether has improved transparency by releasing quarterly attestation reports, but a significant gap remains between these reports and the comprehensive, independent, real-time audit that a traditional financial instrument of comparable size would face. This gap between attestation and audit is a residual trust deficit.

The second is depegging risk. The clearest empirical case is USDC's depeg to $0.87 after the collapse of Silicon Valley Bank (SVB) in March 2023. Circle held $3.3 billion at SVB, roughly 8% of its reserves [35]. During the 48-hour depegging period, cascade effects propagated throughout the DeFi ecosystem: the declining value of USDC-denominated collateral triggered liquidations in lending protocols, traders rushing to exit USDC

positions unbalanced DEX pools, and USDC's function as the default settlement token was temporarily disrupted. The collapse of SVB occurred after banks closed on a Friday, and Circle was unable throughout the weekend to confirm the safety of its $3.3 billion in reserves or to process redemptions—exposing a structural operational break: on-chain stablecoins claim to run around-the-clock and instantly, but redeeming their reserves depends on the working hours of traditional banks (Monday through Friday), so depegging risk peaks precisely when the traditional financial system is least available. The depeg was resolved only after the U.S. government guaranteed SVB's depositors, which means USDC's stability ultimately depended on sovereign intervention rather than any algorithmic or architectural mechanism.

The third is systemic concentration risk. USDT and USDC together hold more than 85% market share. The failure of either issuer (whether due to reserve insolvency, regulatory action, or an operational-security incident) could trigger a systemic crisis across the entire on-chain financial system. This concentration creates a single-point-of-failure architecture, at odds with the decentralization ethos of blockchain. Diversification into alternative stablecoins has made limited progress, precisely because of the liquidity network effect analyzed in Section 3.4.2.

The fourth is the effect on monetary sovereignty. The global adoption of dollar-denominated stablecoins in effect extends U.S. monetary influence into economies that may not welcome it. For emerging-market countries with volatile local currencies, the adoption of USDC and USDT provides citizens and enterprises with de facto dollar accounts, potentially weakening the transmission of local monetary policy and capital-account controls. The central bank digital currency programs of several countries are an explicit defensive response to this dynamic. The EU's MiCA sets capital requirements and reserve rules for stablecoin issuers [23]; the United States is advancing stablecoin legislation, bringing stablecoins into the framework of regulated financial infrastructure.

3.5 Asset tokenization and on-chain settlement

The previous two sections made the macro case: the direction of convergence (Section 3.3) and the evidence of settlement-layer migration (Section 3.4). This section turns to the technical path, analyzing the engineering process of asset tokenization and the architectural advantages of on-chain settlement, to give those macro arguments a micro-technical foundation.

3.5.1 The economic logic of asset tokenization

Asset tokenization is the process of cryptographically mapping the ownership of and associated rights in a real-world asset onto a digital token on a blockchain; its core function is to lower transaction costs through standardization, transforming non-standardized, low-liquidity real-world assets into standardized, highly liquid, globally tradable digital-native assets. Chapter 1 answered why digital assets are distinctive; this section answers how to give traditional assets those distinctive properties as well.

The trust-tax compression curve of Section 3.3 showed that tokenization progress correlates positively with the share of the technical trust tax. Tokenization is not a simple

digital mapping but a multi-stage engineering process spanning law, technology, and compliance, and understanding it helps explain why different asset classes tokenize at systematically different speeds. Figure 3-8 shows the full life-cycle process of a tokenized asset, from a traditional asset to on-chain circulation.

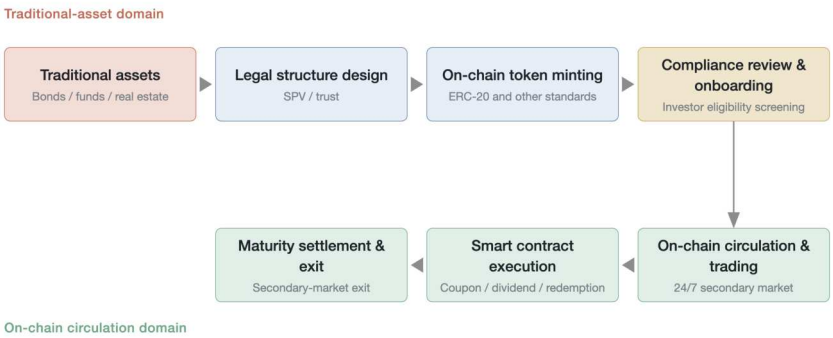


Figure 3-8. The full life-cycle process of a tokenized asset—from a traditional asset to on-chain circulation

The size distribution of tokenized assets is highly concentrated. Figure 3-9 shows that private credit, together with Treasuries and money market funds, constitutes the two largest categories, which together account for the vast majority of total tokenized RWAs; the remaining domains are still below tens of billions of dollars. This concentration reflects the prediction of the trust-tax compression curve: asset classes with a high share of technical trust tax move on-chain first.

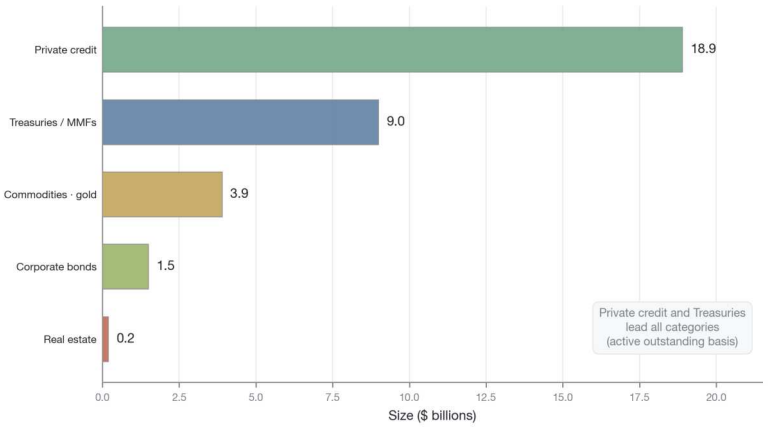


Figure 3-9. A size ranking of the major subsegments of on-chain RWAs (in billions of dollars, as of the end of 2025; private credit leads at roughly $18.9 billion, followed by Treasuries and money market funds at roughly $9 billion, the two together accounting for the vast majority of tokenized RWAs; commodities dominated by gold, corporate bonds, and real estate remain a long tail at tens of billions of dollars and below; private credit uses an active-outstanding basis, dominated by Figure's on-chain home equity line of credit (HELOC) loan book, while the rest use a market-value/distributed basis; data source: RWA.xyz [28])

Table 3-5 summarizes the current progress of on-chain finance in four core domains—

private credit, Treasury funds, sovereign and corporate bonds, and cross-border payments—along three dimensions: size, representative projects, and key advantages relative to traditional finance.

Domain	Size	Representative projects	Key advantages
Private credit	Roughly $18.9 billion	Apollo and Securitize, Centrifuge	Solves poor liquidity, lowers the participation threshold
Treasuries and money market funds	Roughly $9 billion	BUIDL, Franklin Templeton, Ondo	Around-the-clock subscription and redemption, DeFi composability
Sovereign and corporate bonds	Hundreds of millions of euros	European Investment Bank, Siemens	T+0 settlement, lower issuance costs
Cross-border payments	$33 trillion in annual settlement	Circle USDC, Tether USDT	Near-real-time, extremely low cost, globally reachable

Table 3-5. The application progress of on-chain finance in key domains (Data source: RWA.xyz [28])

Table 3-5 reveals a pattern consistent with the trust-tax compression curve: the two largest domains, cross-border payments ($33 trillion in annual settlement) and private credit (roughly $18.9 billion), correspond exactly to the asset classes with a higher share of technical trust tax and an obvious efficiency gap in traditional markets. The pain point of cross-border payments is speed and cost; that of private credit is illiquidity—both domains that code and architecture can directly improve. By contrast, sovereign and corporate bonds remain at the level of hundreds of millions of euros, and their issuance involves more institutional approval steps, confirming the constraint that the institutional trust tax places on on-chain speed.

3.5.2 The architectural advantages of on-chain settlement

On-chain settlement uses a blockchain as the underlying network for value transfer and clearing, on a globally unified, neutral, around-the-clock ledger whose trust is guaranteed by decentralized cryptographic consensus. Traditional payments rest on the serial relay of bilateral trust relationships, in which each additional intermediary node adds a layer of cost; on-chain settlement rests on a single global consensus, in which all participants share the same trust foundation.

Figure 3-10 compares the two settlement paths. In the traditional cross-border payment path, the SWIFT network performs a messaging function, conveying payment instructions rather than moving funds, while the funds themselves move through the nostro/vostro account relationships between correspondent banks. In this multi-node serial relay, each node adds delay and fees, and the failure of any node breaks the entire chain (SWIFT itself is raising payment speed and transparency through innovations such as SWIFT GPI to meet the competition). In the on-chain settlement path, the sender initiates a transfer directly on the blockchain through a wallet, and once blockchain consensus verifies it, the funds arrive at the recipient's wallet—only three

nodes, with the intermediate consensus layer decentralized and free of any single point of failure.

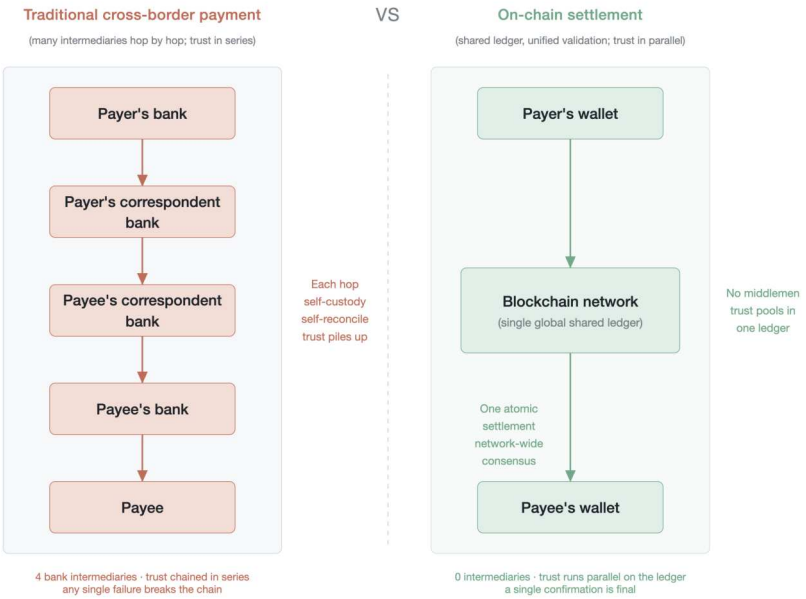


Figure 3-10. A process comparison of traditional cross-border payments and on-chain settlement—the architectural difference between serial trust and parallel trust

Parallel trust is superior to serial trust in both efficiency and resilience. In regions such as Latin America and sub-Saharan Africa, many small and medium-sized enterprises are increasingly adopting stablecoins as their primary means of payment and receipt, cutting the in-transit time of funds from days to minutes and transaction costs by as much as 90%—a cheaper, faster global-payment option for enterprises underserved by the traditional financial system.

3.5.3 Three surmountable challenges

On-chain finance faces three surmountable challenges. They are not independent of one another but form a mutually reinforcing feedback loop. Figure 3-11 uses a triangular structure to depict the bidirectional dependencies among technical maturity, regulatory certainty, and market size. This means that a breakthrough in any single dimension requires the cooperation of the other two to produce a lasting effect; it also means that once one dimension makes substantive progress, a positive feedback loop may accelerate the whole.

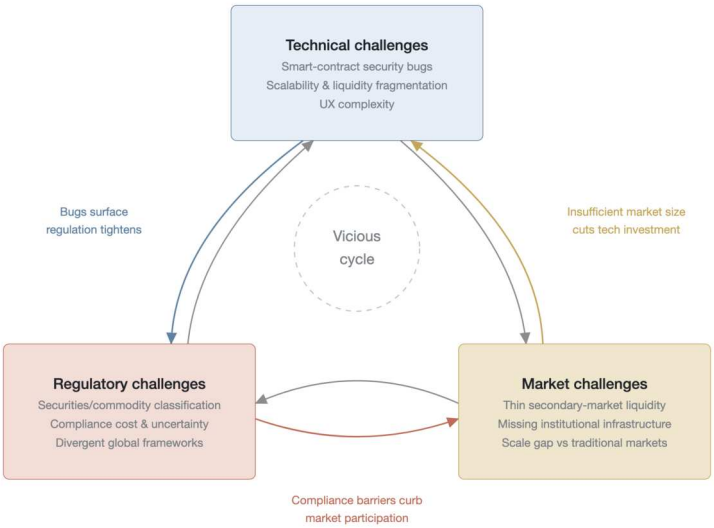


Figure 3-11. The interrelationship of the three major challenges of on-chain finance—the feedback loop among technology, regulation, and the market

On technical maturity, smart-contract security risk remains prominent (see Section 3.1.4 for the scale of security losses in 2025). The operational reliability of underlying blockchains still falls below that of traditional financial infrastructure: Solana suffered several multi-hour network outages between 2021 and 2023, and surging gas fees on the Ethereum mainnet during periods of high load make small transactions uneconomical, whereas the availability of traditional financial infrastructure such as Fedwire typically exceeds 99.99%. TVL on Layer 2 networks—which L2BEAT now reports as total value secured (TVS)—grew from roughly $5 billion to more than $40 billion, with a 2024–2025 peak approaching $50 billion [36], but this brought a fragmentation problem, and the user experience remains overly complex. On regulatory certainty, MiCA leads but is still being implemented, the U.S. jurisdictional dispute between the SEC and the CFTC is unresolved, and fragmentation leaves institutions facing high compliance risk. On market size, roughly $25 billion in tokenized RWAs is still a negligible fraction of the hundreds of trillions in the global traditional financial market, and illiquidity and a weak secondary market form a negative loop.

These three challenges are obstacles that effort can overcome, distinct from the impossibility boundaries of the next section. The essential difference is this: the former are matters of degree—technical security will keep improving, the regulatory framework will gradually clarify, and market size will grow with the adoption rate; their solution paths are known, and the only bottleneck is the speed of execution. The latter are matters of kind—sovereignty will not vanish, the physical world will not become digital, and the logical constraints of decentralization cannot be engineered away; even infinite technological progress cannot break through them. This distinction is the first analytical step in assessing the prospects of any specific on-chain financial innovation.

3.6 The impossibility boundaries

The three structural constraints—the sovereignty boundary, the physical-world interface, and the governance paradox—are not obstacles that can be overcome through technological progress or institutional innovation but permanent constraints that must be lived with. Identifying these boundaries is the premise of a mature analytical framework and the methodological basis for distinguishing surmountable challenges from impassable constraints.

3.6.1 The sovereignty boundary

At bottom, the permissionlessness and globalization of on-chain finance challenge the state's jurisdiction over financial activity. This challenge is not technical but political: financial sovereignty is a core component of national sovereignty, and the transmission of monetary policy, foreign-exchange controls, capital-account management, and sanctions enforcement all depend on jurisdiction over financial activity. If a state cannot track, restrict, or freeze the flow of funds within its territory, its monetary policy, tax system, and sanctions capability will all be fundamentally weakened.

Empirically, no major country has chosen to relinquish financial jurisdiction, and national responses form a full spectrum from comprehensive prohibition to framework-based regulation. In 2021, China comprehensively banned cryptocurrency trading and mining while advancing the research, development, and pilots of the digital yuan. The logic of this strategy is to reject permissionless on-chain finance while retaining the technical potential of programmable money, provided that control over programmability stays with the central bank. The United States has adopted a strategy of selective enforcement: in 2022, OFAC sanctioned Tornado Cash, freezing all Ethereum addresses associated with the mixing protocol—the first time a sovereign state imposed sanctions on immutable open-source smart-contract code rather than on a specific individual or organization [37] (the sanction was later overturned by the Fifth Circuit in November 2024, and OFAC removed it in March 2025, though criminal proceedings against the founders continue as a separate matter). This action sparked a constitutional debate over whether code constitutes protected speech, but its policy signal was clear: even a technically unstoppable on-chain protocol cannot transcend sovereign jurisdiction. The CFTC's enforcement actions between 2023 and 2024 against offshore DeFi protocols such as Opyn, ZeroEx, and Deridex further confirmed its swap-jurisdiction claim over crypto derivatives; U.S. users trading perpetual futures on offshore DEXs may technically violate the CEA, which bears directly on the central subject of this book. The EU, through MiCA, brings crypto assets into a unified regulatory framework—neither prohibiting them wholesale nor letting them run free, but framing jurisdiction and seeking a balance between room for innovation and regulatory capacity [23].

What these different strategies have in common is this: no country has chosen to cede financial jurisdiction. Fully permissionless global finance (anyone able to transact any amount with any counterparty at any time without government review) is politically impossible so long as the system of states endures. The realistic end state is *conditional permissionlessness*: technically permissionless but conditional on compliance. The zero-knowledge-proof compliance credentials discussed in Chapter 30 are one tech-

nical exploration of maximizing freedom within this boundary, letting users prove to a smart contract that they meet compliance requirements without exposing their identity or transaction details. The sovereignty boundary also explains why compression difficulty is so high at the right end of the trust-tax compression curve: the trust tax of equities and real estate is supplied primarily by sovereign institutions—corporate law defines shareholder rights, property law defines real-estate ownership—and sovereign institutions will not voluntarily withdraw from these domains.

3.6.2 The physical-world interface

The efficiency advantage of architectural trust is clearest when on-chain finance handles purely digital assets, but it hits a fundamental interface problem when it touches assets in the physical world: a blockchain can transfer the ownership of a token frictionlessly, but it cannot frictionlessly transfer control over the physical entity that the token represents.

The theoretical root of this problem lies in a generalized form of the oracle dilemma. A blockchain is a closed, deterministic system whose internal state can be fully verified cryptographically, but its interface with the external world depends on oracles—intermediaries that feed off-chain information on-chain. For price data, decentralized oracle networks such as Chainlink have provided relatively reliable solutions. But for the state of physical assets (such as whether a house still exists, whether a batch of goods is still in the warehouse, or who actually possesses a plot of land), no decentralized verification mechanism exists. Verifying the state of the physical world ultimately depends on human observation and institutional enforcement.

Real-estate tokenization illustrates these constraints layer by layer. At the title-confirmation layer, property-registration systems differ enormously across jurisdictions, and many emerging-market countries lack even a complete digitized registry, so the mapping between on-chain tokens and off-chain title needs legal institutions to establish and maintain it. At the rights-enforcement layer, an on-chain token can record a change in home ownership, but when the occupant refuses to leave, a smart contract cannot enforce it; that requires a court summons and the physical action of law-enforcement personnel. At the valuation layer, the value of real estate depends on many off-chain factors—location, physical condition, urban-planning policy—that on-chain data cannot capture. At the cross-border enforcement layer, even if on-chain records were globally unified, the degree of judicial recognition of on-chain title certificates varies across national courts. Tokenization can make a home's transaction records more efficient and transparent, but it cannot make title confirmation and enforcement more efficient; the latter will always depend on the offline legal system.

Commodity tokenization faces similar constraints. Tokenized gold requires trusted third-party custody and periodic audits to ensure a one-to-one correspondence between the on-chain token and the physical gold, and this trust still rests on institutional trust in the auditor and custodian rather than on architectural trust. For physical assets, architectural trust will always require the cooperation of institutional trust, and the optimal solution is a hybrid model of on-chain records plus off-chain enforcement. This also explains why the asset classes making the fastest progress (stablecoins, Treasuries,

private credit, and corporate bonds) are all purely financial assets: their value is defined entirely by contractual terms and cash flows and involves no transfer of control over a physical entity.

3.6.3 The governance paradox

The ideal of on-chain finance is that rules are defined by code, enforced by consensus, and independent of the discretion of any centralized entity. Under normal operating conditions this ideal holds, with smart contracts executing automatically according to predetermined logic and requiring no human intervention. In an extreme crisis, however, ideal and reality collide: a decentralized system is by design averse to centralized decision-making, yet in a crisis it most needs a rapid, discretionary, centralized response.

The 2016 Ethereum DAO incident is the defining case of this paradox. An attacker exploiting a recursive-call vulnerability in the DAO smart contract moved roughly 3.6 million ETH (worth roughly $50 million–$60 million at the time) within hours, about one-third of the DAO's total assets. From the standpoint of "code is law," the attacker's behavior was compliant at the code level: he invoked the contract's functions in a manner the code permitted. The community faced an impossible choice: respect the determinism of the code, thereby accepting the attack's outcome, or modify the blockchain's history to return the stolen assets, thereby repudiating the founding principle that code is law. In the end, the Ethereum community redirected the stolen funds to a recovery contract through a hard fork—in essence, a centralized decision repairing a decentralized system. The cost of this decision was the permanent split of Ethereum into two chains, ETH and ETC, the latter refusing the rollback out of fidelity to the "code is law" principle. The DAO incident proves that when a decentralized system encounters an extreme situation its rules cannot foresee, it must either accept the consequences of its rules or introduce human discretion to repair them; there is no third option.

The March 2025 JELLY incident at Hyperliquid brought this paradox into the domain of perpetual futures. A trader manipulated the liquidation mechanism by opening a large JELLY short position on Hyperliquid while pumping the JELLY price in the spot market, exposing Hyperliquid's insurance vault to tens of millions of dollars in potential losses [25]. Hyperliquid's validators voted to force the delisting of the JELLY token and settle all positions at a specific price, effectively preventing further losses. This decision technically protected the system and most users, but it raised a fundamental governance question: if validators can vote to change the rules in response to a crisis, then the system differs from a traditional centralized exchange only in that the decision-making body is a small set of validators rather than a single management team.

These two cases reveal an intrinsic logical constraint of decentralized governance: a fundamental trade-off between full decentralization and efficient crisis response. Full decentralization means no entity holds the power to modify the rules unilaterally, but crisis response requires action that is rapid, decisive, and potentially rule-breaking. Current practice in on-chain finance is evolving toward a hybrid governance architecture: under normal conditions, smart contracts execute automatically to maximize

decentralization; in an emergency, a multi-signature or validator committee executes interventions of limited scope to introduce the minimum necessary centralization; and all interventions are subject to time-locks and after-the-fact review to limit the abuse of centralization. Chapter 30 discusses this dual architecture of time-locks and emergency powers in detail.

3.6.4 Maximizing within boundaries

The three impossibility boundaries do not negate the value of on-chain finance; they are a necessary premise for building a rigorous analytical framework. System design that lacks boundary awareness will spend resources trying to break through impassable constraints rather than finding the optimal solution within them.

The analytical value of identifying boundaries lies in matching each category of financial function with the most suitable trust paradigm. In the domain of purely digital assets (such as cryptocurrency trading, perpetual futures, and decentralized lending), the architectural trust of on-chain finance has been thoroughly validated, and its advantage lies in permissionless global participation, around-the-clock operation, and automated execution. In the purely sovereign domain (such as currency issuance, equity as defined by corporate law, and property registration), institutional trust is irreplaceable, and the value of on-chain technology lies in raising efficiency rather than replacing institutions. The hybrid domain (such as tokenized Treasuries, compliant stablecoins, and institutional-grade DeFi) requires the synergy of architectural and institutional trust, and the BUIDL fund is a successful case of exactly that.

As an epistemological stance, *maximizing within constraints* has practical significance at three levels. At the level of technical design, protocol designers should treat the compliance interface as a native component of the architecture rather than an after-the-fact patch, reserving programmable access points for regulatory requirements while keeping the core logic decentralized. At the level of business strategy, entrepreneurs and investors should focus first on the left and middle of the trust-tax compression curve—the domains with a high share of technical trust tax and manageable institutional obstacles—rather than confronting sovereign institutions head-on at the right end. At the level of analytical method, in assessing any on-chain financial innovation one should first ask whether it touches the three impossibility boundaries, thereby distinguishing engineering problems that iteration can solve from structural constraints that will not vanish with technological progress. This stance is the epistemological foundation of all analysis that follows in this book.

3.7 Chapter summary

The value of digital assets lies not in creating a better Bitcoin but in providing a systematic technical toolkit for compressing the trust tax of finance. This chapter has organized that argument through five named theoretical concepts, each answering an analytical question at a different level.

The economic reexamination of the four distinctive properties raises the inherent attributes of digital assets from technical description to economic proposition: permission-

lessness raises the contestability of financial markets by several orders of magnitude by lowering sunk costs, although liquidity barriers keep it short of the theoretical standard of perfect contestability; around-the-clock trading transforms the time fragmentation of traditional markets from a hard constraint into a soft one; atomic settlement eliminates the settlement cascade but cannot prevent the liquidation cascade; and programmability, while expanding the design space of financial innovation, introduces code rigidity and governance tension because of the completeness constraint of smart contracts. These four properties do not function independently but combine to produce novel financial mechanisms unachievable in traditional finance: the funding rate of perpetual futures, the automated market-making of AMMs, and the atomic, uncollateralized borrowing of flash loans—all products of this combinatorial effect.

The institutional convergence thesis (which grafts a competitive trust-tax-compression mechanism onto DiMaggio and Powell's institutional-isomorphism theory) reveals a bidirectional process: traditional finance adopts on-chain technology to compress the trust tax, on-chain finance introduces institutional elements to raise the level of trust, and under competitive pressure the two move from opposite directions toward the same equilibrium range. The BUIDL fund demonstrates convergence in both directions at once and is the landmark empirical evidence for this thesis. The trust-tax compression curve further explains the unevenness of convergence: from payments to real estate, the share of the institutional trust tax rises and the difficulty of moving on-chain rises accordingly, and the current sequence of progress closely matches this gradient. On this basis, the three-horizon framework distinguishes an H1 that has reached scale, an H2 that is breaking through, and an H3 that remains a long-term prospect, providing differentiated analytical standards for assessing the on-chain prospects of different domains. The stablecoin settlement-layer migration thesis argues that the $33 trillion in annual settlement volume already marks a structural migration of the financial system's settlement layer, that the threefold network effect gives this migration a self-reinforcing and irreversible character, and that once the settlement layer is established, the on-chain migration of upper-layer applications becomes an architectural inevitability. The impossibility boundaries, in turn, draw clear limits around the optimistic arguments above: the sovereignty boundary means that fully permissionless global finance is politically infeasible, the physical-world interface means that architectural trust cannot independently handle assets involving physical entities, and the governance paradox means that a decentralized system inevitably requires a centralized response in a crisis. Identifying these three constraints is the premise for the epistemological stance of *maximizing within constraints* and the methodological basis for distinguishing engineering problems from structural constraints.

The three chapters of Part One constitute a complete cognitive framework. Chapter 1 established what the differences between digital assets and traditional finance are, Chapter 2 what those differences gave rise to in derivatives, and Chapter 3 why all of this carries systemic significance for the entire financial system. Perpetual futures—the highest-volume product in the first horizon of on-chain finance—will be the central object of analysis for the 30 chapters that follow. The next chapter enters Part Two, turning these macro frameworks into an empirical analysis of microstructure: who trades, how they behave, and how profits flow.

References

[1] BlackRock. (2024). *BlackRock USD Institutional Digital Liquidity Fund (BUIDL).* Securitize. https://securitize.io/buidl

[2] Depository Trust & Clearing Corporation. (2024). *Smart NAV pilot report: Bringing trusted data to the blockchain ecosystem.* https://www.dtcc.com/-/media/Files/Downloads/DTCC-Connection/Smart_NAV-Report.pdf

[3] European Investment Bank. (2021). *EIB issues its first ever digital bond on a public blockchain* [Press release]. https://www.eib.org/en/press/all/2021-141-european-investment-bank-eib-issues-its-first-ever-digital-bond-on-a-public-blockchain

[4] Siemens. (2023). *Siemens issues first digital bond on blockchain* [Press release]. https://press.siemens.com/global/en/pressrelease/siemens-issues-first-digital-bond-blockchain

[5] Visa Inc. (2023). *Visa expands stablecoin settlement capabilities to merchant acquirers* [Press release]. https://usa.visa.com/about-visa/newsroom/press-releases.releaseId.19881.html

[6] Monetary Authority of Singapore. (2023). *Project Guardian: Open, interoperable networks for digital assets.* https://www.mas.gov.sg/schemes-and-initiatives/project-guardian

[7] Artemis Analytics. (2026). *Stablecoin annual transaction volume data* (reported by Bloomberg, January 2026). Artemis. https://www.artemis.xyz/

[8] Brevan Howard Digital. (2024). *On stablecoins.* https://www.brevanhoward.com/research

[9] CoinGecko. (2025). *2025 annual crypto industry report.* https://www.coingecko.com/research/publications/2025-annual-crypto-report

[10] Baumol, W. J., Panzar, J. C., & Willig, R. D. (1982). *Contestable markets and the theory of industry structure.* Harcourt Brace Jovanovich.

[11] Shepherd, W. G. (1984). "Contestability" vs. competition. *American Economic Review*, 74(4), 572–587.

[12] Autonomous Research. (2023). *The cost of running a regulated exchange.* Autonomous Research Reports. (See also Oliver Wyman. (2023). *Global capital markets: Structural shifts.*)

[13] Hyperliquid. (2025). *Hyperliquid documentation.* https://hyperliquid.gitbook.io/hyperliquid-docs

[14] DeFiLlama. (2025). *Hyperliquid protocol statistics.* https://defillama.com/protocol/hyperliquid

[15] CertiK. (2026). *Hack3d: 2025 annual security report.* https://www.certik.com/resources/blog/hack3d-2025-annual-report

[16] Kaiko Research. (2024). *Crypto market liquidity: Intraday patterns and depth analysis.* https://www.kaiko.com/research

[17] Blonien, P., & Ober, A. (2024). *Is 24/7 trading better?* (SSRN Scholarly Paper

No. 4942934). https://doi.org/10.2139/ssrn.4942934

[18] Financial Stability Board. (2024). *The financial stability implications of tokenisation*. https://www.fsb.org/2024/10/the-financial-stability-implications-of-tokenisation/

[19] Szabo, N. (1997). Formalizing and securing relationships on public networks. *First Monday*, 2(9). https://doi.org/10.5210/fm.v2i9.548

[20] Hart, O., & Moore, J. (1999). Foundations of incomplete contracts. *Review of Economic Studies*, 66(1), 115–138. https://doi.org/10.1111/1467-937X.00080

[21] DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48(2), 147–160. https://doi.org/10.2307/2095101

[22] UK Finance & Oliver Wyman. (2023). *Unlocking the power of securities tokenisation*. UK Finance. https://www.ukfinance.org.uk/policy-and-guidance/reports-and-publications/unlocking-power-securities-tokenisation

[23] European Parliament & Council of the European Union. (2023). *Regulation (EU) 2023/1114 on markets in crypto-assets (MiCA)*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023R1114

[24] TRM Labs. (2025). *What the first quarter of MiCA tells us about CASP authorizations*. TRM Labs Blog. https://www.trmlabs.com/resources/blog/what-the-first-quarter-of-mica-tells-us-about-casp-authorizations

[25] Hyperliquid. (2025). *Post-mortem: JELLY token incident, March 2025*. https://hyperliquid.gitbook.io/hyperliquid-docs

[26] Coase, R. H. (1937). The nature of the firm. *Economica*, 4(16), 386–405. https://doi.org/10.1111/j.1468-0335.1937.tb00002.x

[27] Williamson, O. E. (1985). *The economic institutions of capitalism*. Free Press.

[28] RWA.xyz. (2025). *Real-world asset tokenization dashboard*. https://www.rwa.xyz/

[29] Boston Consulting Group & ADDX. (2022). *Relevance of on-chain asset tokenization in 'crypto winter'*. https://www.bcg.com/publications/2022/relevance-of-on-chain-asset-tokenization-in-crypto-winter

[30] McKinsey & Company. (2024). *From ripples to waves: The transformational power of tokenizing assets*. https://www.mckinsey.com/industries/financial-services/our-insights/from-ripples-to-waves-the-transformational-power-of-tokenizing-assets

[31] Atlantic Council. (2025). *Central bank digital currency tracker*. https://www.atlanticcouncil.org/cbdctracker/

[32] People's Bank of China. (2025). *Shuzi renminbi shidian jinzhan baogao* [Digital yuan pilot progress report]. http://www.pbc.gov.cn/

[33] Visa Inc. (2025). *Annual report 2025*. https://www.visa.com/annualreport

[34] Katz, M. L., & Shapiro, C. (1985). Network externalities, competition, and compatibility. *American Economic Review*, 75(3), 424–440. https://doi.org/10.2307/1814809

[35] Circle. (2023). *USDC and Silicon Valley Bank update*. Circle Blog. https:

//www.circle.com/blog/usdc-update

[36] L2BEAT. (2025). *Layer 2 total value locked*. https://l2beat.com/scaling/tvl

[37] U.S. Department of the Treasury. (2022). *OFAC designates Tornado Cash* [Press release]. https://home.treasury.gov/news/press-releases/jy0916

[38] Visa Onchain Analytics. (2026). *Onchain stablecoin transaction volume: Raw and adjusted (bot/MEV-filtered) estimates* (data by Allium). Retrieved June 2026, from https://visaonchainanalytics.com

[39] CoinGecko. (2025). *Rise of perpetuals and perp DEXs*. https://www.coingecko.com/research/publications/rise-of-perpetuals-perp-dexs

Part II
The Ecosystem

Core question: Who trades perpetual futures, why do they come, and on what kind of ecosystem do they depend?

Participants and Behavior in Perpetual Futures Trading

On March 12, 2025, an anonymous wallet opened a 50x-leveraged long position in ETH on Hyperliquid with an initial margin of 4.3 million USDC, then used its unrealized profits to keep adding to the position, expanding its notional value to more than $300 million [1]. Under Hyperliquid's maintenance margin ratio (approximately 0.3%), 50x leverage placed the liquidation price roughly 1.7% away from the entry price, and the trader further narrowed this buffer by repeatedly withdrawing excess margin. The entire trade was publicly visible on-chain; copy traders rushed in to go long, shorts were forced to close, and the price was driven still higher. Yet rather than close the position, the trader withdrew realized profits from the account, deliberately lowering the margin ratio and raising its own liquidation price. Hyperliquid in fact allows traders to withdraw margin balances in excess of the maintenance-margin requirement while a position remains open, and this rule became the linchpin of the strategy. The protocol finally liquidated the position automatically when the ETH price reached $1,915. Because the liquidation was larger than liquidators could absorb through competitive market bidding, and the insurance fund was too small to cover it, the position passed to the HLP market-making vault. In unwinding the position, the vault incurred roughly $4 million in slippage losses and triggered more than $130 million in panic withdrawals by liquidity providers [2]. This episode reveals that, in a permissionless, fully transparent, and high-leverage digitally native environment, the basic assumptions about participant behavior have been rewritten. It raises the central questions of this chapter: which participants populate the perpetual futures market, how do they compete with one another, and what is the ultimate source of the market's profits?

The trader had calculated precisely that the penalty for being liquidated was far smaller than the slippage of unwinding a $300 million position on the open market in the ordinary way, and thereby shifted the price impact systematically onto the protocol's passive liquidity providers. Parasitic behavior in the on-chain environment has thus moved away from the informational advantage of traditional finance and toward a deep exploitation of what protocol rules imply economically.

To understand this new form of competition, this chapter takes the classic tripartite taxonomy of Harris (2003) as its academic starting point and builds a two-layer classification framework, functional and behavioral. It then examines, in turn, the behavioral logic and profit sources of the six core participant types; the game network among

participants and the mapping of profit flows; the threefold reshaping of participant behavior by the digitally native environment; the rise of autonomous artificial intelligence agents as a new type of participant; and a systematic comparison with traditional finance. The chapter's central thesis is that the gains of profit-motivated traders ultimately come from only two channels—the hedging premium that utilitarian traders pay to transfer risk, and the systematic losses that futile traders incur through mistaken decisions—and that the digitally native environment fundamentally reshapes this game structure through the fluidization of roles, the algorithmization of competition, and the transparency paradox.

4.1 The participant classification framework

This chapter repeatedly invokes several concepts introduced earlier. The transparency paradox of Chapter 1 holds that the full disclosure of on-chain data does not amount to an equal distribution of informational advantage; the composability thesis explains how the seamless interoperability among smart contracts creates a capacity for atomic execution that does not exist in traditional finance. The cost of abolishing the expiration date, analyzed in Chapter 2, shows that perpetual futures gain flexibility by eliminating delivery at maturity but at the same time introduce structural costs such as the indefinite horizon of inventory risk. These concepts form the necessary background for understanding this chapter's analysis of participant behavior.

Market microstructure theory offers two complementary classificatory perspectives. One is based on the functional purpose of a trade and answers the question of *why* participants trade, revealing the fundamental economic function that gives the market its reason to exist. The other is based on the specific role a participant plays in the market and answers the question of *how* participants trade, focusing on the particular roles that different participants play in price formation and the supply of liquidity.

4.1.1 Functional classification

Harris (2003) divides market participants into three basic categories by trading motive [3]. Utilitarian traders participate in the market not to profit directly from price movements but to achieve some external economic objective. In the perpetual futures market, the most typical utilitarian trader is the hedger. A Bitcoin miner, for example, may short perpetual futures to lock in future mining revenue, thereby transferring the risk of price fluctuations to other market participants. For this miner, perpetual futures are a risk-management tool rather than a vehicle for speculation. Likewise, a decentralized finance (DeFi) protocol treasury holding a large amount of ETH may use perpetual futures to hedge the price risk on its balance sheet. The defining feature of utilitarian traders is that they are willing to pay a certain cost—for instance, accepting a slightly unfavorable execution price—in exchange for certainty. They are the principal source of demand for market liquidity; their trading needs arise from external economic activity and constitute the fundamental reason the market exists, namely the transfer of risk.

Profit-motivated traders take the extraction of profit from the market as their central

objective; they are the principal source of market liquidity and the primary driver of price discovery. This group can be further subdivided by the source of its profits. Informed traders rely on an informational advantage—whether from deep fundamental research, distinctive analysis of on-chain data, or possession of non-public information—to forecast price movements and profit from them. Arbitrageurs do not depend on judging the direction of prices; instead, they earn low-risk profits by discovering and exploiting pricing discrepancies between different markets or different assets. Market makers earn the spread by simultaneously posting bid and ask quotes, providing other traders with the convenience of immediate execution. The framework also singles out a class of parasitic traders, who do not rely on fundamental information but profit by exploiting the behavior of other traders [3]. In the on-chain environment, maximal extractable value (MEV) searchers and front-running bots are the archetypal parasitic traders: they monitor the pending transaction pool to anticipate other traders' intentions and execute ahead of them to capture profit, and their behavior generally does not improve the market's price-discovery efficiency.

Futile traders are the third category in this taxonomy, and the one most easily overlooked. These traders expect to profit from trading but, owing to a lack of skill, information, or judgment, lose money persistently on average [3]. They often mistakenly believe themselves to be profit-motivated traders. In the crypto perpetual futures market, the presence of futile traders is especially pronounced. The extreme conditions of October 2025 provide direct evidence of their scale: within a mere 24 hours, more than 1.62 million trading accounts were forcibly liquidated, with total liquidations exceeding $19 billion, of which approximately 85%–90% (about 87%) by value were long positions (the precise liquidation statistics come from CoinGlass; exchange data generally understate the true scale) [4][5]. These traders chased the market higher during calm or rising periods, using leverage far beyond their risk tolerance and generally lacking any downside-protection strategy. Research in behavioral finance shows that this trading behavior is, in its psychological mechanisms, highly similar to gambling [6]. Their losses constitute an important source of profit for profit-motivated traders.

4.1.2 The two ultimate sources of profit

The core value of the tripartite taxonomy is that it reveals not only the fundamental reason the market exists (the risk-transfer demand of utilitarian traders) but also the ultimate source of market profits: the gains of profit-motivated traders come, in the final analysis, from only two channels [3]. The profits of arbitrageurs appear, on the surface, to derive from correcting mispricings in the market, seemingly independent of the two sources below. Mispricings, however, are themselves jointly produced by the inelastic demand of utilitarian traders and the pricing errors of futile traders; arbitrage profits can therefore, ultimately, still be traced back to the two sources, of which they are an indirect derivative.

The first source is the hedging premium that utilitarian traders pay to obtain utility. To convert future uncertainty into present certainty, utilitarian traders are willing to pay a premium—for example, by accepting a slightly unfavorable execution price to guarantee immediate execution, or by bearing a positive funding rate over an extended period to

maintain a hedge. These additional costs are essentially the insurance premiums they pay for risk transfer, flowing directly into the bid-ask spreads earned by market makers, the funding rates collected by short counterparties, and the basis profits captured by arbitrageurs. (The detailed mechanism for calculating the funding rate—including the composition of its interest rate and premium components, the typical 8-hour settlement cycle, and the extreme range of its magnitude—is presented in Section 2.3. Under normal market conditions, the funding rate is typically on the order of $\pm 0.01\%$ every 8 hours, but in extreme conditions it can climb to $\pm 0.1\%$ or higher, at which point the funding rate itself becomes a significant source of profit or a substantial cost of carry.) This profit channel is the fundamental economic rationale for the existence of the perpetual futures market: without genuine hedging demand, a derivatives market would degenerate into a purely zero-sum arena. In the crypto perpetual futures market, however, the proportion of utilitarian traders is far lower than in traditional commodity or financial futures markets. Schmeling et al. (2023), drawing on data from the Bank for International Settlements (BIS), studied the structure of carry trading in crypto markets, analyzing primarily the statistical properties of funding rates and the risk-return characteristics of carry strategies [7]. From its findings—that funding rates in crypto perpetual futures markets are predominantly positive but highly time-varying, and that speculative positions dominate—one can tentatively infer that participants in crypto derivatives markets are driven primarily by speculative motives, and that the share of genuine hedging demand may be lower than in traditional commodity futures markets (in most such markets, commercial hedgers typically hold more than half of the open interest, though the precise proportion varies by product and period). This inference, however, involves an additional argumentative step from the structure of carry trades to the share of hedging demand, and should not be regarded as a direct conclusion of that paper. Moreover, in interpreting the phenomenon of low hedging demand in crypto markets, one must distinguish two very different states: the absence of hedging demand, and the presence of hedging demand that cannot be expressed. In fact, a substantial share of institutional investors—including funds, corporate treasuries, and asset managers that hold exposure to crypto assets—have a genuine economic need to hedge risk through perpetual futures, yet cannot translate this need into actual hedging positions because of regulatory restrictions in their jurisdictions, compliance constraints arising from fiduciary duties, or a lack of access to regulated trading venues. This implies that the observable hedging demand in the market may be systematically lower than the actual economic need for hedging. As regulatory frameworks gradually become clearer and regulated trading infrastructure matures, this suppressed demand may be released in the future, thereby altering the relative proportions of the two profit sources.

The second source is the systematic losses that futile traders incur through mistaken decisions. Because they are at a systematic disadvantage in access to information, analytical capacity, or execution speed, futile traders are on average in a state of persistent loss. Their losses flow to profit-motivated traders along several paths: when they are liquidated, their short counterparties reap a direct directional profit; in ordinary trading, market makers safely earn the spread on their non-toxic order flow; and when forced liquidation is triggered, liquidators collect substantial liquidation penalties. From the

perspective of market microstructure, it is precisely the non-toxic order flow supplied by futile traders that subsidizes market makers' adverse-selection losses to informed traders, enabling market makers to remain profitable and continue supplying liquidity (this mechanism is fully elaborated in Section 4.3.3). A structural feature of today's crypto perpetual futures market is that the losses of futile traders constitute the single largest source of profit, while the hedging premium is not yet fully developed. This speculation-dominated profit structure is one root of the market's fragility: when futile traders are depleted by their losses and exit, the market loses its principal source of profit, market makers withdraw because they can no longer offset their adverse-selection losses, and market liquidity dries up in a recurring cycle.

The depletion hypothesis above, however, should not be understood as a one-directional, irreversible process; it should instead be examined within a dynamic-equilibrium framework. The stock of futile traders at any point in time depends on two opposing rates: a replenishment rate and a depletion rate. The replenishment rate is driven by the inflow of new users, which accelerates markedly during bull-market cycles under the combined catalysis of price narratives (the confirmation bias that *a rising price proves one right*), social-media contagion, and accommodative macro liquidity. The depletion rate is determined by loss-driven exits, as futile traders gradually leave after suffering consecutive losses and forced liquidations. When the replenishment rate persistently exceeds the depletion rate, the market displays a boom of ample liquidity, narrowing spreads, and richly profitable market makers; when the depletion rate exceeds the replenishment rate, the market enters a downturn of contracting liquidity, widening spreads, and profit-motivated participants preying on one another. This dynamic framework maps closely onto the bull-bear cycle: a bull market is essentially a period of net inflow of futile traders, a bear market a period of net outflow, and the turning points of the cycle tend to occur in the critical region where the two rates approach balance. More fundamentally, the inflow and outflow of futile traders are not exogenous random events but endogenous products of global macroeconomic conditions. The dollar liquidity cycle (proxied by the size of the Federal Reserve's balance sheet and global dollar funding costs), the swings in global risk appetite (proxied by the VIX index and emerging-market capital flows), and the regulatory cycles of the major jurisdictions (from tolerance to tightening to clarification) jointly determine the rhythm at which participants enter and exit the crypto market. In periods of ample macro liquidity, rising risk appetite, and relatively permissive regulation, large numbers of financially inexperienced new participants are drawn into the high-leverage perpetual futures market, supplying profit-motivated traders with abundant non-toxic order flow; when macro conditions reverse, the inflow of new users collapses, the existing stock of futile traders is cleared out at an accelerating pace, and the market's profit sources dry up accordingly. Any assessment of the sustainability of the perpetual futures market therefore cannot be divorced from a judgment about the macro-liquidity environment. The market's internal game structure matters, but changes in the composition of participants are ultimately governed by how external macroeconomic conditions shape the flows of capital and people.

4.1.3 Behavioral classification

The functional classification explains *why* participants trade, but understanding the market's microstructural mechanics also requires analyzing *how* they trade—the specific roles they play in the market. Building on the theoretical framework above [3], the participants of the perpetual futures market can be divided into seven core roles. Market makers are the suppliers of market liquidity; by posting bid and ask quotes simultaneously, they give other traders the convenience of immediate execution. Informed traders drive price discovery; they trade on their informational advantage and incorporate new information into prices. Arbitrageurs maintain market efficiency; by eliminating price discrepancies between markets, they ensure that the law of one price holds. Speculators are the principal bearers of risk and a source of market vitality; their trading activity provides the market with abundant liquidity but may also amplify volatility. Liquidators safeguard system stability; they forcibly close the positions of leveraged traders when their margin becomes insufficient, preventing risk from spreading. In addition, brokers and agents provide market access, order execution, and trade-protection services to all other types of traders; in the on-chain environment, this role is filled by decentralized exchange (DEX) aggregators, MEV-protection services, and wallet front-ends.

Role	Core function	Profit source	Primary risk	Typical examples
Market maker	Provide liquidity	Bid-ask spread	Adverse selection, inventory risk	Wintermute, AMM LPs, HLP vault
Informed trader	Drive price discovery	Informational advantage	Information decay, regulatory risk	On-chain analysts, insider traders
Arbitrageur	Maintain market efficiency	Spread convergence	Execution risk, counterparty risk	Funding-rate arbitrageurs, cross-market arbitrageurs
Parasitic trader	Profit from others' behavior	Transaction-ordering rights, information theft	Intense competition, mechanism changes	MEV searchers, front-running bots
Speculator	Bear risk, provide liquidity	Judgment of price direction	Leverage liquidation, emotional bias	Retail traders, trend-following funds
Liquidator	Maintain system stability	Liquidation penalties	Market impact, intense competition	Liquidation bots
Broker/agent	Provide market access and execution protection	Service fees, trade-routing optimization	Trust risk, compliance risk	DEX aggregators, MEV protection, wallet front-ends

Table 4-1. Behavioral classification of participants in the perpetual futures market (Data source: author's compilation)

Table 4-1 sets out the core function, profit source, primary risk, and typical examples of the seven participant types. Two structural features stand out: the diversity of profit

sources (from the market maker's bid-ask spread to the liquidator's liquidation penalty) and the asymmetry of risk exposure (market makers and speculators bear persistent, uncertain risk, whereas arbitrageurs and liquidators face competitive pressure mainly at the level of execution). This lays the microfoundation for the analysis of the game relationships that follows.

The speculator is not an independent category within the functional classification; it is a cross-cutting role that spans the two categories of profit-motivated and futile traders. A minority of speculators with genuine judgment belong to the profit-motivated traders, whereas the majority who lack an informational advantage belong to the futile traders. This cross-category attribute makes the behavior of the speculator group highly heterogeneous overall, a point analyzed in detail in Section 4.2.5.

The role of the broker/agent, in turn, differs fundamentally from that of the other six participant types. The first six participate directly in the market game, contending for profit by bearing risk or providing services, whereas the broker/agent belongs to the market's infrastructure layer, providing market access and execution services to other participants and not itself participating directly in the directional game. The subsequent behavioral analysis and construction of the game network in this chapter therefore focus on the six directly contending participants; the functions of the broker/agent are discussed in detail in the service-provider ecosystem of Chapter 5.

Figure 4-1 integrates the two dimensions—functional and behavioral—to present a two-layer classification framework for participants (the top layer, *why they trade*; the second layer, *how they trade*). The core tension the figure reveals is that a single behavioral role (such as the speculator) may fall into different categories along the functional dimension—precisely the complexity that the subsequent analysis must address.

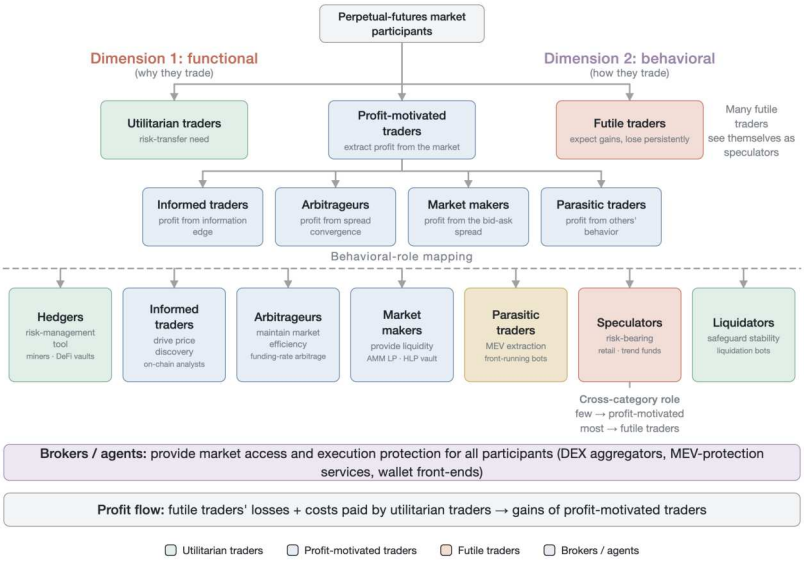


Figure 4-1. A two-dimensional classification framework for participants in the perpetual futures market (conceptual framework diagram; author's synthesis based on the tripartite taxonomy of Harris 2003 [3]; not empirical data)

4.1.4 The hidden costs of the negative-sum game

In a strict mathematical sense, the perpetual futures market is not a zero-sum game but a negative-sum game. Summing the realized and unrealized profit and loss (P&L) of all traders necessarily yields a result below zero. This shortfall represents the friction costs that must be incurred to keep the entire market infrastructure running.

$$L_{\text{neg}} = \sum_{i=1}^{N_f} f_i + \sum_{k=1}^{N_m} m_k + \sum_{l=1}^{N_g} g_l$$

Here, f_i is the transaction fee on the i-th trade, m_k is the amount of MEV captured in the k-th extraction, and g_l is the gas fee on the l-th trade. L_{neg} is the market's negative-sum quantity, representing the negation of the sum of all participants' P&L. The funding rate is not counted in the negative-sum quantity, because on major platforms (Binance, Hyperliquid, dYdX, and others) it is designed as a direct transfer payment between longs and shorts, from which the protocol itself extracts no fee. The formula also does not explicitly list implicit frictions such as slippage costs and the opportunity cost of margin. In the thin-liquidity environment on-chain, the market impact (slippage) of large orders can constitute a significant implicit transaction cost, while the forgone return on capital locked in margin is another cost that is easily overlooked. These implicit frictions further deepen the negative-sum nature of the market.

This negative-sum quantity is analogous to the institutional friction costs of the traditional financial system; only the manner of its collection has been reshaped by the digitally native architecture. Traditional finance levies explicit fees through intermediaries such as exchanges, clearinghouses, and brokers, whereas on-chain finance collects them through protocol-hardcoded fees, the execution fees of the blockchain network, and the MEV extracted by algorithmic bots. In the fully transparent public-mempool environment in particular, MEV constitutes an implicit tax. Every order executed through a decentralized exchange may have part of its value extracted by searchers—via front-running or sandwich attacks—during ordering and block packing, deepening the market's negative-sum character still further.

In the DEX ecosystem there is also a special external source of funds: protocol token incentives. Some protocols subsidize participants through liquidity mining, trade mining, and airdrops, which is essentially a form of seigniorage, with the cost borne by token holders through dilution. Protocol token incentives, however, are not an intrinsic institutional feature of the perpetual futures market (many centralized exchanges operate without relying on this mechanism), and they therefore do not constitute an independent category of profit source; they are instead an external subsidy mechanism specific to DEXs. This external subsidy mechanism has an important effect on the boundary conditions of the negative-sum game described by Equation 4-1. Absent token incentives, the perpetual futures market is strictly negative-sum, and the sum of all participants' P&L is necessarily negative. With token incentives, if the total subsidy exceeds L_{neg}, participants as a whole may exhibit positive-sum returns in the short run; but this positive-sum property rests on continuous token issuance, and token inflation inevitably dilutes the value held by existing holders, so it is unsustainable over the long run. As the token price falls under inflationary pressure, the real value of the subsidy

diminishes, and the market ultimately reverts to its intrinsic negative-sum nature.

4.1.5 Limitations of the classification

In using this classification framework, one must recognize the limitations it faces in the digitally native environment. In traditional financial markets, licensing regimes and a high degree of professional specialization keep participants' roles relatively fixed. In the decentralized finance environment, however, the boundaries between roles become blurred. A single user may, within the same day, be a speculator (opening a leveraged position), a market maker (supplying funds to a liquidity pool), and an arbitrageur (profiting from price discrepancies across protocols). This role fluidity is one of the most distinctive features of the digitally native market, and it is examined in detail in Section 4.4. Moreover, the framework is built primarily on assumptions of human behavior, yet an increasing share of on-chain trades is executed by automated bots whose behavioral patterns differ starkly from those of human traders, which likewise calls for a re-examination of the classical theory.

Despite these limitations, the classification framework retains significant analytical value. It provides a structured starting point for understanding the complex diversity of market behavior, allowing us to dissect systematically the core motives, strategies, and market impact of each type of participant. Acknowledging its limitations does not negate its validity; rather, it keeps the analysis critical—especially in examining how the digitally native environment breaks down these classical categorical boundaries.

4.2 Behavioral analysis of the six core participant types

Within the profit-source framework of Harris (2003), each type of participant can be understood as an economic entity contending for a specific share of the two profit sources. Its behavioral logic is jointly determined by its profit source, its risk exposure, and its game relationships. This section examines, in turn, the six core participant types of the perpetual futures market, revealing how the digitally native environment reshapes the basic assumptions about participant behavior in classical market microstructure theory.

4.2.1 Market makers

Market makers are the core liquidity providers of market microstructure, and their basic business model is to earn the bid-ask spread by quoting on both sides of the order book. Market makers earn the spread primarily from the non-toxic order flow of futile traders, while also capturing income from the hedging premium that utilitarian traders are willing to pay in crossing the spread for immediate execution. This profit pool, however, is continually compressed by informed traders and mechanism parasites.

The core challenge facing market makers can be understood through the three-component decomposition of the spread. Stoll (1989) showed that the bid-ask spread can be decomposed into order-processing cost, inventory-holding cost, and adverse-selection cost [8]. In traditional financial markets, these three costs remain in relatively stable proportion. In the crypto perpetual futures market, however,

the adverse-selection cost is markedly amplified. This judgment rests mainly on theoretical inference: given the empirical fact that the share of informed trading in crypto markets is high, the classical logic of spread decomposition reasonably implies a correspondingly greater weight on the adverse-selection component. Direct empirical measurement of the exact percentage of the bid-ask spread that adverse selection accounts for in crypto perpetual futures markets remains limited, however. Estimating these components systematically from on-chain market data, using classical spread-decomposition methods such as Huang-Stoll or Glosten-Harris, is still a research agenda awaiting deeper development. As noted earlier in the discussion of the cost of abolishing the expiration date, the absence of an expiry renders inventory risk indefinite: market makers cannot know when a trend will end, which makes them more sensitive to adverse-selection risk. Tiniç et al. (2023) documented substantial adverse-selection costs in crypto markets (about 10% of the effective spread in their sample), from which one can infer that market makers face greater adverse-selection risk in crypto markets than in traditional ones—part of the reason crypto perpetual futures spreads are systematically wider than those of traditional futures [9]. The probability of informed trading (PIN) model proposed by Easley et al. (1996) provides a classic tool for quantifying order-flow toxicity [10]. In the on-chain environment, because all trade data are publicly visible, the assumptions of the traditional PIN model change: informed trading is no longer hidden within an anonymous order flow but can be partly observed through on-chain analysis. This requires researchers to apply classical order-flow-toxicity measures with due caution about their conditions of applicability. In addition, Milionis et al. (2022) formally analyzed the losses incurred by automated market makers (AMMs) on decentralized exchanges, showing that the losses passive liquidity providers bear from adverse selection are structurally higher than the inventory losses of traditional market makers [11].

Figure 4-2 decomposes the market maker's bid-ask spread into three cost components—order processing, inventory holding, and adverse selection—and compares their relative weights in traditional versus crypto perpetual futures markets. Adverse selection accounts for a markedly larger share in crypto markets, reflecting how the high proportion of on-chain informed trading compresses market makers' profitability.

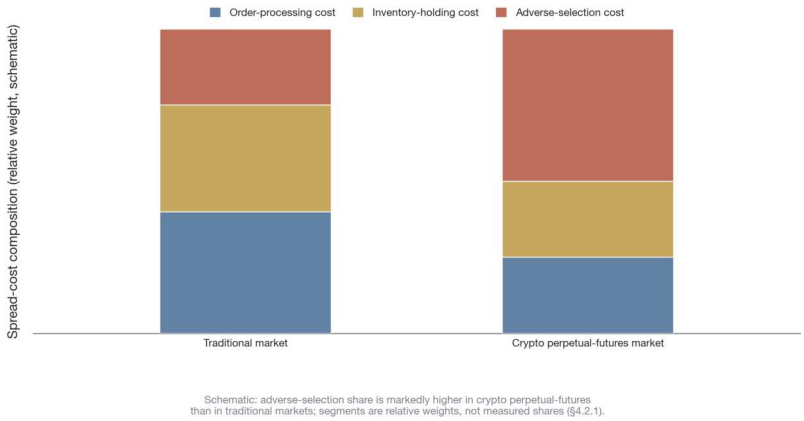


Figure 4-2. The three-component cost decomposition of the market maker's bid-ask spread (illustrative cost structure: the height of each cost segment within a bar reflects only qualitative relative weight, not a measured share; the larger adverse-selection segment on the crypto side merely indicates the direction in which this cost is significantly amplified and does not represent its actual share of the spread; Tiniç et al. 2023 measured adverse selection at about 10% of the effective spread in their sample [9], which serves as an independent empirical anchor)

The market-making paradigms of centralized and decentralized exchanges have already diverged. On centralized exchanges (CEXs), market making is dominated by professional firms that deploy high-frequency algorithms and cross-exchange hedging strategies; their core competitive strengths are speed and information-processing capacity, and the greatest unknown they face is the opacity of the exchange's order-flow quality. On decentralized exchanges, the market-making model diverges further into a passive-liquidity-pool model and a model that combines an on-chain central limit order book (CLOB) with a vault. In the passive model, liquidity providers passively bear all of the adverse-selection risk; the mechanism is simple but capital-inefficient. In the vault model, protocol algorithms make markets actively, seeking a balance between a low entry barrier and risk management.

The execution architecture of decentralized perpetual futures exchanges is not monolithic, and differences in the underlying consensus and ordering mechanisms have a profound effect on market makers' strategic choices and risk exposures. The prevailing DEX execution architectures can be summarized into three paradigms. The first is the fully on-chain model, in which both trade matching and settlement rely on the consensus mechanism of the underlying public chain; the archetypal examples are AMM protocols deployed on general-purpose L1s such as Ethereum or Solana. Under this architecture, all trading intentions propagate through the public mempool, block producers hold ultimate control over transaction ordering, and market makers are therefore exposed to the most severe MEV risk, with sandwich attacks, front-running, and back-running constituting persistent implicit costs. At the same time, block-confirmation time imposes an incompressible lower bound on latency, making it difficult for market makers to cancel orders or adjust quotes in time amid sharp volatility, so that inventory risk is amplified. The second is the application-chain or centralized-sequencer model, of which Hyperliquid is the archetype. Hyperliquid runs its own L1 blockchain (based on HyperBFT consensus), but its validator set is relatively concentrated, the

sequencer is controlled by the protocol team, and trade matching is completed off-chain at sub-second latency before being confirmed on-chain in batches. This architecture gives market makers execution speed and deterministic ordering close to those of a centralized exchange, effectively eliminating the MEV risk arising from a traditional public-chain mempool, but at the cost of introducing a trust assumption about the sequencer operator—in theory the sequencer holds unilateral power over transaction ordering and censorship, which constitutes a centralization risk analogous to the internalization of order flow within an exchange in traditional finance. The third is the application-chain model built on the Cosmos SDK, of which dYdX v4 is representative. dYdX v4 runs an off-chain order book on its dedicated Cosmos chain, with validators simultaneously performing the dual roles of order matching and block proposal; maintenance of the order book is done off-chain in memory, and only the final execution results are submitted to on-chain consensus. This hybrid architecture disperses ordering power through the decentralization of validators, but it also introduces a subtle MEV vector: a validator, as block proposer, could in theory order strategically after observing the state of the order book, though dYdX constrains this through economic incentives and reputation mechanisms. The latency profiles of the three architectures also differ significantly—the fully on-chain model's latency is measured in seconds, Hyperliquid's sequencer latency in sub-seconds, and dYdX v4's hybrid model falls between the two—which directly determines the upper bound on how frequently market makers can update quotes and the timeliness of their inventory hedging. For a professional market maker, the choice of architecture is essentially a trade-off between the trust model and the latency profile: the fully on-chain model offers the strongest trustless guarantees but the highest latency and the greatest MEV exposure; the centralized-sequencer model offers the lowest latency but requires trusting the sequencer's fairness; and the Cosmos application-chain model seeks a compromise between the two.

The differences in execution architecture extend further to how the mark-price mechanism shapes market-maker behavior and liquidation triggers. The mark price is the benchmark price a perpetual futures protocol uses to compute unrealized P&L and to determine liquidation triggers, and its construction differs substantially across platforms. Most protocols adopt an external oracle price (such as the aggregated spot price provided by Chainlink or Pyth) as the core anchor of the mark price, so as to prevent the protocol's internal execution price from being manipulated to trigger improper liquidations. The update frequency of the mark price, its smoothing algorithm, and its tolerance for deviation from the protocol's internal execution price all affect market makers' behavioral decisions, however. In protocols where the mark price updates slowly or the smoothing window is long, the protocol's internal execution price may deviate significantly from the mark price over short intervals, and market makers can adjust their quoting strategy accordingly to avoid the risk of being wrongly liquidated because of a lagging mark price; conversely, in protocols where the mark price is highly sensitive and updated in real time, market makers' liquidation buffer is more precise but also more fragile, and any price spike from an external oracle can instantly trigger large-scale liquidations, in turn forcing market makers to bear excess market impact as they absorb the liquidation flow. Hyperliquid adopts a mark-price mechanism anchored to a volume-weighted price from major centralized exchanges and

sets an instantaneous-deviation protection threshold, which to some extent reduces the systemic liquidation risk caused by oracle manipulation, but also means that in extreme conditions the mark price may lag the true market price, introducing additional uncertainty into pricing when the market-making vault takes over liquidated positions. The design of the mark-price mechanism thus becomes one of the key variables by which market makers assess protocol risk: it determines not only the precise conditions for triggering liquidation but also, by shaping market makers' quoting behavior, the liquidity depth and price-discovery efficiency of the entire market. Market makers on centralized and decentralized exchanges also face informational disadvantages of different natures. On centralized exchanges, third-party market makers may face a potential conflict of interest with the exchange's proprietary market-making service, which, by virtue of its priority visibility into order flow, may hold a structural informational advantage; on decentralized exchanges, the principal informational threat facing market makers comes from the targeted sniping of MEV searchers, who conduct front-running and sandwich attacks by monitoring publicly visible trading intentions. The two informational disadvantages are entirely different in nature: the former stems from information asymmetry in an opaque environment, the latter from asymmetry in execution speed in a fully transparent one; yet both systematically erode market makers' profitability.

The double-sided nature of Hyperliquid's market-making vault is a paradigmatic case for understanding the market maker's dilemma. Facing uninformed forced-liquidation flow, a market maker can earn substantial profits; but facing strategic actors with a mechanism-parasitic advantage, it suffers systematic losses. This shows that a market maker's P&L is essentially a function of the toxicity concentration of its order flow: the same market maker exhibits starkly different profitability against counterparties whose order flow carries different information content. The case also reveals the deeper difference between the professional-market-maker and the passive-vault models. Professional market makers typically deploy an entire suite of sophisticated risk controls, including real-time detection of order-flow toxicity (dynamically identifying the concentration of informed trading using indicators such as the volume-synchronized probability of informed trading, or VPIN), strict inventory caps, tiered quoting strategies (dynamically adjusting the depth posted at different price levels according to order-flow quality), and cross-asset correlation hedging. By contrast, passive-vault models such as HLP lack a real-time toxicity-filtering mechanism and, in essence, absorb all types of order flow indiscriminately under a uniform strategy. This structurally exposes them to the spike risk of adverse selection: when toxic order flow floods in en masse, a passive vault cannot rapidly narrow its quotes or suspend market making as a professional market maker would, and it therefore systematically bears a disproportionate loss.

4.2.2 Informed traders

Informed traders are the principal driver of market price discovery. Their profits come mainly from the directional game against futile traders, while they also extract value from market makers' inventory through adverse selection. The presence of informed traders enables market prices to reflect the latest information, but it also raises market makers' costs.

The classic model of Kyle (1985) describes how informed traders strategically execute their trades, gradually incorporating private information into prices [12]. In the digitally native environment, this logic is accelerated and complicated. The rise in the speed of information transmission and execution has sharply shortened the traditional window of informational advantage. More importantly, the source of the informational advantage has been reshaped. The traditional informational advantage based on fundamental research and industry knowledge still exists, but the share of the digitally native informational advantage—based on on-chain data analysis, whale-address tracking, smart-contract-interaction monitoring, and scanning of the MEV pending pool—has risen markedly.

This digitally native informational advantage is a concrete embodiment of the transparency paradox. On-chain information is public private information: everyone can see a whale's positions and trading intentions, but only participants with the specific data-processing capacity and algorithmic infrastructure can correctly interpret its meaning and convert it into trading signals. The informed trader's barrier therefore shifts from the privilege of acquiring information to the capacity to process and execute on public information.

Informed traders have a double-sided character in market microstructure. On one hand, by trading, they transmit information into prices, promoting price discovery and improving the market's pricing efficiency. On the other hand, their trading raises market makers' adverse-selection costs, forcing market makers to widen spreads to offset potential losses and thereby reducing liquidity for all market participants. One of the central trade-offs in market mechanism design is precisely how to find the optimal balance between encouraging informed trading to promote price discovery and protecting market makers to sustain liquidity.

The on-chain informational advantage decays rapidly. As ever more participants deploy similar on-chain analytics tools and whale-tracking algorithms, the alpha derived from on-chain data transparency declines at an accelerating pace. At the same time, the spread of private trading channels (such as private mempools and trade-privacy-protection mechanisms like Flashbots Protect) further erodes the transparency-based informational advantage: trading intentions are hidden until the moment of block confirmation, rendering useless those informed strategies that rely on scanning the pending-transaction pool. This constitutes a continual arms race over informational advantage, in which the half-life of any advantage progressively shortens, forcing informed traders to invest ever-higher technical costs to sustain an increasingly fragile marginal edge.

4.2.3 Arbitrageurs

Arbitrageurs profit at no risk or low risk by seeking price discrepancies between different markets or instruments. As argued in Section 4.1.2, arbitrageurs' profits appear on the surface to come from systematic mispricings, but mispricings are themselves jointly created by the inelastic demand of utilitarian traders and the pricing errors of futile traders; arbitrage profits can therefore, ultimately, still be traced to the two sources. By correcting mispricings, arbitrageurs indirectly serve the realization of the hedging premium, allowing hedgers to obtain more accurate execution prices.

In the perpetual futures market, arbitrage takes three core forms. Funding-rate arbitrage is the most common strategy: an arbitrageur establishes a short position in the perpetual futures market while buying an equal amount of the asset in the spot market, thereby earning a positive funding rate. The BIS working paper (Schmeling et al., 2023) shows that crypto carry can generate positive returns in most periods, but its return distribution exhibits pronounced negative skewness and is highly time-varying; much like traditional foreign-exchange carry trades, such strategies face the risk of sudden and severe losses during tail events, with an asymmetric crash exposure hidden beneath the appearance of stable returns [7]. The empirical study of Werapun et al. (2025) further shows that, under certain market conditions, funding-rate arbitrage can even generate excess returns [13]. Cross-market arbitrage profits from the transient price discrepancy of the same asset across different exchanges, a discrepancy usually driven by differences in liquidity and in the speed of price updates. Triangular arbitrage, within a single exchange, exploits inconsistencies in the cross rates of three assets to complete a circular arbitrage.

Characterizing funding-rate arbitrage and cross-platform arbitrage as "risk-free," however, is a dangerous oversimplification. These strategies are more accurately described as low-risk but not risk-free. Their risk exposure manifests along at least three dimensions. First, rate-reversal risk has a significant historical frequency. The sign and magnitude of the funding rate can switch abruptly with market sentiment; in extreme conditions, a funding rate that was positive can turn deeply negative within hours, and the negative period often lasts beyond the arbitrageur's expected tolerance, turning the strategy from a source of stable income into one of persistent losses. Historical data show that during major market events the intraday swing in the funding rate can reach dozens of times its normal magnitude, and this nonlinear jump causes rate-forecasting models built on assumptions of historical mean reversion to fail at precisely the moment they are most needed. Second, on platforms using a unified margin system, the floating loss on the spot long position shares a single margin pool with the perpetual futures short position; when the price of the underlying asset rises sharply, the book loss on the perpetual short and the book gain on the spot long theoretically offset each other, but in extreme volatility the transient deviation between the mark price and the spot price, together with lag effects in margin computation, can cause the account's overall margin ratio to fall below the maintenance level in an instant, triggering forced liquidation. An arbitrageur being liquidated while in a "hedged" state is nearly a paradox in traditional financial terms, yet it is a structural vulnerability of the on-chain unified-margin system. Third, cross-platform arbitrage faces the friction costs of margin fragmentation and withdrawal delay. When an arbitrageur holds a long on platform A and a short on platform B, a sharp one-sided floating loss requires topping up margin on the losing side, while withdrawal of funds on the winning side is often constrained by withdrawal cooling-off periods, on-chain confirmation delays, and congestion on liquidity bridges. This timing mismatch creates a liquidity trap at precisely the moment the market most needs funds to be redeployed, exposing a theoretically perfectly hedged position to one-sided liquidation risk in practice.

The limits-of-arbitrage framework of Shleifer and Vishny (1997) illuminates these risks [14]. The framework's core insight is that arbitrage in the real world is far from the

costless, risk-free activity described in textbook models; arbitrageurs face capital constraints, margin-call pressure, and principal-agent problems, and these frictions can allow a mispricing to widen further before it is corrected. In the crypto perpetual futures market, the applicability of this framework requires a key distinction along the dimension of atomicity. At one end is on-chain atomic arbitrage: under the atomicity guarantee of a single transaction, borrowing, trading, and repayment are completed within the same block, so there is no price risk during a holding period, and the only uncertainty the arbitrageur faces is whether the transaction can be successfully included in a block. Such arbitrage approaches the theoretical zero-risk benchmark, and the capital constraints and margin-call pressure of the Shleifer-Vishny framework do not apply. At the other end is non-atomic position arbitrage, including funding-rate arbitrage and cross-exchange spread arbitrage, in which the arbitrageur must maintain several independent positions over a period of time, bearing the risks of price fluctuation, margin calls, and liquidity mismatch—precisely the dangerous zone the Shleifer-Vishny framework warns against. In tail events, such nominal "arbitrage" strategies may be forced to close at the least favorable moment under margin-call pressure, so that behavior originally intended to correct mispricing itself becomes an amplifier of market volatility. The arbitrageur group in the perpetual futures market therefore spans a broad risk spectrum, from near-zero risk to something indistinguishable from directional speculation.

The composability thesis is directly embodied in arbitrage behavior, but its scope of application must be strictly bounded. The atomic-execution feature of the on-chain environment, combined with the flash-loan mechanism, allows atomic arbitrage between spot decentralized exchanges to be executed with zero capital and zero risk: an arbitrageur can complete the full loop of borrowing, cross-pool trading, and repayment within a single transaction block, something unimaginable in traditional financial markets. This mechanism does not apply to position arbitrage in perpetual futures, however. Funding-rate arbitrage and cross-platform spread arbitrage require maintaining independent positions at different points in time and on different platforms; the holding period spans multiple blocks or even several days and cannot be compressed into a single atomic transaction. Even spot atomic arbitrage itself is not truly "zero-cost": fierce competition among arbitrageurs for the same spread opportunity gives rise to priority gas auctions (PGAs), and the winner often has to pay most of the arbitrage profit as gas fees and tips to the block builder, so the economic rent of arbitrage tends to be consumed by execution costs in the competitive equilibrium. This digitally native atomic-arbitrage mechanism greatly improves the price-convergence efficiency of on-chain spot markets, but its efficiency gains do not automatically extend to the domain of perpetual futures arbitrage, which requires holding positions over time.

4.2.4 Mechanism parasites

The concept of the parasitic trader was proposed by Harris (2003); in traditional financial markets, it manifests primarily as informational parasitism [3]. Informational parasites profit from non-public information obtained by improper means, with typical behaviors including insider trading and front-running client orders—behaviors generally prohibited by law. The on-chain environment of perpetual futures, however, has

given rise to an entirely new form of parasitism: mechanism parasitism. This is a core theoretical contribution to understanding the evolution of on-chain market microstructure.

Mechanism parasitism and informational parasitism differ fundamentally in the source of their advantage, the nature of their behavior, and their legality. To raise this concept from descriptive language to an operational analytical tool, this chapter offers the following definition.

> Mechanism parasitism refers to behavior in which a market participant, while fully complying with the explicit rules of a protocol, exploits the unintended economic consequences produced by the combination of those rules to systematically shift transaction costs or risks onto other participants. It obtains only when three conditions hold simultaneously: (1) the behavior lies entirely within what the protocol's rules permit; (2) the economic consequence exploited is not the intended purpose of the protocol's designers; and (3) the behavior produces a net transfer of value from passive participants to the active exploiter.

Condition (2)'s criterion—"the intended purpose of the protocol's designers"—requires a qualification. In the context of decentralized governance, a protocol's rules are often produced through community proposals and voting, and there is no single, clearly attributable "designer's intent," which makes condition (2) difficult to apply in practice. A more operational formulation is this: the behavior's negative externalities (the harm to other participants and to the overall functioning of the ecosystem) exceed the positive externalities it brings to the ecosystem. This net-externality criterion requires no tracing of subjective intent; it reaches a determination by assessing the behavior's objective economic consequences, and it applies more readily in an on-chain environment where governance is dispersed.

The advantage of mechanism parasites comes not from knowing information others do not know but from understanding the economic implications of the protocol's rules more deeply than others do. What they exploit is a cognitive asymmetry: the protocol's rules are public to all, but their deeper economic effects are not fully understood by every participant. This distinction sets mechanism parasitism apart from ordinary arbitrage: arbitrage corrects market mispricings and pushes prices toward equilibrium, carrying a positive externality; mechanism parasitism exploits rule loopholes to shift costs, does not improve market efficiency, and carries a negative externality.

Figure 4-3 presents, in a side-by-side layout, the evolution from informational parasitism to mechanism parasitism, at the core of which lies a paradigm shift: the source of advantage moves from *exploiting information others do not have* to *understanding public rules more deeply than others do*; its persistence changes from dissipating as information diffuses to enduring structurally; and defense correspondingly shifts from legal punishment to mechanism repair.

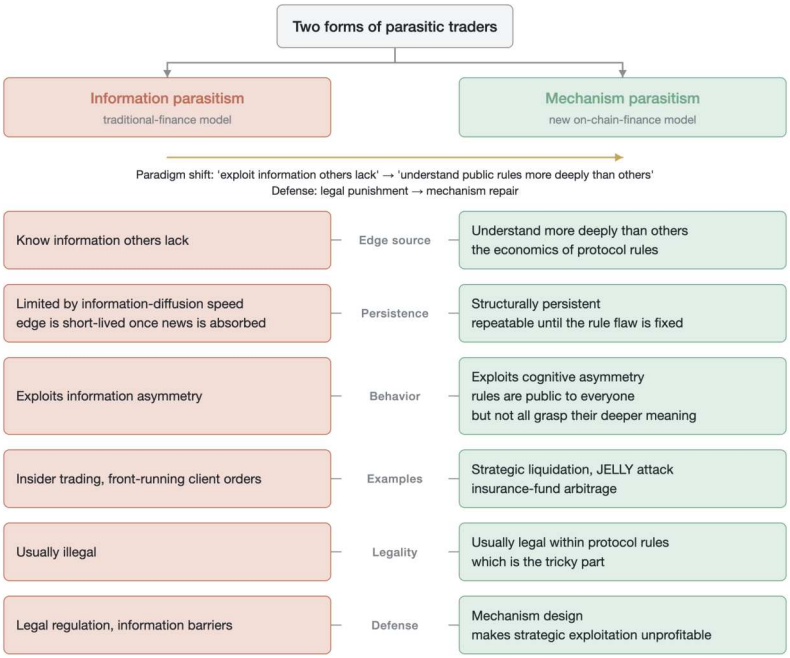


Figure 4-3. An evolutionary comparison of mechanism parasitism and informational parasitism (conceptual comparison diagram; author's synthesis; not empirical data)

The Hyperliquid whale event described in the introduction is a paradigmatic case of mechanism parasitism. The trader's behavior fully satisfies the three conditions of Definition 4-1: the behavior lies within the protocol's rules (condition 1); it exploited the unintended economic consequence that the liquidation penalty was lower than the slippage of closing on the market (condition 2); and it shifted the price-impact cost from itself to passive liquidity providers (condition 3). Actively triggering liquidation thus became a better exit strategy than closing the position oneself. The existence of such an inequality is an inherent loophole in the protocol's mechanism design, not a moral failing of the trader.

Mechanism-parasitic behavior presents a spectrum in on-chain markets, ranging from the mild MEV sandwich attack, to strategic positions exploiting extreme funding rates, to aggressive strategic liquidations, and even to sophisticated attacks exploiting joint loopholes in the liquidation and insurance-fund mechanisms. What these behaviors share is that they are compliant at the level of the protocol's rules. Such compliance, however, is not the same as compliance with the law. The legal standard for market manipulation differs across jurisdictions. The EU Market Abuse Regulation (MAR) adopts an effects-based standard: creating false or misleading signals, or producing an abnormal or artificial price, can constitute manipulation, with no requirement of subjective intent; whereas price manipulation under the U.S. Commodity Exchange Act (Rule 180.2) has traditionally required a specific intent to create an "artificial price," so that a price effect alone does not constitute manipulation, though the post-Dodd-Frank fraud-based Rule 180.1 requires only recklessness. Under these frameworks, even

behavior that fully complies with a protocol's explicit rules may face legal prosecution if its effect constitutes artificial manipulation of the market price or systematic harm to other participants. The legal status of mechanism parasitism therefore remains undetermined, and this constitutes a twofold challenge for defending against it: at the level of mechanism, the economic loopholes that make such behavior profitable must be repaired through redesign; at the level of law, the regulatory framework must respond with new classifications of on-chain behavior and effects-based regulatory principles.

4.2.5 Speculators and futile traders

Speculators seek to profit from judging the direction of prices, and they bear the principal risk in the market. Within the speculator group, however, there is significant differentiation. A minority of informed directional traders have genuine judgment about fundamentals or technicals and can achieve average profits; in economic terms, they can be regarded as a subset of informed traders. The vast majority of speculators, however, are futile traders: they trade on emotion, noise, or herd mentality, use leverage far beyond their risk tolerance, and as a result systematically become the counterparties of profit-motivated traders, bearing persistent losses.

The systematic losses of futile traders constitute the core mechanism of profit transfer in the perpetual futures market. The October 2025 liquidation data cited in Section 4.1.1 (more than 1.62 million accounts liquidated, totaling more than $19 billion, about 87% of them long [4][5]) are precisely its macro evidence: this was not mere market volatility but a systematic wealth-transfer event—futile traders gradually accumulated leverage during a calm period and, during the deleveraging period, transferred their previously accumulated capital, through the liquidation mechanism, to market makers, informed traders, and liquidators.

The extreme one-sidedness of the 87% long liquidations cannot be explained by individual-level overconfidence or leverage abuse alone; it requires the analytical framework of collective behavior. The theory of information cascades proposed by Banerjee (1992) provides a rigorous microfoundation for understanding such extreme uniformity [15]. The theory holds that when individuals make sequential decisions under incomplete information, the rational strategy is to ignore one's own private signal and follow the behavior of predecessors, because a predecessor's choice is taken to embody the aggregate of their private information. Once such following reaches a critical mass, subsequent participants will conform regardless of the private information they hold, forming a self-reinforcing information cascade. In the crypto perpetual futures market, the conditions for an information cascade are structurally reinforced. A persistently positive funding rate constitutes the most intuitive signal of collective sentiment: when the funding rate stays for a long time at an annualized 30% or higher, the implicit information it conveys to the market is that "the overwhelming majority of participants are bullish," and latecomers tend to read this signal as a market consensus rather than a warning of crowding risk. At the same time, the crypto market relies heavily on social media for the transmission of information, and the public trade calls of key opinion leaders (KOLs) constitute a powerful cascade trigger on platforms such as Twitter and Telegram; when a KOL with a large following

publicly announces opening a long position, a considerable proportion of followers will copy the trade directly without independent judgment, further pushing up the price and the funding rate, attracting more imitators, and thereby forming a positive feedback loop. The full transparency of on-chain data plays a dual role here: it makes the position direction of whales and the trading behavior of KOLs into publicly observable signal sources, accelerating the spread of the information cascade, while also making the extreme one-sided concentration of positions visible to all, though no one is willing to be the first to act against it. When the entire market's long positions accumulate to a critical level under the drive of an information cascade, any exogenous shock is enough to trigger the liquidation cascade described in Section 4.3.4, and the one-sidedness of the cascade is precisely a mirror image of the one-sidedness of the earlier herding behavior.

This behavioral pattern arises not merely from inadequate market education but is systematically induced by the product-design features of perpetual futures. Skwarek et al. (2025) confirm that a variety of behavioral biases significantly affect the decision-making of crypto investors [16]. Delfabbro et al. (2021) further note that trading behavior in crypto markets is, in its psychological mechanisms, highly similar to gambling [6]. Specifically, the psychological stickiness of perpetual futures trading can be explained through three classic mechanisms. First, the variable-ratio reinforcement effect: the occasional large profits produced by high-leverage trading constitute an intermittent, large positive reinforcement whose reinforcing power far exceeds that of steady small gains—isomorphic, in psychological principle, to the reward mechanism of a slot machine. Second, the near-miss effect: a trader's experience of narrowly escaping near the liquidation price (for example, when the price rebounds less than 1% from the liquidation line) is subjectively coded as "almost succeeding" rather than "nearly failing," thereby reinforcing the willingness to continue high-leverage trading. Third, the mental-accounting effect: traders tend to treat each trade as an independent mental account rather than part of an overall portfolio, which leads them to take on risk far beyond a rational standard on a single trade while their perception of cumulative losses is systematically blunted. More critically, the around-the-clock, uninterrupted trading of the perpetual futures market eliminates the natural "cooling-off period" provided by the market-closure mechanism of traditional financial markets. In traditional stock or futures markets, the hours between the close and the next open provide traders with a time window to reflect on decisions, assess losses, and restore rationality; in the 24/7 perpetual futures market, this buffer is entirely absent, impulsive revenge trades after a loss can be executed within seconds; the chain of emotion-driven trading can extend indefinitely; and the destructive power of the three psychological mechanisms above is thereby markedly amplified. High leverage, permissionless access, and an emotional community culture together build an environment that attracts and retains futile traders. How these persistent losses economically subsidize market makers' adverse-selection losses is detailed in Section 4.3.3.

4.2.6 Liquidators

In the perpetual futures ecosystem, liquidators clear away systemic bad debt. Their core motive is to profit from the systematic losses of futile traders—that is, to earn

the liquidation penalties paid by those being liquidated. By promptly taking over and closing insolvent positions, liquidators ensure the protocol's solvency and prevent bad debt from turning into systemic risk.

Liquidators' behavior is double-sided under extreme market conditions, however. In a normal market, liquidators and speculators form a functional complement that keeps the system running in good health. But when the market swings violently, liquidators' individually rational behavior can become an accelerator of the liquidation cascade. To trigger liquidation as quickly as possible and earn the penalty, liquidators dump the liquidated assets on the market, and this large-scale selling pressure pushes the market price down further, driving more positions that were originally safe into liquidation.

The specific design of the liquidation mechanism decisively shapes the identity and behavioral patterns of those who execute liquidations, and different platforms differ fundamentally in how they position the liquidation executor. On centralized exchanges, liquidation is essentially part of the platform's internal risk-control process: the exchange's liquidation engine, as a built-in component of the protocol, automatically executes forced closures; the process is invisible to external participants; residual margin (if any) is returned to the trader; the loss shortfall is absorbed by the exchange's insurance fund; and an external independent liquidator does not exist in this model. In decentralized lending protocols (such as Aave and Compound), liquidation is designed as a fully externalized, economically incentivized activity: when a borrower's collateralization ratio falls below the liquidation threshold, any external participant can call the smart contract's liquidation function to repay part of the debt and acquire the collateral at a discount, so the liquidator is a permissionless independent economic actor whose participation is driven entirely by the profit motive. The liquidation mechanism of a perpetual futures DEX falls between these two paradigms, forming a hybrid model. Take Hyperliquid: its liquidation process adopts a layered, escalating architecture. When a trader's margin ratio falls below the maintenance-margin requirement, the protocol first attempts to close the position gradually on the market via limit orders through the internal liquidation engine; if market liquidity is insufficient to complete the closure within a reasonable slippage range, the remaining position is transferred to the HLP market-making vault to be absorbed collectively by passive liquidity providers; and only in extreme cases does the insurance fund step in as the last loss-absorbing layer. In this hybrid model, the "liquidator" is not a single external independent role but a multi-layered defense system jointly composed of the protocol's liquidation engine, the market-making vault, and the insurance fund, and the room for external independent liquidation bots is significantly compressed. dYdX v4 adopts another hybrid path, in which liquidation is executed by validator nodes during the block-proposal process, with liquidation orders embedded in the block-construction flow; this couples liquidation execution with block-ordering power, and a validator can in theory choose strategically when and how to execute a liquidation. Characterizing liquidators broadly as external independent profit-seekers therefore applies only to the purely externalized model of lending protocols; in the actual operation of centralized exchanges and most perpetual futures DEXs, the identity of the liquidation executor is far more complex, involving an interweaving of roles such as the platform's internal engine, passive-vault participants, and validators. Traditional cliff-edge liquidation of-

ten causes a precipitous drop in price, whereas progressive liquidation or vault-takeover designs seek to smooth the liquidation shock. Specifically, the prevailing liquidation mechanisms can be grouped into four types. The first is cliff-edge (full) liquidation, in which the entire position is forcibly closed at once when the margin ratio hits the threshold; its execution is simple but its instantaneous impact on the market price is greatest, and it readily triggers a cascade. The second is progressive (partial) liquidation, in which only the minimum share of the position needed to restore the margin ratio above the maintenance requirement is closed, dispersing the price impact across multiple time windows through incremental reductions. The third is vault-takeover liquidation, in which the protocol's vault directly absorbs the position on the brink of liquidation and closes it out gradually when later market conditions are more favorable, thereby converting the immediate market impact into inventory risk for the vault. The fourth is the auto-deleveraging (ADL) mechanism, in which, when the insurance fund is exhausted or market liquidity is severely insufficient, the protocol forcibly reduces the unrealized profits of traders on the winning side to cover the negative-equity losses of the losing side; although this mechanism effectively prevents systemic default at the protocol level, it does so at the cost of sacrificing the reasonable expectations of winning traders, raising fairness disputes over the redistribution of counterparty risk. In the high-leverage crypto derivatives market, the design of the liquidation mechanism bears not only on the stability of an individual protocol but is also a key line of defense against cross-market risk contagion; the European Securities and Markets Authority (ESMA) (2025), in its analysis of MEV, likewise notes that on-chain microstructure issues such as transaction ordering and cross-venue spreads carry systemic implications [17]. The liquidation executor—whether its identity is the platform's internal engine, an external independent bot, or a passive vault—is the actual operator of this mechanism, and a subtle tension always exists between its pursuit of profit and the goal of system stability.

The legal characterization of liquidation activity is coming under increasingly strict scrutiny across jurisdictions, adding a regulatory-risk dimension to liquidator behavior that cannot be ignored. In the traditional financial system, forced closure is an internal risk-control action of a regulated financial institution, and the licensing qualifications and compliance obligations of the executing entity have a clear legal framework. In the permissionless on-chain environment, however, the legal nature of liquidation activity becomes ambiguous: when an anonymous wallet address forcibly closes another's position by calling a smart-contract function and derives an economic gain from it, this behavior may in some jurisdictions be regarded as providing a financial service. The EU's Markets in Crypto-Assets Regulation (MiCA), for example, imposes authorization and compliance obligations on crypto-asset service providers (CASPs); if liquidation activity were classified as such a service, an anonymous participant would face a fundamental contradiction between compliance requirements and anonymity. At the same time, the Travel Rule promoted by the Financial Action Task Force (FATF) requires virtual asset service providers (VASPs) to transmit the identity information of the originator and beneficiary in a transaction, which poses a direct challenge to on-chain liquidation bots that operate on the basis of anonymity. Although most jurisdictions have not yet explicitly brought on-chain liquidation activity within the regulatory scope of financial-

service licensing, the tightening regulatory trend means that the behavioral pattern of participating in liquidation and extracting profit anonymously may, in the medium term, face a fundamental reassessment of its legal status. This regulatory uncertainty affects not only liquidators' behavioral incentives but may also have a profound effect on the participant structure of the liquidation market: if liquidation activity is ultimately required to be licensed, the current competitive landscape dominated by anonymous bots will be forced to shift toward compliant professional institutions, and both the efficiency and market coverage of liquidation may undergo structural change.

4.3 The game network among participants

Having established the two ultimate sources of profit in the perpetual futures market and the basic features of the six core participant types, we can now construct a dynamic picture of the market. A participant's survival and prosperity in the market depend not on a static identity label but on its position within a complex game network and its capacity to intercept profit flows. This game takes on dynamic features in the digitally native environment that are entirely different from those of traditional financial markets, with high leverage, full transparency, and the determinism of protocol mechanisms jointly reshaping the rules of the game. This section analyzes four core game relationships and, finally, reveals the complete network structure of profit flows across the ecosystem.

4.3.1 Market makers and informed traders

The game between market makers and informed traders is the cornerstone of all market microstructure theory, but in the on-chain environment of perpetual futures its intensity is pushed to a new extreme. As set out in the three-component decomposition of the spread in Section 4.2.1, a market maker's bid-ask spread must compensate for the three costs of order processing, inventory holding, and adverse selection [18], and in the crypto-asset perpetual futures market the nature of this adverse selection has fundamentally changed.

The full transparency of the on-chain environment paradoxically aggravates the market maker's adverse-selection risk: publicly visible positions and liquidation prices are public private information, which requires extremely high data-processing capacity to convert into a genuine trading advantage [12]. Informed traders use this to forecast short-term price movements more precisely and gain the upper hand in the game against market makers (the specific sources of their informational advantage were elaborated in Section 4.2.2). Research has documented substantial adverse-selection costs in crypto markets, from which one can infer that the adverse-selection environment market makers face in supplying liquidity is more severe than in traditional markets [9].

The absence of an expiration date in perpetual futures further amplifies the market maker's inventory risk—as analyzed in Section 4.2.1, inventory risk is thereby rendered indefinite: market makers cannot know when a one-sided trend will end and can only demand a wider spread to compensate for potentially enormous losses [19]. This double pressure is concentrated in the case of Hyperliquid's HLP vault. In the violent market

volatility of October 2025, facing a large volume of non-toxic order flow triggered by liquidations, the vault realized a single-day profit of approximately $40 million [20] (this October 2025 figure comes from an industry analysis by the market maker Kappa Lab; it is not peer-reviewed, and its methodology and data coverage have not yet been independently verified). Yet, facing a liquidation event meticulously engineered by a strategic actor, the same vault bore a loss of approximately $4 million [2]. The concept of "non-toxic" requires a precise distinction here. Liquidation-triggered order flow is indeed low in toxicity along the dimension of informational toxicity, because liquidation is a forced action mechanically triggered by the margin ratio and carries no private information about the asset's fundamentals, so a market maker accepting such flow does not face adverse selection in the classical sense. Along the dimension of liquidity toxicity, however, liquidation order flow is highly toxic, because the one-directional, concentrated selling pressure produced by large-scale liquidations severely shocks the market price, and a market maker—even if it correctly judges that the flow carries no information signal—may still suffer serious inventory losses by absorbing a huge one-sided position. This contrast reveals the fragility of the market maker's profit: its P&L essentially depends on the information content of the counterparty's order flow, and in a permissionless, anonymous network, screening order flow for toxicity is extremely difficult.

4.3.2 Market makers and mechanism parasites

Beyond the traditional informed trader, perpetual futures market makers also face a new type of counterparty unique to the on-chain environment: the mechanism parasite. Following the definition of mechanism parasitism in Section 4.2.4—whose advantage comes not from exclusive information but from understanding the economic implications of protocol rules more deeply than others—this paradigm shift from informational parasitism to mechanism parasitism constitutes one of the most far-reaching structural changes in the microstructure of decentralized finance.

Unlike the informational parasitism of traditional finance, which illegally exploits non-public information (such as insider trading and front-running client orders [3]), the mechanism parasite of the on-chain market operates entirely within the protocol's open and transparent framework of rules, keenly capturing the economic loopholes of the protocol's design in extreme situations and converting them into risk-free or low-risk arbitrage opportunities. The Hyperliquid whale event described in the introduction is a classic case of mechanism parasitism. By deliberately withdrawing margin to raise its own liquidation price, the trader exploited the rule loophole that the liquidation penalty was lower than the slippage of closing on the market, systematically shifting an enormous price-impact cost onto the protocol's passive liquidity providers [20].

This mechanism-parasitic behavior constitutes an unprecedented toxicity for market makers. The traditional adverse-selection model assumes that the informed trader's advantage is temporary and dissipates as information is gradually incorporated into prices [21]. The advantage of the mechanism parasite, however, is structural: as long as the protocol's rule loophole remains unrepaired, the parasitic behavior can be executed repeatedly. What the market maker faces is no longer a random game based

on price expectations but systematic extraction based on a deterministic rule loophole. This new toxic order flow not only directly erodes the market maker's capital base but, more seriously, destroys the economic rationale for passive liquidity provision, forcing market makers to adopt more defensive quoting strategies and ultimately causing a systematic decline in the liquidity of the entire market. Defending against such mechanism parasitism cannot rely on ex post punishment of the actor but must rely on tighter mechanism design that ensures that, in any state, the cost of strategically exploiting the rules exceeds its potential gain.

4.3.3 Profit-motivated traders and futile traders

Once the complexities of strategy and mechanism are stripped away, the profit structure of the perpetual futures market reveals a core fact: the vast majority of the profit that keeps the entire ecosystem running comes, in the end, from the systematic losses of futile traders. The wealth accumulation of all profit-motivated participants—whether market makers, informed traders, or mechanism parasites—depends heavily, at the macro level, on this profit source of futile traders.

Futile traders are usually defined as market participants who expect to profit from trading but, owing to systematic deficiencies in skill, information-processing capacity, or emotional control, persistently lose in the long-run game [3]. In the crypto-asset perpetual futures market, the behavioral patterns of this group closely match the overconfidence theory of behavioral finance. Research shows that investors who lack genuine stock-picking ability but are overconfident tend to trade more frequently, use higher leverage, and ultimately earn worse investment returns [22]. The high leverage of perpetual futures, their around-the-clock uninterrupted trading, and their strong sentiment-driven dynamics provide systematic inducing conditions for such overconfident behavior. The behavioral patterns of many futile traders have already blurred the line between investing and gambling; they tend to chase the market higher during rising periods and, lacking effective hedging strategies, bear volatility exposure far beyond their risk tolerance [6].

From the perspective of market microstructure, the presence of futile traders is the key economic force that resolves the market maker's survival dilemma. In the game against informed traders and mechanism parasites, market makers must bear adverse-selection losses; if the market were composed entirely of rational and fully informed participants, market makers could not sustainably supply liquidity [18]. It is precisely the steady stream of non-toxic order flow supplied by futile traders that allows market makers to earn the spread safely and thus remain profitable overall. Further, when futile traders' positions are forcibly liquidated by adverse price movements, their losses convert directly into penalty income for liquidators and directional profit for short counterparties. The crypto perpetual futures market therefore exhibits a profound dependence: the high running costs of the market infrastructure, the risk compensation of market makers, and the profits of the various arbitrageurs are to a large extent subsidized by the futile traders who lack a competitive edge yet blindly use high leverage.

4.3.4 Speculators and the liquidation mechanism

Leverage is the core driving force of the perpetual futures market, but it is also the key mechanism that transforms individually rational behavior into system-level nonlinear risk. The interaction between speculators and the liquidation mechanism constitutes the most destructive microstructural phenomenon in crypto markets: the liquidation cascade. The market crash of October 2025 provides a paradigmatic observational sample, showing how the leverage cycle evolves through four distinct phases and ultimately triggers a systemic evaporation of liquidity.

In the first phase, the accumulation phase, the market is in a relatively calm uptrend. Attracted by the wealth effect, futile traders continually chase the market higher and gradually increase their leverage multiples [4]. The leverage inflation of the accumulation phase is not a linear process but is driven by a characteristic reflexive positive-feedback mechanism [23]: a rising price pushes the funding rate positive; the positive funding rate is itself read by the market as a signal of long dominance; the bullish narrative on social media spreads accordingly; this narrative in turn attracts new participants with fresh capital; the new participants adopt higher leverage in pursuit of higher returns; and their buying further lifts the price, forming a self-reinforcing loop from price to narrative to capital inflow. In this phase, the system's fragility accumulates in the dark, with the liquidation prices of a large number of long positions densely distributed just below the current market price. To meet the vigorous trading demand, market makers keep expanding their risk exposure. As the price reaches a local high, the market enters the second phase, the critical phase. The first scattered liquidation signals begin to appear; sensitive market makers perceive the rise in order-flow toxicity and begin to reduce their quote depth out of individual rationality; and market liquidity quietly evaporates.

Figure 4-4 presents the four-phase evolution of the leverage cycle: the horizontal axis is time, and the vertical axis shows both the price trajectory and the level of system-wide leverage. The durations of the phases are highly asymmetric—the accumulation phase can last from weeks to months, whereas the cascade phase is often completed within hours to days—and this asymmetry in time scale is the core expression of the nonlinear risk of leveraged markets.



Figure 4-4. A four-phase model of the leverage cycle (conceptual model constructed by the author; the price and system-leverage curves are qualitative illustrations, not real time series; the scale of the October 2025 event is given in the main text [4][5])

When the initial substantive liquidations are triggered, the market plunges rapidly into the third phase, the cascade phase. This is a textbook positive-feedback loop: the liquidation engine automatically dumps the liquidated party's collateral onto the market, and this cost-insensitive market selling triggers a violent price decline in the already thin liquidity pool; the further fall in price then reaches lower liquidation lines, triggering a new round of forced closures [24]. In this phase, market makers withdraw entirely to protect their own capital, liquidity dries up completely, and the price exhibits a sharp, nonlinear decline. The October 2025 event is an extreme embodiment of this process, whose liquidation scale and one-sidedness (more than 1.62 million accounts, more than $19 billion, about 87% long) are detailed in Section 4.1.1 [4][5].

The destructive power of the cascade phase depends largely on microstructural differences in the margin system and on platform-level risk-mitigation mechanisms. Under the isolated-margin model, the liquidation of a single position does not spill over to the trader's other holdings, and the cascade effect is relatively localized; the cross-margin (unified-margin) model that is increasingly prevalent on today's major platforms, however, has an advantage in capital efficiency that is precisely a key accelerator of the cascade phase: under cross margin, a sharp fall in one asset erodes the account's overall margin level, in turn triggering the collateral liquidation of cross-asset positions that were originally safe, forming a cross-asset contagion path. Facing cascade pressure, the platform's insurance fund serves as the first buffer in the third phase: when the liquidation engine cannot complete a closure on the open market at a price better than the bankruptcy price, the insurance fund absorbs the shortfall, preventing losses from being passed directly onto winning traders. The insurance fund is finite in size, however, and in an extreme tail event, when the loss rate of successive liquidations exceeds what the insurance fund can bear, the protocol is forced to activate the auto-deleveraging mechanism, forcibly reducing the positions of counterparties in order of profit to achieve a systemic release of risk. Although the auto-deleveraging mechanism itself effectively

prevents default at the protocol level, its forced closure of winning traders' positions carries a significant negative externality, further intensifying the crisis of trust among market participants and the contraction of liquidity.

Another key feature of the cascade phase is cross-platform contagion. In today's highly fragmented crypto derivatives market, centralized exchanges and decentralized protocols form contagion chains through three main channels. First, cross-platform arbitrageurs hold hedged positions simultaneously across multiple venues; when liquidation on one platform forces them to close, their hedging positions on other platforms are unwound as well, transmitting selling pressure to markets that were originally unaffected. Second, market makers' synchronized withdrawal under market stress has a cross-platform character: when order-flow toxicity spikes at one venue, market makers, out of risk-management considerations, simultaneously reduce their quote depth at all venues, causing liquidity to evaporate synchronously across the whole market. Third, the mark price of perpetual futures is usually computed from a spot-price index across multiple exchanges, so a price collapse on one platform is transmitted, through the index weights, to the mark prices of other platforms, triggering liquidations that should not have occurred. At the same time, the liquidation process itself has several failure modes: an oracle may lag, malfunction, or even be manipulated during extreme volatility, causing the mark price to deviate from the true market price. The zero-price liquidation event at MakerDAO in March 2020 is a paradigmatic case: under conditions of severe congestion on the Ethereum network, a liquidator won an auction for collateral worth millions of dollars with a zero bid, exposing the systemic risk that the liquidation mechanism may fail completely when the network is congested [25]. A deeper nonlinear effect is that the price impact of the cascade phase is not proportional to the scale of liquidation but is amplified exponentially as liquidity dries up: when order-book depth contracts to one-tenth of its normal level, the price slippage caused by a market sale of the same size may be dozens of times that under normal conditions, and this nonlinear price impact in a liquidity vacuum is precisely the fundamental reason the cascade phase is far more destructive than a linear expectation would suggest.

These factors together point to the fourth phase, the recovery phase: after the deleveraging process is complete, the market enters recovery. The level of system-wide leverage falls sharply, market makers slowly restore liquidity provision after reassessing risk, the price stabilizes at a new low, and the market awaits the start of the next cycle. This nonlinear, accelerated decline driven by the liquidation mechanism reveals the structural feature of decentralized markets in which the superposition of individually optimal behavior leads to systemic fragility [26].

4.3.5 The complete map of profit flows

Having analyzed the four core game relationships in depth, we can now draw the complete topological structure of profit flows in the perpetual futures market. This is not a simple zero-sum game but a complex directed network with clear tiers and directions, in which each type of participant occupies a distinct niche based on its informational advantage, capital size, and depth of understanding of the mechanisms.

In this ecological network, futile traders and utilitarian traders constitute the system's

foundational profit sources. The hedging premium that utilitarian traders voluntarily pay to hedge risk, and the vast systematic losses that futile traders incur through directional misjudgment and liquidation, converge into the two profit sources that drive the entire market. Market makers occupy the central hub of the network: on one side they earn a stable spread from non-toxic order flow, and on the other they inevitably transfer profit, through adverse selection, to those with an informational advantage. Informed traders, by virtue of their informational advantage; mechanism parasites, by virtue of their deep understanding of the rules; and arbitrageurs, by virtue of their speed and capital advantages, together constitute the group of participants with the strongest profit-extraction capacity in the ecosystem. They profit not only directly from futile traders' directional errors but also continually extract value from market makers' inventory.

Figure 4-5 charts the complete topology of profit flows: the direction and thickness of each edge indicate the direction and relative magnitude of a profit flow, revealing a three-tier structure—source (utilitarian and futile traders) → market-maker relay → terminal (informed traders, mechanism parasites, and arbitrageurs). The contrast between the market maker's inbound edges (spread income) and outbound edges (adverse-selection losses) vividly captures the *squeezed-from-both-sides* predicament. The protocol fees and MEV extracted throughout the network are precisely the microstructural source of the negative-sum quantity in Section 4.1.4.

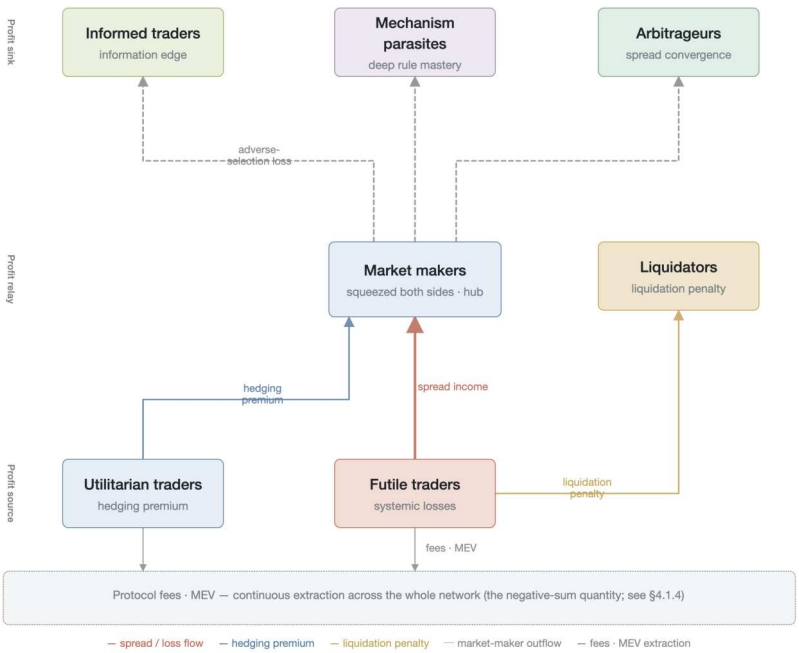


Figure 4-5. The complete map of profit flows (conceptual diagram of profit flows: nodes and edges represent mechanistic relationships; edge thickness denotes qualitative relative magnitude, not a measured scale)

At the same time, the entire game network is subject to continual external extraction.

Protocols and exchanges extract system value through trading fees, while the underlying blockchain validators and miners capture another portion of the profit through network fees and extractable value. This gives the perpetual futures market its essentially strict negative-sum character (as captured by the negative-sum formula in Section 4.1.4). Understanding this overall picture of profit flows is essential to grasping market structure: it shows that, in a permissionless, fully transparent, and highly leveraged digitally native environment, wealth flows with relatively high efficiency from participants who lack information and mechanism understanding to those who command advantages in algorithms, speed, and rules. The topological structure above is a simplified model that omits several important feedback loops and spillover paths. First, the funding rate of perpetual futures feeds back into the pricing of the spot market, and changes in the spot price feed back, through oracle price feeds, into the mark price of perpetual futures, forming a cross-market cyclical feedback that may produce a self-reinforcing spiral in extreme conditions. Second, the model does not incorporate inter-protocol competition effects: when multiple perpetual futures protocols compete for the same pool of liquidity providers and traders, the direction and scale of profit flows adjust dynamically with changes in the incentive structures across protocols. Finally, the profit-flow network is not closed within the crypto market but spills over to the external financial system through several paths. An extreme deleveraging event, for example, may trigger depegging pressure on stablecoins, and the crypto exposure that institutional investors hold through ETFs or structured products may transmit the shock of on-chain deleveraging to traditional financial markets, forming cross-system risk contagion. These omitted dimensions do not alter the core structure of the profit-flow topology, but they remind the reader to weigh the model's boundary conditions carefully when applying it to extreme scenarios. This picture of profit flows provides the empirical basis for subsequent chapters to explore how market quality can be improved through mechanism design.

4.4 The threefold reshaping by the digitally native environment

By rewriting the basic rules of trading, the digitally native environment fundamentally changes participants' behavioral patterns. In traditional financial markets, participants' identities are fixed, the game plays out mainly among human traders, and the information environment is highly opaque. In the permissionless, fully transparent, and highly composable on-chain perpetual futures market, however, these basic assumptions are thoroughly overturned. This environment not only lowers the barrier to entry but also gives rise to entirely new forms of the game. Specifically, the digitally native environment reshapes participant behavior along three core dimensions: the shift of roles from fixed identities to fluid states, the evolution of the game from human decision-making to algorithmic execution, and the transformation of transparency into systemic fragility in a high-leverage environment. Together, these three transformations construct an entirely new ecosystem that resembles traditional finance on the surface but operates on a fundamentally different underlying logic.

4.4.1 The fluidization of roles

In the traditional financial system, participants' roles are highly rigid. This rigidity stems mainly from strict licensing regimes, high capital requirements, and complex compliance-review mechanisms. Market makers, brokers, clearing institutions, and speculators each occupy a clearly defined market niche, and crossing these boundaries demands very high institutional costs. The classification of participants in the traditional market therefore tends to present a static, tree-like structure, in which each entity typically plays only one or a few fixed roles over its life cycle.

The permissionless nature of on-chain finance thoroughly breaks down this role rigidity. In a perpetual futures protocol, a participant's identity is no longer a fixed label but a fluid state that can be switched at any time. A single wallet address can play multiple roles simultaneously at different points in time, or even within a single complex composite transaction. This role fluidity can be abstracted as a state-transition diagram containing multiple nodes—speculation, market making, arbitrage, liquidation, governance, and others—between which the cost of transition approaches zero.

Figure 4-6 uses a state-transition diagram to show how on-chain participants switch market roles at near-zero cost (five nodes: speculation, market making, arbitrage, liquidation, and governance). Its defining feature is near-complete connectivity—a direct transition path exists between almost any two nodes—in sharp contrast to the sparsely connected state-transition diagram of traditional finance, where licensing regimes and compliance requirements prevail.

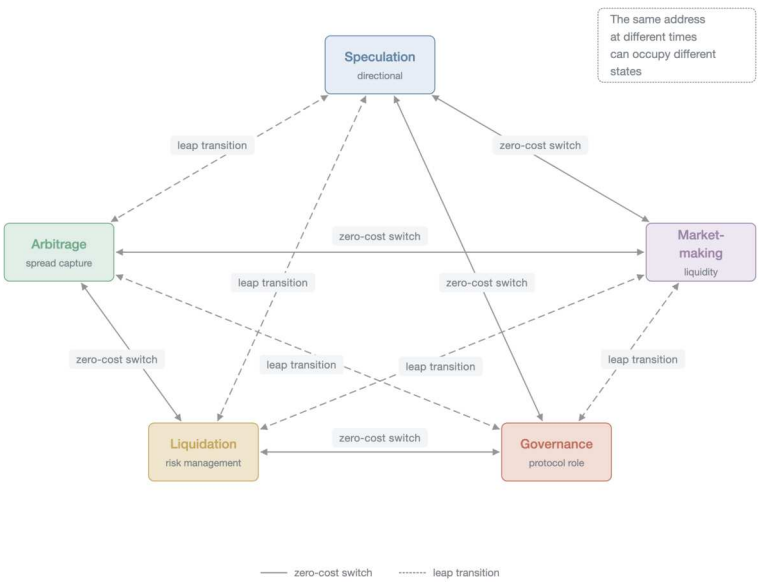


Figure 4-6. A state-transition diagram of participants (conceptual illustration of role-state transitions; not empirical data)

This free transition between states is not merely a theoretical possibility; it has already

found preliminary support in on-chain data. An independent analysis of Hyperliquid trading data indicates that the roughly 15% most active addresses in the market participated in at least two different types of market activity within a single calendar month—for example, conducting directional speculation while also supplying liquidity to the protocol's vault—and that these addresses contributed more than 40% of total trading volume [27]. These data, however, come from a non-peer-reviewed independent analysis whose sample coverage, method of clustering address ownership, and robustness of statistical inference the academic community has not independently verified; the specific figures above should therefore be regarded as order-of-magnitude estimates rather than precise conclusions. Even so, this preliminary evidence points to a trend worth attention: role fluidity is driven mainly by professional participants. They exploit the low-friction feature of the digitally native environment to adjust their functional positioning in the market dynamically, so as to maximize capital efficiency and returns.

The fluidization of roles profoundly changes the meaning of specialization. On-chain finance greatly lowers the institutional threshold for switching roles, enabling professional participants to integrate cross-domain strategies more efficiently while also making it possible for ordinary users to engage in complex activities such as passive market making. This does not mean, however, that the market is moving toward de-specialization. On the contrary, the barrier of specialization shifts from license-based institutional admission to a cognitive and technical barrier based on mechanism understanding, algorithmic technology, and capital scale. Under this new paradigm, a participant's competitiveness no longer depends on the identity label it holds but on its ability to switch efficiently between states and capture profit. Role fluidity may in practice take two starkly different behavioral forms. The first is rational strategic diversification, in which a participant, based on a deep understanding of market microstructure and its own comparative advantage, makes prudent dynamic allocations across states to optimize its overall risk-adjusted return. The second is overconfidence-driven strategy drift, in which a participant, after achieving short-term success in one role, rashly expands into unfamiliar territory without adequate cognitive preparation—for example, a successful directional speculator turning to active market making without fully understanding adverse-selection risk. The former is a positive product of the low-friction environment; the latter may accelerate the erosion of capital. The line between the two is often hard to draw clearly ex ante, but it usually shows up as a significant difference in the ex post P&L data.

4.4.2 The algorithmization of the game

The digitally native environment not only changes the form of participants' roles but also fundamentally reshapes how the game is executed. In on-chain markets, the transparent rules defined by smart contracts and the public mempool provide structurally favorable conditions for an algorithmic game. Economic bots (including MEV searchers, liquidation bots, and arbitrage bots) have become a decisive force in market microstructure, shaping market behavior even more than human traders do.

The behavioral mechanism of MEV searchers is a paradigmatic example of the algorithmization of the game. On general-purpose blockchains such as Ethereum mainnet,

these algorithmic programs continually scan the public mempool for large orders or pricing discrepancies they can exploit. Once an opportunity is found, they execute front-running or sandwich attacks by paying a higher gas fee to block builders to ensure their transactions are packaged first. According to data from Flashbots and EigenPhi, the cumulative MEV extracted on Ethereum mainnet (including front-running, sandwich attacks, and arbitrage) has been estimated on the order of $1 billion to $2 billion since 2020; a further distinction matters here: the payments validators receive via MEV-Boost reach several billion dollars more, and the two figures rest on different bases and should not be conflated [28]. Although these data mainly reflect the scale of value extraction in Ethereum-mainnet spot trading and general DeFi scenarios, the core logic they reveal—that the algorithmic game has evolved from a fringe strategy into a core profit-distribution mechanism of on-chain finance—applies equally to the perpetual futures market.

The specific forms of MEV in perpetual futures protocols, however, differ fundamentally from the mempool game on Ethereum mainnet. In perpetual futures protocols that adopt an application-chain architecture (such as Hyperliquid), transaction-ordering power is concentrated in the protocol's own sequencer, and the resulting ordering-power MEV constitutes a distinctive attack surface: the sequencer operator, or a participant colluding with it, can extract excess profit by manipulating the order of transaction execution. In addition, front-running oracle updates is another important form specific to the perpetual futures market: an attacker exploits the time-difference window between the oracle's price update and the actual change in the market price, establishing a position ahead of the new price taking effect to capture a deterministic profit. These two forms differ in mechanism from the gas-auction-based front-running of the Ethereum mempool: the latter depends on the visibility of trading intentions in the public mempool and on block builders' ordering auctions, whereas the former exploit, respectively, the privileged position of a centralized sequencer and the discrete temporal structure of oracle updates. This distinction matters theoretically, because different attack surfaces call for different defensive strategies and institutional designs.

The essence of MEV is a direct manifestation, at the level of trade execution, of the transparency paradox discussed in Chapter 1. The public mempool makes all trading intentions visible to the entire network, but this visibility does not translate equally into everyone's benefit. Only searchers with the fastest computation and the optimal network topology can convert this transparency into actual profit. In other words, in an algorithm-dominated game, all participants can observe the same information, but only the algorithm with the highest computation speed can convert it into a trading advantage. This dynamic places ordinary users at a systematic disadvantage in the process of trade execution; they often pay an implicit cost for the searchers' services without being aware of it.

Faced with the negative externalities of the algorithmic game, the crypto community's response is not to eliminate MEV entirely but to manage and regulate it through institutionalized means. Flashbots, for example, has introduced a sealed-bid auction mechanism, the MEV-Share protocol achieves a partial return of profit to users, and the Ethereum protocol layer has advanced the architectural design of proposer-builder separation (PBS) [29]. The core goal of these mechanism innovations is to reduce, through

market-based means, the degree of exploitation of ordinary users, while acknowledging that the algorithmic game is unavoidable. This process of institutionalization bears a deep structural resemblance to the regulatory evolution of high-frequency trading in traditional financial markets, both reflecting the market's difficult balance between the pursuit of efficiency and the preservation of fairness.

4.4.3 The transparency-fragility paradox

In the low-leverage spot market, the full transparency of on-chain data is generally regarded as a positive attribute, one that makes the market highly auditable, monitorable, and verifiable. When this transparency is combined with the high-leverage feature of the perpetual futures market, however, its nature fundamentally reverses. This section proposes the core concept of the transparency-fragility paradox to clarify how, under specific conditions, the absolute openness of information transforms into structural fragility for the system.

In the on-chain perpetual futures market, every trader's position size and margin balance are directly visible to all through the public state of the smart contract, and although the liquidation price is not stored directly in on-chain state, it can be computed precisely from the public maintenance-margin parameters and the mark-price formula, so that in practical effect it is fully public. When a trader uses very high leverage, the liquidation price is often very close to the current market price. This combination creates a highly attractive attack vector: if an attacker with sufficient capital can precisely compute the liquidation price of a high-leverage long, and assesses that the cost of pushing the market price down to that trigger point is lower than the potential gain from the liquidation, then targeting that liquidation becomes a rational strategy with positive expected value. More importantly, this attack logic is not confined to the targeted sniping of a single position but can extend to the system level. Because all position data can be aggregated and analyzed, an attacker can construct a complete liquidation heatmap—marking the total volume and cumulative scale of positions that would be triggered for liquidation at each price level. Such a heatmap reveals the critical points of the liquidation cascade along a downward price path: when the price is pushed down to the first dense liquidation band, the forced selling pressure released will push the price down further and trigger the next tier of liquidations, forming a self-reinforcing cascade. A system-level liquidation-cascade attack—one that targets not a specific position but the triggering of the entire liquidation chain—therefore becomes a rational behavior with higher expected return in an environment where transparency and high leverage coexist.

Figure 4-7 depicts the transparency-fragility paradox as a side-by-side comparison: the same set of fully transparent data enhances market auditability and risk assessment in a low-leverage setting, yet in a high-leverage setting combines with the attack capacity of large capital to make the targeted triggering of others' liquidations a strategy with positive expected value. The key dividing line between the two settings is precisely the level of leverage.

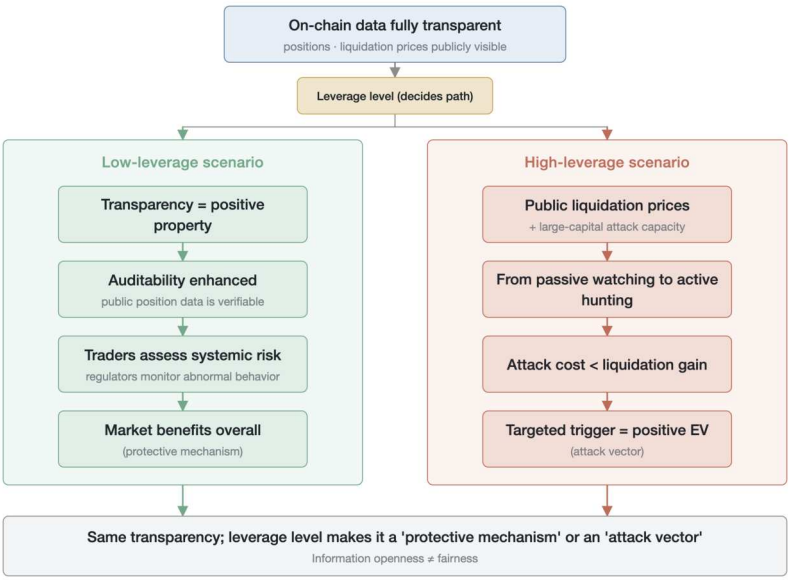


Figure 4-7. The mechanism of the transparency-fragility paradox (schematic of the transparency-fragility paradox mechanism; not empirical data)

The Hyperliquid whale event described in the introduction is an extreme demonstration of this paradox. In that event, the trader was not passively liquidated but actively exploited the fact that its own liquidation price was publicly visible, turning liquidation into a strategic tool [30]. This shows that in a transparent, high-leverage environment, the depth of one's understanding of the economic implications of transparent data is a more decisive competitive advantage than the acquisition of information itself.

The transparency-fragility paradox reveals a deep contradiction in on-chain market design: the openness of information is not necessarily equivalent to the fairness of the game. When all participants' position information is fully public, and some participants possess the capacity to use this information to carry out targeted liquidations, transparency instead intensifies asymmetric profit extraction in the market. This paradox not only makes concrete, at the level of individual participants, the macro transparency thesis of Chapter 1; it also provides a core theoretical starting point for subsequent chapters exploring the redesign of the liquidation mechanism and the protection of liquidity.

4.4.4 The compounding effect of the threefold reshaping

The fluidization of roles, the algorithmization of the game, and the transparency-fragility paradox—these three transformations do not occur in isolation but are interwoven and act jointly, ultimately shaping an entirely new trading ecosystem that resembles traditional finance in form but is fundamentally different in nature.

Table 4-2 summarizes, along five dimensions—role form, principal actors in the game, information environment, competitive advantage, and source of fragility—the systematic differences in participant behavior between traditional finance and the on-chain

perpetual futures market. The thread running through all dimensions is this: the core shift of the on-chain market is not the elimination of the constraints of traditional finance but the transformation of the nature of those constraints, from institutional barriers to cognitive and technical barriers—the underlying logic of the financial game is unchanged, but the microstructural mechanisms that realize this logic have fundamentally changed.

Dimension	Traditional financial markets	On-chain perpetual futures market
Role form	Fixed identities (based on licensing and admission)	Fluid states (based on zero-cost state transitions)
Principal actors in the game	Dominated by human decision-makers	Dominated by algorithms and economic bots
Information environment	Opaque and asymmetric	Fully transparent, with asymmetric understanding of mechanisms
Competitive advantage	Proprietary information and execution channels	Depth of mechanism understanding and speed of algorithmic execution
Source of fragility	Hidden leverage and opaque counterparty risk	Public high leverage and the risk of being targeted for liquidation

Table 4-2. A comparison of participant behavior in traditional finance and the on-chain perpetual futures market (Data source: author's compilation)

In this new ecosystem, where the three transformations compound one another, traditional market microstructure theory is partly invalidated. The adverse-selection risk market makers face, for example, no longer comes only from informed traders with inside information but increasingly from mechanism parasites who can precisely compute liquidation cascades and deeply understand protocol loopholes. Likewise, retail investors' losses arise not only from misjudgment but also because they are at a systematic disadvantage, in a fully transparent environment, against high-frequency algorithms and targeted-liquidation strategies.

The purpose of this comparison is not to judge the relative merits of the two systems but to establish a key cognitive benchmark: the underlying rules of the game in the on-chain perpetual futures market have changed irreversibly. Understanding this paradigm shift—from fixed identity to fluid state, from a human game to algorithmic confrontation, and from opaque protection to transparent fragility—is a necessary premise for analyzing the market's ecological structure in depth, assessing its systemic risk, and designing more robust protocol mechanisms.

4.5 AI agents: the rise of a seventh participant type

The six participant types analyzed in Section 4.2 (market makers, informed traders, arbitrageurs, mechanism parasites, speculators, and liquidators) all follow a common pattern: humans design strategies, which are then executed by humans or by simple algorithms. The emergence of the AI agent breaks this pattern: it can autonomously

generate strategies beyond executing preset ones, process unstructured information, and discover and exploit opportunities across markets, protocols, and chains. The rise of AI agents pushes the threefold reshaping identified in Section 4.4 (role fluidity, algorithmization of the game, and the transparency-fragility paradox) to its extreme, and gives rise to a series of unprecedented problems.

4.5.1 From automation to autonomous agency

The economic bots analyzed in Section 4.4.2 (MEV searchers, liquidation bots, and arbitrage bots) are automated agents. They execute deterministic operations under deterministic conditions according to deterministic rules preset by humans. The logic of an arbitrage bot, for example, might be "if the spread exceeds X, execute the arbitrage; otherwise, wait." The behavior of such agents is entirely predictable, because humans write the rules and those rules do not change.

AI agents based on large language models (LLMs) are autonomous agents, and they differ from automated agents in three essential respects. In the source of strategy, an automated agent's strategy is designed by humans and executed by the machine, whereas an autonomous agent can autonomously generate, evaluate, and adjust strategies, continually optimizing them through market interaction. In information-processing capacity, an automated agent processes only structured data (prices, order books, on-chain state), whereas an autonomous agent can process unstructured information (news, research reports, social media, regulatory texts). In adaptive capacity, an automated agent's rules are fixed and require manual updating when the environment changes, whereas an autonomous agent continually learns and adapts within market interaction.

This leap is not a quantitative change but a qualitative one. Formally, the behavior of an automated agent can be modeled as a deterministic finite automaton—given the state and input, the output is entirely determined—whereas the behavior of an autonomous agent is closer to a partially observable Markov decision process, whose strategy adjusts dynamically based on incomplete observation of the environment and whose output carries intrinsic uncertainty.

Industry data corroborate the speed of this leap. Messari (2024), in its *Crypto Theses 2025*, offers a qualitative assessment of the explosive growth of AI agents but gives no specific target for assets under management [31]; as of mid-2026, there is still no authoritative statistical basis for the on-chain assets managed by AI agents, and the specific figures circulating in the market are mostly third-party estimates that should be treated with caution. Even so, a conservative estimate of the growth rate hints at a rapidly approaching turning point: when the assets managed by AI agents reach some critical proportion of the total market, the market's microstructure will change qualitatively.

4.5.2 AI agents' competition for the two profit sources

When AI agents are incorporated into the profit-source framework of Section 4.1, they do not simply add a new type to the six participant types; rather, they launch multidimensional competition simultaneously across both profit sources and in multiple roles.

In the competition for the losses of futile traders, AI agents first participate as informed traders. An AI agent can simultaneously read and understand all public news, on-chain data, social-media sentiment, and regulatory developments, and generate directional judgments on this basis, with a breadth and speed of information processing far beyond those of a human informed trader. The study of Lopez-Lira and Tang (2023) shows that when LLM agents capable of reading news are introduced, the market's price-discovery efficiency improves significantly and the degree of information asymmetry falls sharply [32]. This means that traditional informed traders may become the informationally disadvantaged party relative to AI agents; the source of the informational advantage shifts from exclusive information to reasoning and association capacity, and the dimension of competition shifts from an information gap to an intelligence gap. At the same time, the presence of AI agents also accelerates the losses of futile traders, because AI counterparties price more precisely and leave less room for human error. The more widespread AI agents become, the more the total losses of futile traders may in fact decline—not because there are fewer futile traders, but because each one loses faster, survives for a shorter time, and thus contributes fewer ongoing losses. This may have an indirect negative effect on market liquidity. There is, however, a countervailing effect worth attention: the "AI narrative" itself may attract a new wave of futile traders into the market—that is, retail investors drawn to AI-driven trading bots, harboring unrealistic fantasies about their expected returns and thereby becoming a new source of losses. More subtly, poorly designed AI trading agents may themselves evolve into an entirely new category of "futile AI traders," which incur systematic losses through overfitting to historical data, model drift, or adversarial attack; their behavioral patterns appear formally as algorithmic rationality, but in economic consequence they are no different from human futile traders, and may even amplify the total volume and speed of losses through large-scale deployment.

In the competition for the hedging premium, AI agents achieve, in the role of market maker, a multidimensional optimization that human market makers cannot reach. An AI agent can manage inventory across multiple exchanges and multiple trading pairs simultaneously, assess the toxicity probability of each order in real time, and adjust quote width and depth within milliseconds. The cost identified in Section 2.2.5—that abolishing the expiration date renders market-making risk indefinite—may be partly mitigated in the hands of an AI market maker, because the AI's inventory management can hedge dynamically across assets and markets, achieving a multidimensional dispersion of risk beyond the reach of a human market maker. Yet AI market makers also intensify the competitive pressure on human market makers. In a market where AI market makers can offer narrower spreads and greater depth, the survival space of human market makers is structurally compressed, and market making may evolve from a professional activity with a technical barrier into an oligopolistic activity whose barrier is AI infrastructure.

In addition, AI agents can identify and exploit a protocol's token incentives more precisely, automatically switching between protocols to maximize returns, thereby further accelerating the efficiency-driven mining of protocol token incentives and raising the cost to a protocol of acquiring liquidity through token incentives.

4.5.3 Differentiated impact on the six participant types

The rise of AI agents has an asymmetric impact on the six participant types of Section 4.2. Some roles are strengthened, some weakened, and some face questions of survival.

Figure 4-8 presents, as impact vectors, the differentiated effects of AI agents on the six participant types (a highly uneven distribution): market makers and arbitrageurs polarize; informed traders and mechanism parasites experience an *arms-race escalation* (the dimension of competition rises from an information gap to an intelligence gap); speculators and futile traders are eliminated at an accelerating pace; and liquidators move toward full automation (AI displaces rule-based liquidators, penalties are competitively compressed, and homogenization risk is introduced). The overall effect is a further concentration of the participant structure.

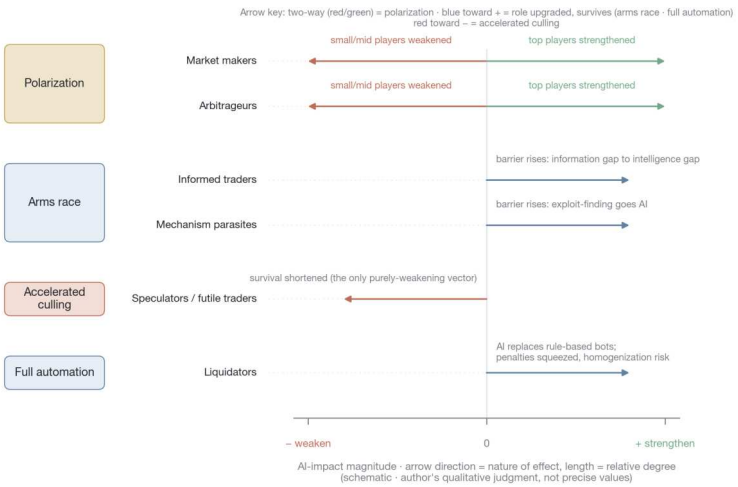


Figure 4-8. The differentiated impact of AI agents on the six participant types (the AI impact reflects the author's qualitative judgment; direction = strengthening/weakening, length = relative intensity; not empirical data; the horizontal-axis scale does not represent measured values)

Market makers: polarization. Top market makers gain an overwhelming advantage by adopting AI, able to operate in more markets simultaneously, manage inventory more efficiently, and control risk more precisely. Small and medium market makers are gradually eliminated, unable to compete with AI market makers on speed and precision. The overall result is a rise in market-making concentration; liquidity may be deeper but comes from more concentrated sources, and system resilience may decline.

Informed traders: the informational advantage redefined. The informed advantage based on processing public information is leveled by AI, and the space for survival shifts toward a combination of exclusive data sources, deeper reasoning, and lower latency—advantages that are themselves being gradually eroded by AI. The dimension of competition shifts from an information gap to an intelligence gap, and the barrier to informed trading rises rather than falls.

Arbitrageurs: greater efficiency but compressed profits. AI makes arbitrage

execution faster and more precise, but competition also intensifies. Arbitrage profits trend toward zero, and market efficiency improves. The sustainability of arbitrage as a maintainer of market efficiency is called into question.

Mechanism parasites: an arms-race escalation. AI makes the discovery of mechanism loopholes faster and more systematic. The precision of strategic liquidation and insurance-fund arbitrage rises. A protocol's mechanism design must contend with AI-level loophole-discovery capacity, and the liquidation designs of Chapter 11 and Chapter 29 face a higher standard.

Speculators and futile traders: accelerated elimination. Facing AI counterparties, human speculators' informational disadvantage widens further and their losses accelerate. Futile traders' losses are generated faster but may decline in total volume, and the market may face the long-term problem of futile-trader depletion.

Liquidators: full automation. AI-driven liquidation bots far surpass rule-based bots in speed and strategic precision. Competition in the liquidation market intensifies, and liquidation penalties are competitively compressed. AI liquidation systems, however, also introduce new operational risks: adversarial inputs (such as carefully constructed price sequences) may mislead the decisions of AI liquidation bots, causing them to execute liquidations at unfavorable moments or to misjudge the value of liquidated positions; a population of liquidators built on similar model architectures carries homogenization risk and may fail simultaneously or make the same misjudgment under extreme market conditions; and the continual evolution of market microstructure may cause model drift, rendering parameters trained on historical data ineffective in a new market state.

A key insight is this: the impact of AI agents is not uniform—it amplifies the market's inequality. The gap between participants who have AI infrastructure and those who do not may be larger than the gap between institutions and retail investors in traditional finance. This raises a question of fairness at the level of Chapter 25 (market quality): "Does the speed gulf between AI participants and human participants require an institutional response?"

4.5.4 Three entirely new forms of the game

AI agents not only change the competitive landscape of existing participants but also give rise to three entirely new forms of the game that do not exist in the game network of Section 4.3.

The first form of the game arises between AI and AI, and its core risk is strategy homogenization. When most of the market makers and arbitrageurs in a market are AI agents built on similar large models, their response patterns to the same signal may be highly similar. When an anomalous event occurs, all the AIs may simultaneously widen spreads, skew quotes to the same side, hit their risk-control thresholds, and reduce exposure. This is faster and more thorough than the synchronized liquidity contraction of human market makers analyzed in Section 2.4.2 and Section 4.3.4, because the AIs' synchronization is at the millisecond level, whereas human liquidity contraction is dispersed over at least minutes to hours. The AIs' recovery may also be faster, however,

because AI is not subject to behavioral biases and can immediately resume quoting once the triggering condition disappears, which may give crises a sharper but shorter V-shaped profile. Strategy homogenization arises because identical model architectures produce identical outputs for identical inputs; their predictability is higher than that of human behavior, and defenses must be designed at the level of model diversity. This risk is not a purely theoretical inference but has historical precedent in traditional quantitative finance. In the quant crisis of August 2007, quantitative hedge funds using similar multi-factor models unwound positions synchronously within a few days, causing what were originally highly diversified portfolios to suffer anomalous losses of several standard deviations [33]. The core lesson of that event is that strategy homogeneity is itself a source of systemic risk: when a sufficient number of participants make the same decision based on the same signal, the liquidity depth the market exhibits in normal times evaporates in an instant during a stress event. In an AI-dominated market, the time scale of the cascade dynamic may be fundamentally compressed. A traditional quant crisis usually follows a four-phase model: weeks of position accumulation, the emergence of a triggering event, several days of cascade unwinding, and weeks of market recovery. When the principal actors of the game shift from human quant traders (whose decision cycles are hours to days) to AI agents (whose reaction times are at the millisecond level), however, the four phases above may collapse into a two-phase structure: slow accumulation of homogeneous positions and a near-instantaneous cascade collapse. The intermediate "gradual unwinding" phase is eliminated because all the AIs hit their risk-control thresholds almost simultaneously, turning the crisis from a gradual evolution into an abrupt, phase-transition-like event.

The second form of the game arises between AI and humans, and its core feature is a structural asymmetry in information-processing capacity. When an AI agent can process thousands of data sources within milliseconds—simultaneously assessing fundamentals, technicals, on-chain data, sentiment, and cross-market signals—while human traders still rely mainly on limited information and intuition, the capability gap between the two is systematic. This gap is especially pronounced in the processing of unstructured information. Take the appearance of a piece of regulatory news: a human needs minutes to hours to read, understand, and judge its impact, whereas an AI agent can complete processing and execute a trade within seconds. This is a structural difference arising from a fundamental asymmetry in information-processing capacity, not an asymmetry in the right to acquire information. From the perspective of the profit-source framework of Section 4.1, this means AI agents can extract value more efficiently from the systematic losses of futile traders, but it may also cause this profit source to shrink because of the accelerated exit of futile traders. The transparency-fragility paradox of Section 4.4.3 acquires a new dimension here: the threshold for exploiting transparency rises further, from computation speed to reasoning intelligence.

The third form of the game arises between AI agents and protocol mechanisms, and its essence is the intelligent-scaling of mechanism parasitism. The mechanism parasitism defined in Section 4.2.4 will be markedly amplified in the AI era. Current mechanism parasitism is realized mainly by human strategists who study protocol documentation in depth, and discovering an exploitable loophole usually takes days to weeks; an AI agent, by contrast, can systematically analyze the smart-contract code of all on-chain

protocols, simulate different attack paths and economic consequences, discover in hours the mechanism loopholes that would take a human weeks to find, and automatically construct and execute strategic attacks. This means a protocol's mechanism design must be able to withstand AI-level loophole-discovery capacity. The study of Dou, Goldstein, and Ji (2025) further notes that even without explicit instruction, AI agents trained under the same reinforcement-learning paradigm may develop tacit-collusion strategies in interaction—for example, jointly maintaining a high bid-ask spread [34]. Such collusion is harder to detect and prove than human collusion, because it is achieved not through explicit communication or agreement but emerges naturally from the same learning process. A distinction must be drawn here between two concepts that are often conflated: tacit collusion refers to independent AI agents spontaneously converging, through repeated play in the learning process, on a cooperative equilibrium (such as jointly maintaining a high spread), in which there is genuine strategic interaction and mutual adaptation; strategy convergence refers to AI agents based on similar model architectures and training data independently producing similar outputs, with no cooperative dynamic. The two may be equally harmful in their market impact but are entirely different in causal mechanism. Traditional collusion-detection methods, whether based on communication monitoring or behavioral-pattern analysis, may fail in both cases, because there is neither an explicit coordination signal that can be intercepted nor any way for the similar patterns in price behavior to distinguish whether they arise from interactive learning or from independent convergence.

4.5.5 Revisions to the analytical framework

The introduction of AI agents requires several revisions to the framework established in the first four sections of this chapter.

At the level of profit sources, the spread of AI may cause a structural change in the relative proportions of the two profit sources. The systematic losses of futile traders may shrink as they are eliminated at an accelerating pace, while the hedging premium may expand as AI makes more precise hedging strategies widespread, so that more participants can become utilitarian traders. This change in proportion has direct implications for diagnosing the health of the market. At the level of participant classification, the AI agent is not simply a seventh type of participant; it is more accurately positioned as a multi-role participant that can perform market making, informed trading, arbitrage, liquidation, and other functions simultaneously within the same time window. It accelerates the role-switching speed of the state-transition diagram described in Section 4.4.1 from the human minute-to-hour level to the millisecond level, turning state transitions from discrete jumps into something approaching continuous flow. At the level of the game network, when both sides of the game are AI, the nature of the game may evolve from a game of incomplete information into an algorithmic game of incomplete predictability, in which each side's strategy space itself changes dynamically, the equilibrium concepts of traditional game theory may no longer apply, and new demands are placed on analytical tools. The threefold reshaping identified in this chapter is likewise pushed to an extreme by the participation of AI: role fluidity accelerates to millisecond-level switching, the algorithmization of the game upgrades from the rule-based to the autonomously learning type with declining behavioral predictability, and

the threshold for exploiting transparency in the transparency-fragility paradox rises from computation speed to reasoning intelligence.

AI agents are not an "exception" to this chapter's analytical framework but its logical extreme. They push each of the three reshapings identified in Section 4.4 toward an extreme form beyond the reach of human participants. A systematic analysis of these extreme forms is developed in Chapter 31.

The autonomous decision-making capacity of AI agents also raises a fundamental question that existing legal frameworks have yet to resolve: when an autonomously operating AI agent causes harm in the perpetual futures market (for example, by triggering a flash crash through a high-frequency strategy, systematically extracting liquidity by exploiting a mechanism loophole, or forming tacit collusion with other AI agents), to whom should legal liability be attributed? To the developer who wrote the underlying model, to the operator who deployed the agent in a specific market, or to the agent itself as a kind of "electronic person" bearing liability? The EU Artificial Intelligence Act (EU AI Act) adopts a risk-tier-based classification framework for regulation but does not place high-frequency or algorithmic trading in its high-risk category (algorithmic trading is regulated mainly by RTS 6 of MiFID II and by DORA); both the AI Act and the trading rules above presuppose human controllability of AI behavior, whereas the decision paths of an autonomously learning agent often cannot be fully predicted or explained even by its developers. The current U.S. regulation of algorithmic trading (such as the SEC's Market Access Rule, Rule 15c3-5, and the CFTC's electronic-trading risk principles—its earlier automated-trading rule, Reg AT, having been withdrawn in 2020) is likewise built on the assumption that "the algorithm is a tool for executing human intent," tracing liability to the natural or legal person who deployed it. When an AI agent's behavior arises not from preset rules but emerges from reinforcement learning, and when its strategic choices exceed any prior human design or expectation, however, the line between "tool" and "independent actor" becomes fundamentally blurred. This gap in legal attribution is not merely a theoretical problem; it directly affects market participants' risk assessment and behavioral incentives: if liability is unclear, deployers may take excessive risk; if liability is too heavy, innovation may be suppressed. The attribution logic of existing legal systems, built on the core assumption of a human decision-maker, must fundamentally re-examine the relationship among "behavior," "intent," and "liability" when facing AI agents endowed with autonomous decision-making capacity.

4.6 Comparison with traditional finance

Systematically comparing the on-chain perpetual futures market with traditional financial markets is not meant to judge their relative merits but to identify precisely: what is a universal law of financial markets that holds in both; what is a structural innovation brought by digitally native technology; and what is entirely new—a phenomenon that does not exist in traditional finance.

4.6.1 The invariant underlying logic

Adverse selection is the perennial dilemma of the market maker. However technology changes, the market maker always faces a basic problem of information asymmetry: it cannot distinguish "non-toxic" from "toxic" orders in advance. This dilemma exists in traditional futures markets, exists in on-chain perpetual futures, and still exists in the AI era. Profit ultimately comes from "providing a service" or from "others' mistakes." This underlying logic does not change with the move on-chain. In traditional finance, market makers profit from the spread, arbitrageurs from mispricings, and informed traders from an informational advantage. In on-chain perpetual futures, these profit sources still exist; only the manner of their realization has changed. High leverage amplifies both returns and risk. Margin and liquidation mechanisms are necessary components of a leveraged market. Whether in traditional futures or on-chain perpetual futures, high leverage demands strict risk-management mechanisms. The zero-sum or negative-sum nature of "for someone to gain, someone must lose" is a mathematical fact of trading markets. In the perpetual futures market, the sum of all traders' P&L is below zero, with the shortfall going to pay transaction fees, gas fees, and MEV costs. This negative-sum nature is the same as in traditional futures markets. While acknowledging this invariant underlying logic, one must also carefully mark the limits of the classical models. The continuous-time assumption of the Glosten-Milgrom model, the premise in the Kyle model that the market maker bears a quoting obligation, and the continuous price-adjustment mechanism on which standard equilibrium analysis relies all face structural deviations in the on-chain environment: a discrete block-production mechanism replaces continuous-time trading; permissionless liquidity provision means there is no market-maker quoting obligation in the traditional sense; and block-level atomic execution (whereby the transactions in a single block either all succeed or all roll back) creates execution dynamics that do not exist in classical theory. These differences do not invalidate the underlying logic of adverse selection, profit sources, and leverage dynamics, but they do require researchers, when applying classical market microstructure models to on-chain perpetual futures, to mark explicitly where the models' assumptions depart from on-chain reality and to remain cautious about how far the models' conclusions extend.

4.6.2 The changed modes of implementation

Table 4-3 compares the structural transformations of on-chain perpetual futures and traditional finance along six core dimensions—mode of admission, source of informational advantage, nature of parasitic behavior, publicness of liquidation, speed of the game, and market-making risk—and distills the core direction of change in each dimension.

Dimension	Traditional finance	On-chain perpetual futures	Core shift
Admission	License + capital → fixed identity	Wallet address → fluid state	Specialization shifts from *licensed admission* to *knowledge/technology*
Informational advantage	From non-public information + interpersonal networks	From the capacity to process public data + understanding of mechanisms	Competition shifts from *who knows* to *who understands faster and more deeply*
Parasitic behavior	Insider trading (illegal → regulated by law)	Mechanism parasitism (compliant at the protocol layer but of undetermined legal status → requires both mechanism design and a legal framework)	Defense shifts from *legal punishment* to *mechanism repair + adaptation of the legal framework*
Liquidation	Internal risk-control process → not public	Permissionless external market → fully public	Liquidation shifts from an *internal function* to an *external profit opportunity*
Speed of the game	Millisecond-level → but with market closures	Millisecond-level + 24/7 + no closure as a shock absorber	Crisis propagation shifts from *hours to days* to *minutes*
Market-making risk	Has an expiration date → finite-horizon inventory risk	No expiration date → indefinite inventory risk	Market making shifts from a *time-bounded bet* to an *indefinite exposure*

Table 4-3. The core shifts between on-chain perpetual futures and traditional finance (Data source: author's compilation based on the literature)

From Table 4-3 a deep pattern can be distilled that runs through all dimensions: the "core shift" column of every row points in the same direction—from a market that depends on external institutional guarantees to one that depends on endogenous technical capacity. Admission shifts from license approval to a knowledge-and-technology barrier; the informational advantage shifts from exclusive channels to a competition in processing capacity; the defense against parasitic behavior shifts from legal punishment to mechanism repair; liquidation shifts from an opaque internal process to a public profit opportunity; the speed of the game loses market closure as a natural shock absorber; and market-making risk is rendered indefinite by the disappearance of the expiration date. The co-directional shift across these six dimensions is no coincidence but the concrete manifestation, at different levels, of a single underlying logic: the replacement of intermediary functions by smart contracts.

The detailed argumentation of the six dimensions above has been developed in Section 4.4 (role fluidity, algorithmization of the game, the transparency-fragility paradox) and Section 4.2 (market-maker inventory risk, mechanism parasitism, the liquidation mechanism); the table here merely provides a summary comparison. These transformations are not independent of one another but mutually reinforcing. The de-licensing of admission, for example, makes role fluidity possible, and role fluidity in turn intensifies the algorithmization of the game.

4.6.3 Entirely new species and phenomena

Three phenomena that have reached an unprecedented scale and systematicity in the on-chain perpetual futures market deserve special analysis. Structural problems of a similar kind are not entirely absent from traditional finance (the problem of market makers' false quotes, for example, and the early transparency deficiencies of the credit default swap market both exhibit some comparable features), but the permissionless admission, full transparency, and automatic code execution of the on-chain environment have produced a qualitative leap in the form, scale, and systemic impact of these phenomena.

Mechanism parasitism is a form of parasitism unique to the on-chain market (see Section 4.2.4): in an environment where rules are fully transparent and enforced by code, systematically discovering and exploiting the economic loopholes of a protocol's rules becomes a profitable, specialized activity; the rules of traditional finance, by contrast, are enforced by people and leave room for discretion, so the conditions for precisely arbitraging the literal meaning of a rule do not exist. The Hyperliquid whale event described in the introduction is a paradigmatic case satisfying the three conditions of Definition 4-1.

The transparency-fragility paradox (see Section 4.4.3) likewise has no counterpart in traditional finance: fully transparent position information, combined with high leverage and algorithmic execution, makes the targeted triggering of others' liquidations a strategy with positive expected value; in traditional finance positions are opaque, so such targeted liquidation is infeasible at the informational level in the first place.

The atomic switching of roles embodies the reshaping of participant behavior by smart-contract composability (the composability thesis of Chapter 1): within a single transaction, a participant can borrow funds via a flash loan, execute arbitrage on a price discrepancy, and provide liquidity in an AMM; this atomic, multi-role parallelism within a single transaction cannot be realized in traditional finance because of technical-architecture constraints.

4.7 Chapter summary

The participant ecosystem of the perpetual futures market can be understood through the profit-source framework of Harris (2003). The gains of profit-motivated traders ultimately come from only two channels: the hedging premium that utilitarian traders pay to transfer risk, and the systematic losses that futile traders incur through mistaken decisions. The former represents hedging demand and is the economic reason the market exists; the latter is the largest source of profit in today's crypto perpetual futures market. All profit-motivated participants (market makers, informed traders, arbitrageurs, mechanism parasites, and liquidators) contend for a share of these two profit sources. To understand this framework is to understand why the profit of certain participants must correspond to the loss of others, and why the market's sustainability depends on the relative size of the two profit sources.

The six distinctive features identified in Chapter 1 produce three structural changes at the level of participants. Role fluidity turns participants from fixed identities into fluid

states: a single address can be in different states at different moments, and the cost of transition between states approaches zero. The algorithmization of the game turns participants from human decision-makers into machine executors, with the influence of economic bots in the market steadily rising. The transparency-fragility paradox means that full transparency no longer implies fairness but instead creates the possibility of targeting the liquidation of other participants; the combination of high leverage and full transparency creates structurally favorable conditions for targeted-liquidation strategies.

"Mechanism parasitism," as an entirely new form of parasitism, challenges the traditional "information asymmetry" analytical paradigm. It exploits not information that others do not know but a deeper understanding of the economic implications of the protocol's rules than others have. The correct way to defend against mechanism parasitism is not to "punish the actor" but to "repair the mechanism loophole that makes the behavior profitable." This places higher demands on mechanism design.

The rise of AI agents pushes the threefold reshaping above to its extreme. It not only changes the competitive landscape of existing participants but also gives rise to entirely new forms of the game: the strategy-homogenization risk of AI vs. AI, the cognitive gulf of AI vs. humans, and the intelligent-scaling parasitism of AI vs. protocol mechanisms. The systematic impact of AI agents on market microstructure (including their effects on liquidity, price discovery, and market stability) is developed in depth in Chapter 31.

What kind of infrastructure do these participants rely on to trade? How does different infrastructure affect their behavior and risk? The next chapter focuses on the service-provider ecosystem that serves these participants.

References

[1] CoinDesk. (2025, March 12). Hyperliquid loses $4M after whale's over $200M ether trade unwinds. *CoinDesk*. https://www.coindesk.com/markets/2025/03/12/hyperliquid-loses-usd4m-after-whale-s-over-usd200m-ether-trade-unwinds

[2] Arkham Intelligence. (2025). Hyperliquid whale passes $4M loss to HLP vault. https://info.arkm.com/research/hyperliquid-whale-passes-4m-loss-to-hlp-vault

[3] Harris, L. (2003). *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press.

[4] FTI Consulting. (2025, December 24). *Crypto crash Oct 2025: Leverage meets liquidity*. https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity

[5] CoinGlass. (2025). *Crypto derivatives liquidation data, October 10–11, 2025*. CoinGlass. https://www.coinglass.com/

[6] Delfabbro, P., King, D. L., & Williams, J. (2021). Cryptocurrency trading, gambling and problem gambling. *Addictive Behaviors*, *122*, 107021. https://doi.org/10.1016/j.addbeh.2021.107021

[7] Schmeling, M., Schrimpf, A., & Todorov, K. (2023). *Crypto carry* (BIS Working Papers No. 1087). Bank for International Settlements. https://www.bis.org/publ/wor

k1087.htm

[8] Stoll, H. R. (1989). Inferring the components of the bid–ask spread: Theory and empirical tests. *The Journal of Finance*, *44*(1), 115–134. https://doi.org/10.1111/j.1540-6261.1989.tb02407.x

[9] Tiniç, M., Sensoy, A., Akyildirim, E., & Corbet, S. (2023). Adverse selection in cryptocurrency markets. *The Journal of Financial Research*, *46*(2), 497–546. https://doi.org/10.1111/jfir.12317

[10] Easley, D., Kiefer, N. M., O'Hara, M., & Paperman, J. B. (1996). Liquidity, information, and infrequently traded stocks. *The Journal of Finance*, *51*(4), 1405–1436. https://doi.org/10.1111/j.1540-6261.1996.tb04074.x

[11] Milionis, J., Moallemi, C. C., Roughgarden, T., & Zhang, A. L. (2022). Automated market making and loss-versus-rebalancing. *arXiv preprint arXiv:2208.06046*. https://arxiv.org/abs/2208.06046 https://doi.org/10.48550/arXiv.2208.06046

[12] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

[13] Werapun, W., Karode, T., Suaboot, J., Arporntthip, T., & Sangiamkul, E. (2025). Exploring risk and return profiles of funding rate arbitrage on CEX and DEX. *Blockchain: Research and Applications*, *6*(4), 100261. https://doi.org/10.1016/j.bcra.2025.100261

[14] Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35–55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

[15] Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, *107*(3), 797–817. https://doi.org/10.2307/2118364

[16] Skwarek, M. (2025). Why do investors behave irrationally in the cryptocurrency and emerging stock markets? *SAGE Open*, *15*(3), 21582440251361212. https://doi.org/10.1177/21582440251361212

[17] European Securities and Markets Authority. (2025, July). *Maximal extractable value: Implications for crypto-asset markets* (Report ESMA50-481369926-29744). https://www.esma.europa.eu/

[18] Stoll, H. R. (2000). Presidential address: Friction. *The Journal of Finance*, *55*(4), 1479–1514. https://doi.org/10.1111/0022-1082.00259

[19] Hendershott, T., & Menkveld, A. J. (2014). Price pressures. *Journal of Financial Economics*, *114*(3), 405–423. https://doi.org/10.1016/j.jfineco.2014.07.012

[20] Kappa Lab. (2025). HLP vs LLP: How perpetual DEX vaults performed in the $20B crypto crash (October 2025). https://kappalab.io/insights/hlp-vs-llp-how-perpetual-dex-vaults-performed-in-the-20b-crypto-crash-october-2025

[21] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100. https://doi.org/10.1016/0304-405X(85)90044-3

[22] Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The Journal of Finance*,

55(2), 773–806. https://doi.org/10.1111/0022-1082.00226

[23] Shiller, R. J. (2019). *Narrative economics: How stories go viral and drive major economic events*. Princeton University Press. https://doi.org/10.2307/j.ctvdf0jm5

[24] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[25] MakerDAO. (2020, April). *Black Thursday post mortem*. https://blog.makerdao.com/the-market-collapse-of-march-12-2020-how-it-impacted-makerdao/

[26] Lehar, A., & Parlour, C. A. (2021). *Decentralized exchange: The Uniswap automated market maker* (Working Paper). University of Calgary & UC Berkeley. https://doi.org/10.2139/ssrn.3905316

[27] Liu, J.-H. (2025). Hyperliquid's trading behavior: An interdisciplinary analysis of whale trading behavior. https://medium.com/@gwrx2005/hyperliquids-trading-behavior-f867c897d970

[28] Flashbots. (2025). *Flashbots: Mitigating the negative externalities posed by Maximal Extractable Value*. https://www.flashbots.net/

[29] Gramlich, V., Jelito, D., & Sedlmeir, J. (2024). Maximal extractable value: Current understanding, categorization, and open research questions. *Electronic Markets*, *34*. https://doi.org/10.1007/s12525-024-00727-x

[30] Bawa, N. (2026, January). How hedge funds, basis traders, and a $200M whale extracted alpha from crypto's $19 billion October. https://medium.com/@navnoorbawa

[31] Selkis, R., & Messari Research. (2024, December). *The crypto theses 2025*. Messari. https://messari.io/report/the-crypto-theses-2025

[32] Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*. https://arxiv.org/abs/2304.07619 https://doi.org/10.48550/arXiv.2304.07619

[33] Khandani, A. E., & Lo, A. W. (2011). What happened to the quants in August 2007? Evidence from factors and transactions data. *Journal of Financial Markets*, *14*(1), 1–46. https://doi.org/10.1016/j.finmar.2010.07.005

[34] Dou, W. W., Goldstein, I., & Ji, Y. (2025). *AI-powered trading, algorithmic collusion, and price efficiency* (NBER Working Paper No. 34054). National Bureau of Economic Research. https://www.nber.org/papers/w34054

Service Providers and the Ecosystem of Perpetual Futures Trading

On November 11, 2022, the cryptocurrency exchange FTX formally filed for bankruptcy protection, revealing a customer-funds shortfall of as much as $8.7 billion on its balance sheet [1]. John J. Ray III, who subsequently took over as chief executive officer and oversaw the bankruptcy proceedings, submitted a sworn declaration to the court stating that, in more than 40 years of legal and restructuring experience, he had never seen such a complete failure of corporate control or such a total absence of trustworthy financial information [2]. The public and the media have often reduced this episode to the personal fraud of the founder. Viewed through the lens of market microstructure and system architecture, however, the emergence of an $8.7 billion customer-funds shortfall is fundamentally an architectural problem rather than a purely moral one. FTX's collapse did not stem from a hack, a market crash, or sophisticated financial engineering. It stemmed from an institution that simultaneously played the roles of exchange, broker, clearinghouse, and custodian and that, absent effective checks and balances, systematically misappropriated customer funds. Under such a heavily function-bundled architecture, conflicts of interest shift from occasional misconduct to a structural norm, and any sufficiently large temptation can produce the same outcome.

This chapter takes neither the one-sided position that centralized exchanges (CEXs) are inherently inferior to decentralized exchanges (DEXs) nor the noncommittal stance that every architecture simply has its own merits and drawbacks. Instead, it uses a precise analytical framework to help readers understand how different service-provider architectures determine the types of risk that users face and how to choose rationally among those risks. To this end, the chapter develops five analytical tools. The risk substitution theorem formalizes the core insight that risk cannot be eliminated, only substituted. The trilemma characterizes the structural tension among performance, transparency, and decentralization. The bundling tax quantifies the hidden cost behind the efficiency advantage of centralized exchanges. The verifiability ladder establishes a five-level assessment standard running from trusting an institution to verifying code. And old problems in new forms reveals the isomorphic mapping between the new problems on-chain and the old problems of traditional finance.

Having established this framework, the chapter analyzes in turn the internal operat-

ing logic of three architectures—traditional financial derivatives markets, centralized cryptocurrency exchanges, and on-chain protocols—reveals the risk isomorphism that cuts across them, and examines the systematic effects of architectural evolution on four core dimensions of market microstructure. Chapter 4 analyzed who trades in perpetual futures markets; this chapter turns to who provides the infrastructure for that trading, and Chapter 6 will further examine how different architectures evolve dynamically through competition. After reading this chapter, readers should be able to use the risk substitution theorem and the trilemma to evaluate the architectural choices of any trading service provider, understand the technical criteria and applicable boundaries of each level of the verifiability ladder, and, when facing a concrete platform choice, make rational decisions grounded in architectural analysis rather than brand trust.

5.1 The risk substitution theorem and the service provider's trilemma

The architectural design of a crypto-asset trading platform does not take place in a vacuum; it seeks an optimum within a set of mutually constraining variables. Before analyzing specific service-provider models in depth, this section first establishes a unified set of analytical tools to serve as the coordinate system for the rest of the chapter. This section presents, in turn, the four functional dimensions, the risk substitution theorem, the service provider's trilemma, and the verifiability ladder—together forming a complete theoretical framework that runs from functional decomposition to risk assessment and from characterizing constraints to quantifying trust.

5.1.1 The functional dimensions of the trade lifecycle

Any perpetual futures service provider, whatever its technical form, must perform four core functions. Together these dimensions answer the fundamental question of how a trade moves safely from intent to finality, and their theoretical foundation traces back to Harris's classic treatment of the functions of market microstructure [3]. Harris decomposes trading services into core functions such as providing immediacy, managing counterparty risk, and facilitating price discovery, and he notes that trading venues also perform the economic functions of aggregating information and reducing search costs. The four-dimensional framework of this chapter is a concrete application of that idea to perpetual futures markets. The information-aggregation function in Harris's framework is folded mainly into the trade-execution dimension (price discovery realized through the matching engine), while the search-cost-reduction function is distributed across two dimensions: asset custody, which lowers trust costs, and risk management, which lowers the cost of risk assessment.

Asset custody answers the question of whether traders' funds are safe. It concerns how traders' assets are held and who controls them, and it directly determines the type and degree of counterparty risk. In traditional finance, asset custody is typically handled by heavily regulated custodian banks or central securities depositories, with a legal separation between client assets and the institution's own assets. In crypto markets, the mode of custody determines whether a trader retains control of the private keys.

Centralized exchanges use a pooled-custody model: traders deposit assets into wallets the platform controls, and the platform records user balances in an internal database. The risk this model exposed in the FTX episode is that, when the internal database can be altered by management, a user's *balance* may be nothing more than a number with no guarantee of solvency behind it. Decentralized protocols, by contrast, achieve non-custodial trading through smart contracts: the trader retains control of the assets throughout the trade, and funds move only under the strict conditions set by the smart contract.

Trade execution answers the question of whether an order is executed fairly. It covers the matching of buy and sell orders, price discovery, latency optimization, and execution fairness. The performance of the matching engine directly determines an exchange's throughput and the quality of market liquidity. Centralized exchanges achieve millisecond- or even microsecond-level matching through high-performance, custom in-memory matching engines. In 2024, the top 10 centralized exchanges recorded more than $58 trillion in total perpetual futures volume, with Binance's market share falling from about 43% at the start of the year to 34% by year-end [4]. Decentralized exchanges must instead contend with the latency of blockchain state updates, giving rise to automated market makers (AMMs), on-chain order books, and hybrid models that match off-chain and settle on-chain. Execution fairness is especially fraught on-chain, because the public transparency of a blockchain lets miners or validators observe and exploit a trader's unconfirmed orders, producing front-running and sandwich attacks. This is the technical re-emergence of an old problem: the long-running controversy in traditional finance over payment for order flow (PFOF) shares the same underlying structure: a broker sells client order flow to a market maker that uses its informational advantage to position itself ahead of execution. Maximal extractable value (MEV) is the on-chain, transparent-environment version of this old problem, made more systematic by the blockchain's openness.

Clearing and settlement answers the question of whether post-trade obligations are honored. For a highly leveraged derivative such as perpetual futures, clearing and settlement encompasses the real-time calculation of margin requirements, the mark-to-market processing of unrealized profit and loss (P&L), and the forced liquidation of positions once they become insolvent. In traditional derivatives markets, central counterparty (CCP) clearing absorbs default risk by building a default-waterfall structure that draws in sequence on the defaulter's margin, its default-fund contributions, the clearinghouse's own capital, and the default funds of non-defaulting members, thereby safeguarding the stability of the system [5]. In on-chain finance, smart contracts replace the traditional clearinghouse, executing real-time, deterministic clearing and settlement through preset code logic; this reconfigures the form of counterparty risk found in traditional clearing and eliminates settlement delay. On-chain clearing does not, however, fully eliminate counterparty risk; it transforms it into new forms. Shortfall losses are socialized to other users through the insurance fund or the auto-deleveraging (ADL) mechanism, the liquidator network may respond with delay under extreme conditions as gas fees spike, and the nondeterministic window of oracle price feeds introduces the risk of mark-price deviation.

Risk management answers the question of whether the system will collapse under ex-

treme conditions. It covers the continuous monitoring of margin, the triggering and execution of forced liquidation, the absorption of shortfall losses, and the defense against systemic risk. In a centralized exchange, the risk engine is usually tightly coupled with the matching engine, able to respond to market swings and trigger liquidations within milliseconds. In on-chain protocols, risk management relies on the timely response of the liquidator network and the accuracy of oracle price feeds; under extreme conditions, liquidation delays or improper liquidations caused by price manipulation can occur.

These four dimensions are not independent. The mode of custody directly affects the effectiveness of risk control: when assets are held by a smart contract, the risk engine can enforce margin requirements at the level of code, without relying on the self-restraint of the platform operator. Execution mode and settlement are likewise closely linked: the atomic-settlement capability of an on-chain order book lets a trade and its settlement complete simultaneously within a single blockchain transaction, eliminating the lag between execution and final settlement found in traditional markets. Understanding the interactions among the dimensions is the key to understanding the relative merits of different service-provider architectures. Traditional finance assigns the four functions to separate, independent institutions to achieve a separation of duties; centralized cryptocurrency exchanges concentrate them in a single entity to achieve functional bundling; and on-chain protocols reconfigure them with code and cryptography [6].

5.1.2 The risk substitution theorem

Observing how trading service providers evolved from traditional finance to centralized cryptocurrency exchanges and then to on-chain protocols, one general pattern gradually emerges. Each architectural transition successfully eliminated certain core risks of the preceding architecture, but each also inevitably introduced new types of risk. This pattern can be stated as the risk substitution theorem: for any transition of a service-provider architecture from A to B, the transition eliminates the risk set R_A specific to A while introducing the risk set R_B specific to B. The key proposition is that R_B is always nonempty—that is, no architecture drives all types of risk to zero simultaneously. Architecture choice is therefore not about eliminating risk but about choosing the type of risk. The rational criterion for the choice is whether, for a particular user, the risks in R_B are more identifiable, more quantifiable, and more manageable than those in R_A.

The risk substitution theorem describes a change in the type of risk, not a conservation of total risk. Risks under different architectures can differ by orders of magnitude in expected loss (probability $\times$ magnitude) and in tail characteristics. The operational risk of a centralized exchange, for instance, is relatively low in frequency, but once it materializes—as with FTX's \$8.7 billion shortfall—it can produce a loss of nearly 100% of principal; the loss from a single smart-contract exploit is usually smaller in scale but occurs more frequently. Tail-risk measures such as conditional value at risk (CVaR) can reveal that certain architectural transitions eliminate low-magnitude risks already effectively mitigated by regulation while introducing high-magnitude risks that lack external constraint. As a result, users with different risk preferences—tail-averse users versus mean-variance users—may reach entirely different optima from the same set of architectural choices.

Applying this theorem to the three architectural transformations of trading service providers reveals a clear symmetry, at each transition, between the risk set eliminated and the risk set introduced (Figure 5-1).

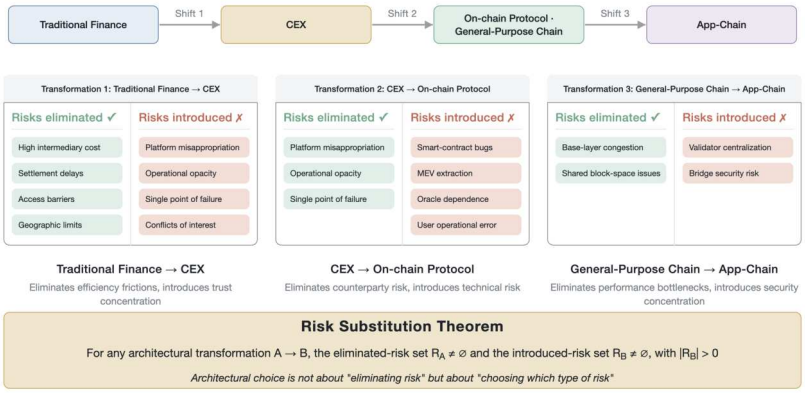


Figure 5-1. The risk substitution theorem — risk mapping across three architectural transitions (conceptual mapping, not empirical data; Data source: Dell'Erba [6], Yadav [7], Three Sigma [8])

Traditional financial markets, through the division of labor among multiple layers of intermediaries and a strict regulatory framework, effectively controlled the risks of fraud and misappropriation of funds. That architecture, however, introduced high intermediation costs, long settlement cycles (typically one to two business days), and significant barriers to access and geographic restrictions. The emergence of centralized cryptocurrency exchanges eliminated these efficiency frictions by adopting a vertically integrated architecture that concentrates all functions in a single entity, enabling around-the-clock global trading and instant transfers of funds. That transition, however, introduced serious risks of platform misappropriation, operational opacity, single points of failure, and conflicts of interest [6]. FTX's $8.7 billion customer-funds shortfall [1] and the Bybit hack of February 2025—in which a North Korean hacking group stole about $1.5 billion in Ethereum assets by manipulating the signing workflow of a multi-signature wallet [9]—are both extreme expressions of this centralization risk.

As markets evolved further toward on-chain protocols, smart contracts and blockchain consensus mechanisms replaced centralized servers. This architectural transition eliminated platform-misappropriation risk and operational opacity through cryptographic verification, achieving trustless trade execution and settlement. Risk did not disappear, however; it shifted to the technical layer. On-chain protocols introduced the risk of smart-contract vulnerabilities, the risk of exploitation through MEV (earlier also called miner extractable value), the risk of dependence on external oracle price feeds, and the risk of users irreversibly losing assets through operational mistakes. In 2024, smart-contract exploits in DeFi caused about $730 million in losses, down from $1.15 billion in 2023 but still a significant systemic threat [8]. Notably, this year-over-year decline may not reflect a simple improvement in security but rather the confluence of several factors: total value locked (TVL) on-chain fluctuated considerably over the same period, and the composition of attack vectors shifted from smart-contract logic

flaws to private-key theft and social engineering, the latter of which does not appear in contract-vulnerability statistics. According to a July 2025 report from the European Securities and Markets Authority (ESMA), cumulative MEV revenue on Ethereum since the Merge in September 2022 has amounted to roughly $1 billion to $1.3 billion, with arbitrage—rather than sandwich attacks—contributing the highest monthly revenue [10].

As DEXs on general-purpose public blockchains evolve further into purpose-built application-specific blockchains (app-chains), the logic of risk substitution appears once more. An app-chain, by building a specially optimized blockchain, eliminates the congestion of the underlying chain, competition over gas fees, and the MEV problems created by shared block space; at the same time, it introduces the risk of validator centralization, the security risk of cross-chain bridges, and the reflexivity problem of economic security. This reflexivity deserves particular attention: an app-chain's economic security is usually anchored to the market capitalization of its native token, and that market capitalization depends in turn on user adoption and confidence, forming a positive feedback loop of "token market cap → staking security → user confidence → adoption → token market cap." On the way up, this loop reinforces itself; on the way down, once it falls below some critical threshold, declining security triggers user flight, which depresses the token's market cap and further weakens security, forming a downward spiral that is hard to reverse (Section 5.4.3 analyzes this mechanism further). The security record of cross-chain bridges likewise warrants attention: as of 2025, cross-chain bridge attacks have caused cumulative losses of more than $2.5 billion, nearly 40% of all funds stolen across decentralized finance [11].

The practical import of the risk substitution theorem is this: when someone asserts that "decentralized exchanges are safer than centralized exchanges," the claim is incomplete. The full statement should be that decentralized exchanges eliminate the counterparty risk of centralized exchanges but introduce smart-contract risk and MEV risk; if a user is able to audit code and to use the appropriate protective tools, then a decentralized exchange may be "safer" for that user, and otherwise not necessarily so. The rational criterion for architecture choice is not the absolute quantity of risk but its identifiability, quantifiability, and manageability.

5.1.3 The service provider's trilemma

In the trade-offs of architectural design, a trading service provider faces a core structural dilemma: performance, transparency, and decentralization cannot all be optimized simultaneously under current underlying technical constraints, and a platform must make trade-offs among them. The root of this dilemma lies in the basic constraints of distributed systems: raising performance requires reducing the number of nodes participating in consensus and simplifying verification to lower communication overhead, whereas strengthening decentralization requires precisely the opposite—expanding the validator set and adding redundant verification. Achieving transparency, in turn, requires that all trade data be public to the entire network, which not only increases the verification burden but also exposes unconfirmed trade information to potential front-runners. The tension among the three is rooted in the inherent contradiction between

the cost of replicating information and the efficiency of reaching consensus: requiring more independent nodes to verify the same state necessarily consumes more time and bandwidth, while compressing the verification process necessarily reduces the number of independent verifiers or the completeness of verification.

The performance dimension covers trading latency, system throughput, and user interaction experience. For high-frequency traders and market makers, a difference of milliseconds in latency can determine the success or failure of a trading strategy. The transparency dimension requires that the entire trade lifecycle, the platform's asset reserves, and the operating logic of the risk engine be fully visible and independently verifiable to all market participants. The decentralization dimension measures the degree to which the system does not depend on a single entity or a small group of controllers, including the distribution of network nodes, the location of governance rights, and censorship resistance.

Different types of service provider occupy markedly different positions within this triangle (Figure 5-2), as analyzed below.

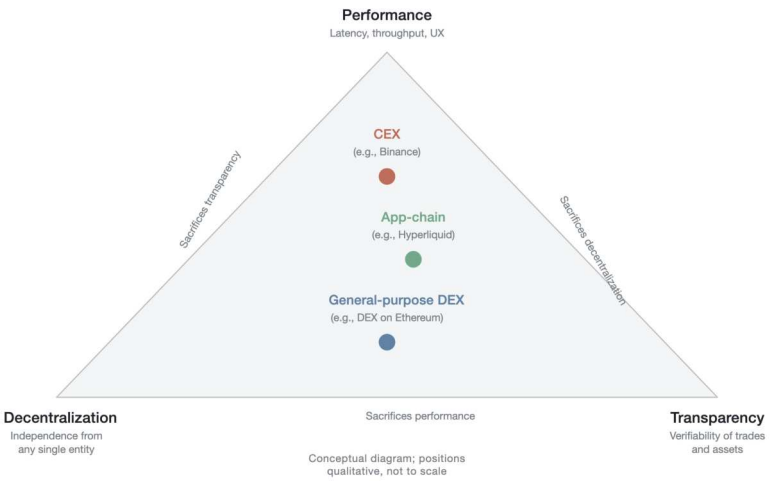


Figure 5-2. The service provider's trilemma — positioning of three paradigms (conceptual triangle; qualitative positions, not measured coordinates; placements reflect the author's qualitative judgment, not empirical data; Data source: CoinGecko [4], The Block [12])

Centralized exchanges prioritize performance and make concessions on transparency and decentralization. Binance handled the largest share of perpetual futures volume in the 2024 market, offering microsecond-level matching latency and deep liquidity, but the internal workings of its matching engine, risk-control logic, and asset management are entirely opaque, and users cannot independently verify whether the platform faithfully executes the rules it claims to follow [4]. Decentralized exchanges deployed on general-purpose public blockchains, by contrast, maximize transparency and decentralization: all trading logic runs openly on-chain, anyone can audit the smart-contract code and verify the execution of every trade, but they are constrained by the block time and computational capacity of the underlying blockchain, so their performance lags far

behind their centralized competitors. Ethereum mainnet's block time of roughly 12 seconds means that a perpetual futures exchange built on that chain cannot support complex trading strategies requiring high-frequency order-book updates.

A new generation of derivatives app-chains seeks a better balance between performance and transparency, but usually at some cost to decentralization. By building a sovereign blockchain optimized specifically for order-book trading, Hyperliquid adopts an on-chain state-machine matching model executed by a validator network, achieving sub-second latency while preserving on-chain settlement transparency. In its HyperBFT consensus mechanism, the leader node drives the execution of matching computation while other validators verify the state-transition result, rather than every node independently executing the full matching process. As of early 2025, however, the network was maintained by only about 16 validator nodes (later gradually expanded to about 24) [12]; compared with Ethereum's more than one million validators, its degree of decentralization remains limited. Ethereum's "one million validators," however, refers to the number of 32-ETH staking units, not the number of independent node operators; a large share of validators is concentrated under liquid staking protocols such as Lido, so the actual number of independent operators is far below that figure. The fourth version of dYdX (v4) likewise adopts a sovereign, Cosmos-based blockchain architecture, but its order book uses a design different from Hyperliquid's: the order book is maintained in validator memory (an off-chain order book), and only fill results pass through consensus onto the chain, while order submission and cancellation do not go through blockchain consensus. This design brings dYdX v4's performance close to that of a centralized exchange, but the size and diversity of its validator set face similar constraints.

The analytical value of the trilemma lies in its explanation of why no perfect service provider exists. Every design is a chosen point within the triangle, and the location of that point is determined by the priorities of its target users. The triangle also predicts the direction of future technical progress: zero-knowledge proofs, modular blockchains, and data-availability layers are all, in essence, ways to expand the feasible area of the triangle so that the three goals can be achieved simultaneously at a higher level—yet the triangle itself, as a structural constraint, will not disappear. This trilemma is related to, but distinct from, the classic blockchain trilemma of security, decentralization, and scalability [13]: the latter describes the constraints of the underlying consensus protocol, whereas this chapter's triangle describes constraints at the level of the trading service provider. The constraints of the underlying triangle propagate upward, shaping the boundary of the feasible region for the service-provider triangle.

5.1.4 The verifiability ladder

To assess more precisely how different trading service providers differ in their trust assumptions, it is useful to build a verifiability-ladder model. The model classifies a trading platform's architecture, according to the degree to which its functions are publicly verifiable on-chain, into five progressive levels from L0 to L4. An increase in verifiability is not a binary switch but a graduated continuum. Each step upward means that the user's trust cost falls by one notch but is replaced by a greater cost of

technical understanding.

The five levels of the ladder run progressively from the fully unverifiable L0 (such as FTX before its collapse) to the fully verifiable L4 (which no platform has yet reached), and the technical criteria at each level span the mode of asset custody, the degree of disclosure of matching logic, the auditability of risk-control parameters, and the level of decentralization of the governance mechanism (Figure 5-3).

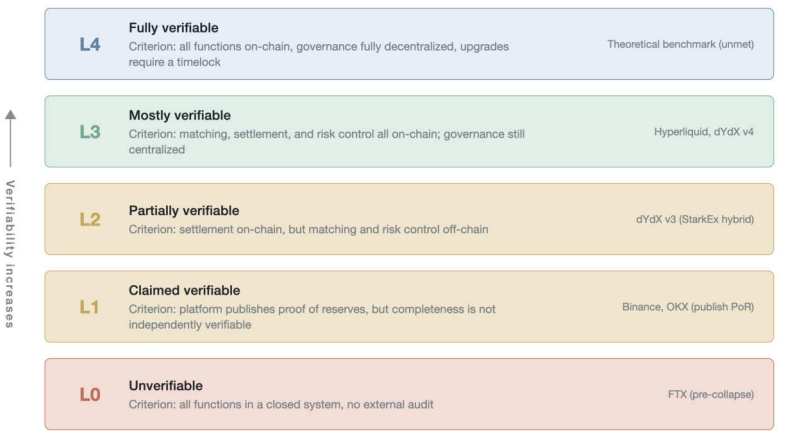


Figure 5-3. The verifiability ladder — the five-level standard from L0 to L4 (conceptual grading framework; the author classifies platforms by technical criteria, not empirical data; platform examples as of 2025, including the proof of reserves (PoR) published by Binance and OKX; Data source: Ray [2], Vidal-Tomás [14], The Block [12])

L0 denotes the unverifiable architecture. At this level, all of the platform's functions run inside a closed, centralized system, and no substantive external audit or verification is possible. FTX before its collapse was a typical L0 platform. In his sworn declaration, John J. Ray III disclosed that FTX did not even maintain a complete balance sheet and used instant-messaging software rather than a formal accounting system to handle financial records [2]. Under an L0 architecture, the user's only option is to trust, and that trust lacks any verifiable basis.

L1 denotes the attestation-based architecture. After FTX's collapse, many centralized exchanges began publishing proof of reserves (PoR) at regular intervals, using Merkle-tree technology to prove their asset positions to users [14]. This form of verification, however, has inherent limitations: proof of reserves is essentially a static point-in-time snapshot, and it often covers only the asset side while being unable to fully verify the liability side (its snapshot nature and liability-side limitations are detailed in Section 5.3.5). Although transparency improves over L0, the system's core operations still depend on the commitments and self-restraint of a centralized entity.

L2 denotes the partially verifiable architecture. At this level, the platform begins to split the key stages of the trade lifecycle, entrusting fund custody and final settlement to on-chain smart contracts while order routing and matching are still handled by high-performance centralized servers. The third version of dYdX (v3) is the archetype of this

architecture; it uses StarkEx validity-proof technology to achieve on-chain settlement on Ethereum while running a high-performance matching engine off-chain. Validity proofs provide a mathematical guarantee of certainty: the operator cannot submit an incorrect state transition to Ethereum mainnet, because any settlement that does not conform to the proof rules is rejected by the mainnet contract. This brings the L2 architecture close to L3's level of protection on the dimension of fund safety. The disadvantage of this hybrid architecture, however, is that the matching process itself is not verifiable and carries censorship risk, so traders must still trust the fairness and censorship resistance of the centralized matching engine. Users can audit the code of the on-chain settlement contract but cannot verify whether order priority is manipulated during off-chain matching.

L3 denotes the largely verifiable architecture. The platform migrates matching, settlement, and the risk engine entirely on-chain, achieving public verifiability across the whole trade process. Both Hyperliquid and dYdX v4 adopt this architecture based on a sovereign app-chain [12]. dYdX v4's off-chain order-book design (see Section 5.1.3) means that order submission, cancellation, and matching themselves do not pass through blockchain consensus, but every fill result can be traced and verified through the blockchain's public record; within L3, it therefore makes a greater concession on matching transparency—relative to Hyperliquid's full-state consensus model—in exchange for performance. Although the core trading logic is already fully transparent, the core development team may still retain some degree of centralized control over node-admission mechanisms, emergency upgrade authority, and the adjustment of governance parameters. Such control has real security value in a protocol's early stages, because when a serious vulnerability is discovered, a rapid emergency response may protect user assets better than a fully decentralized governance process; at the same time, it constitutes a trust assumption that must be explicitly flagged.

L4 denotes the fully verifiable architecture. This is currently the highest standard in theory: it requires the platform not only to run entirely on-chain at the trading and settlement layers, but also to make its governance mechanism and protocol upgrades fully decentralized and subject to strict time locks and community consensus. At this level, any modification of a protocol parameter requires an on-chain vote, and the outcome must pass through a public waiting period before execution, giving all stakeholders ample time to review the change and, if necessary, exit the system. At present, no platform in perpetual futures trading fully meets this standard; it represents the direction of the industry's evolution rather than a current reality.

The core insight of the verifiability ladder is this: for ordinary users who lack the technical ability to audit code, the advantage of "verifying code" over "trusting an institution" may be difficult to realize in practice; but for institutional investors with the requisite ability, verifiability offers a depth of due diligence that is impossible under the centralized-exchange model. Yet verifiability is a necessary but not sufficient condition for trust substitution: even if a system provides complete verification capability, information overload and interface complexity may leave most users with no willingness to verify, and the label "code audited" may instead create a false sense of security, so that "willingness to verify" systematically lags "ability to verify." As cryptographic technology and blockchain performance continue to improve, trading service providers

keep climbing this ladder toward higher levels, converting ever more trust assumptions into deterministic logic provable by mathematics. This trend not only reshapes the competitive landscape of trading service providers but also fundamentally changes how trust is produced in financial infrastructure.

5.2 The functional-separation architecture of traditional finance

The architecture of traditional financial derivatives markets rests on an institutional foundation of functional separation. Brokers, exchanges, clearinghouses, and custodian banks each perform their own role, achieving risk isolation through physical and legal boundaries. This multi-layered intermediary architecture did not form overnight; it is the institutional accumulation of the lessons of two centuries of financial crises. Starting from the complete flow of a single futures trade, this section analyzes the institutional rationale and the efficiency cost of functional separation, together with the central counterparty's core role in managing systemic risk.

5.2.1 The multi-intermediary flow of a futures trade

The flow of a standardized futures trade through the traditional financial system is not completed within a single institution as a closed loop; it relies on a service network made up of several independent entities. Consider the example of a buy order in S&P 500 futures. The investor places a buy instruction with the futures commission merchant (FCM) that holds the account; after verifying that the client's margin is sufficient, the FCM routes the order to the electronic matching engine of the Chicago Mercantile Exchange (CME). The CME recorded an average daily volume of about 26.5 million contracts in 2024 [15], and its matching engine completes price-time-priority order pairing at the microsecond level. The instant the trade is struck, the CME's clearing division steps in as central counterparty, using novation to split the original buy-sell contract into two independent contracts: one between the buyer and the central counterparty, and one between the central counterparty and the seller. Thereafter the central counterparty computes daily profit and loss and collects margin, while the final transfer of funds is executed by an independent settlement bank, ensuring that client assets and the trading institution's own assets are strictly separated at the legal and physical levels.

Within this flow, each participating institution performs a distinct trust function. The FCM, acting as an agent, owes a fiduciary duty to the client and must place the client's interests above its own; this includes the duty of best execution—routing orders to the venue offering the best execution—and the risk-management responsibility of managing the client's margin account and issuing margin calls when necessary [3]. The FCM also performs a credit-intermediation function, extending leverage to clients and continuously assessing credit risk; when a client defaults, the loss is absorbed by the FCM rather than by the market as a whole, and this buffer prevents an individual default from transmitting directly to the systemic level. The exchange's core value lies in its neutrality: it does not trade, holds no positions, and merely provides a fair

and transparent price-discovery venue; its revenue comes from fees rather than trading profits, and its rules and matching algorithms are publicly subject to regulatory review [3]. The central counterparty focuses on risk management, reducing total systemic exposure through multilateral netting and absorbing the default losses of individual members through the default-waterfall mechanism. The custodian bank stands apart from all trading activity; its sole responsibility is to safeguard client assets and to execute transfers of funds and securities on clearing instructions.

In this architecture, the broker sits at the client-interface layer, bearing fiduciary duty and performing credit intermediation; the exchange provides a neutral price-discovery venue at the matching layer; the central counterparty manages systemic default risk at the clearing layer; and the custodian bank ensures the legal segregation of client assets at the asset-safekeeping layer. Clear functional boundaries exist among the four layers, and the institutional barriers of functional separation among layers ensure that the failure of any single layer does not run straight through to the others (Figure 5-4).

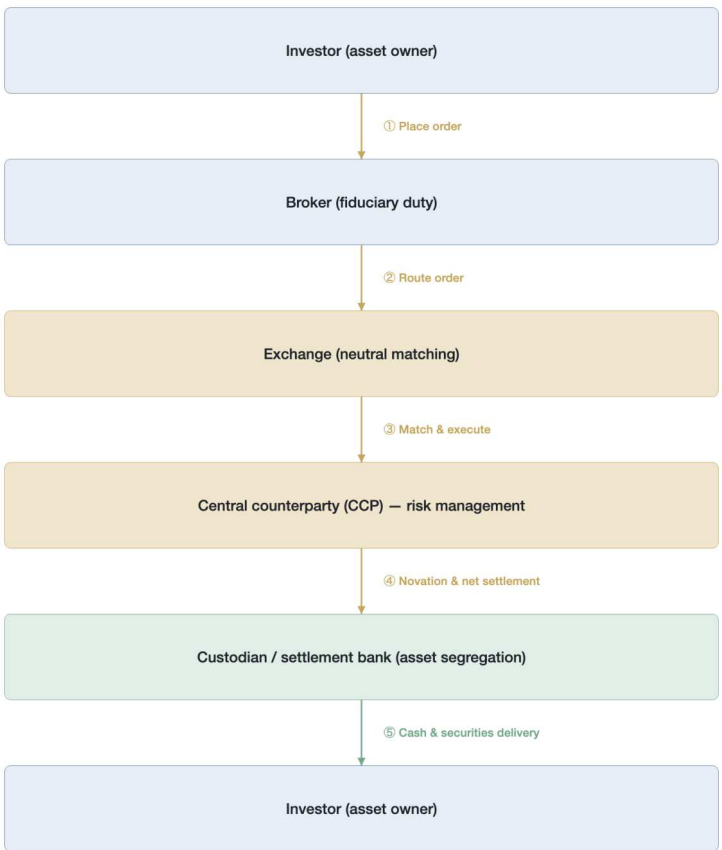


Figure 5-4. The functional-separation architecture of traditional financial service providers, by layer (conceptual architecture based on Harris's 2003 functional-separation framework, not empirical data; Data source: Harris's market microstructure framework [3])

This flow of business through at least four independent institutions in sequence generates time delays and friction costs at every handoff. The FCM collects a commission, the exchange collects a matching fee, the central counterparty collects a clearing fee, and the custodian bank collects a custody fee. On the surface, each layer of friction lowers the efficiency of capital turnover; at a deeper level, however, these are the costs necessarily paid to achieve risk isolation. The core purpose of this architecture is not to pursue the highest trading speed but to prevent the failure of a single node from triggering a global systemic collapse, through physical and legal boundaries. When an exchange converts from a membership-based mutual to a listed company, the tension between commercial interest and market-oversight duty grows increasingly acute; but the institutional framework of functional separation keeps that tension within a controllable range, rather than concentrating every role in one entity and turning conflicts of interest into a structural norm, as with the centralized cryptocurrency exchanges analyzed later.

5.2.2 The institutional rationale for functional separation

The current functional-separation architecture of traditional financial markets is the institutional accumulation of the lessons of two centuries of financial crises. In the early, unregulated markets, institutions that combined client custody, proprietary trading, and brokerage under one roof were common, but this model was repeatedly shown to be prone, under extreme market conditions, to the misappropriation of client assets or to an institution's own speculative losses spilling over onto innocent clients. The Glass-Steagall Act, passed in the United States in 1933, was precisely the legislative response to the systemic risk of the 1929 Great Depression, in which commercial banks used client deposits for speculative trading; the act legally mandated the separation of commercial banking from investment banking [16]. Notably, the Gramm-Leach-Bliley Act of 1999 partially repealed Glass-Steagall's separation requirement, allowing commercial and investment banking to reintegrate. This loosening is widely regarded as one of the institutional triggers of the 2008 global financial crisis, and after the crisis, calls to restore functional separation rose again. This repeated history confirms a core proposition: once the institutional barrier of functional separation is removed, systemic risk reaccumulates along the paths of conflict of interest. The legislative principle of Glass-Steagall was continually extended over the following decades to every part of the securities and derivatives markets, ultimately producing the current institutional configuration in which brokers, exchanges, clearinghouses, and custodian banks each perform their own role.

The institutional rationale for functional separation is expressed mainly along three mutually reinforcing dimensions: risk isolation, checks and balances, and auditability. Risk isolation ensures that the failure of a single institution does not directly destroy the whole system, with losses confined within that institution's functional boundary. The 1995 collapse of Barings Bank offers a classic case: the bank's head of derivatives trading and settlement in Singapore controlled both front-office trading and back-office settlement—two functions that should have been separated—and, absent independent oversight, accumulated hidden losses of up to £827 million, ultimately bankrupting a bank with 233 years of history [17]. The fundamental lesson of Barings is not the

moral failing of a single trader but that the combination of functions eliminated internal checks and balances, allowing risk to swell to an irrecoverable scale before it was discovered. Checks and balances are achieved through mutual monitoring among different institutions: the FCM monitors the exchange's execution quality, the exchange is subject to review by regulators, and the central counterparty stands apart from commercial interests and focuses on risk management. As for auditability, each independently operating institution constitutes a clear audit node, and regulators can separately examine the custodian bank's asset reports and the central counterparty's risk models without confronting a single, highly complex, internally interconnected black box.

Fiduciary duty is the central legal cornerstone of the functional-separation architecture. In his classic market microstructure theory, Harris (2003) notes that brokers owe clients a strict fiduciary duty and must place the client's interests above their own [3]. This legal constraint requires the broker, as an independent credit intermediary, not to bet against its clients, thereby suppressing conflicts of interest at the root. The irreplaceability of fiduciary duty lies in the legal protection it provides to the weaker party in an agency relationship: when a broker breaches its fiduciary duty, the client can obtain redress through the courts. This legal constraint disappears entirely in the centralized-cryptocurrency-exchange model, because the exchange simultaneously plays the dual role of agent and trading counterparty, so that conflicts of interest are no longer occasional misconduct but a structural feature embedded in the business model.

Robustness rests on the premise that each participating institution operates independently within its functional boundary, with no cross-boundary interest linkages. The 2011 bankruptcy of MF Global offers a classic cautionary case: facing a liquidity crisis, this FCM improperly misappropriated nearly $1 billion in segregated customer funds to cover its own trading losses on European sovereign bonds [18]. An investigation by the U.S. Commodity Futures Trading Commission (CFTC) found that MF Global had systematically violated customer-fund segregation rules in the final week before its bankruptcy. This episode not only prompted regulators to comprehensively strengthen the customer-fund segregation regime for FCMs but also demonstrated that even within a functional-separation framework, once the bottom line of physical segregation is breached, systemic risk spreads quickly. Yet the MF Global episode also validated the resilience of the functional-separation architecture: because customer funds enjoy priority in bankruptcy under the law, the vast majority of customers ultimately recovered all their funds through the bankruptcy proceedings [18]—a marked contrast to the near-total inability to recover client assets in the centralized-cryptocurrency-exchange bankruptcies discussed later.

5.2.3 The CCP default waterfall mechanism

The central counterparty is the core hub for managing counterparty credit risk in traditional financial markets, and its operation rests on novation (see Chapter 1 for the discussion of the central counterparty and the default waterfall): once a trade is matched, the central counterparty steps in immediately and replaces the original buy-sell contract with two independent contracts to which it is the counterparty, so that the two sides need not assess each other's credit and need only trust the central counterparty's

solvency. After the 2008 global financial crisis, the G20 leaders reached a consensus in 2009 requiring all standardized over-the-counter (OTC) derivatives to enter central clearing, establishing the central counterparty's systemically important status in the modern financial system [19]. In 2012, the Committee on Payment and Settlement Systems (CPSS) of the Bank for International Settlements (renamed the Committee on Payments and Market Infrastructures, or CPMI, in 2014) and the International Organization of Securities Commissions (IOSCO) jointly issued the Principles for Financial Market Infrastructures (PFMI), setting 24 international standards for the governance, risk management, and financial resilience of central counterparties worldwide [20].

The central counterparty sharply reduces total exposure within the system through multilateral netting, but it also concentrates counterparty credit risk on itself. To manage this highly concentrated risk, the central counterparty builds a multi-layered loss-absorption mechanism that follows the primary *defaulter-pays* principle and introduces mutual support only in extreme cases—the default waterfall. This is a five-layer structure that begins with the defaulting member's initial margin and proceeds layer by layer through default-fund contributions and the central counterparty's own capital up to recovery and resolution measures; see Chapter 1 for the discussion of the central counterparty and the default waterfall [21]. The third layer, the contribution of the central counterparty's own capital, is commonly referred to as *skin in the game*; it is intended to guard against the central counterparty's moral hazard and to ensure that it has sufficient incentive to maintain strict margin standards and prudent risk models. Every layer of the default waterfall is backed by explicit legal force and regulatory review; its parameters are public and transparent, adjustments require formal procedures, and users' rights enjoy legal protection.

The interplay of novation and the default waterfall constitutes the core process of central counterparty risk management: after a trade is matched on the exchange, the central counterparty steps in through novation to replace the original bilateral contract with two independent contracts to which it is the counterparty; thereafter, if one side defaults, the five layers of the waterfall are triggered in the predefined sequence until the loss is fully absorbed or a recovery mechanism is invoked (Figure 5-5).

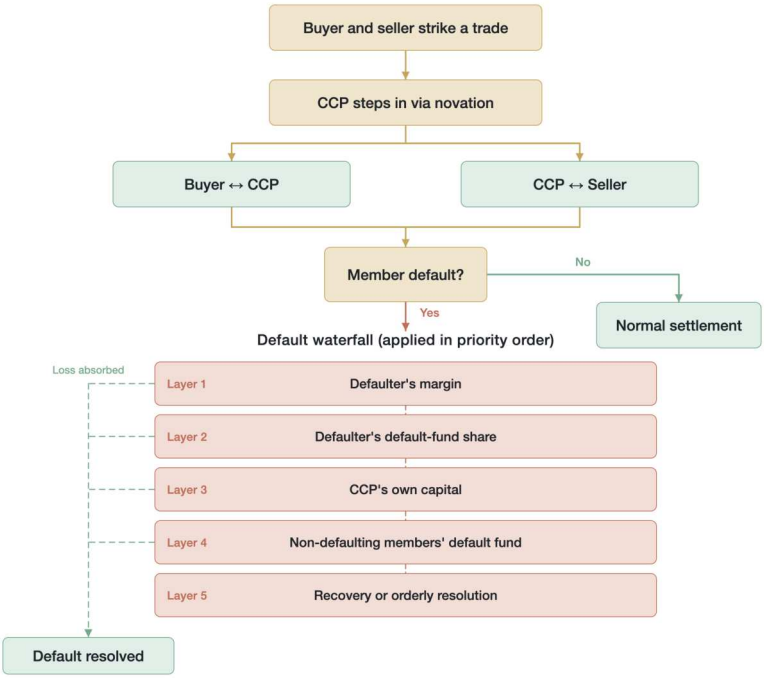


Figure 5-5. CCP novation and the default waterfall (mechanism flow based on the CPMI-IOSCO PFMI [20], not empirical data; Data source: CPMI-IOSCO Principles for Financial Market Infrastructures [20])

The effectiveness of the default waterfall depends heavily on the characteristics of the portfolio and on market-liquidity conditions. When Lehman Brothers failed in 2008, its interest rate swap portfolio cleared in London comprised 66,390 trades with a notional principal of as much as $9 trillion; but because the portfolio was relatively balanced and the interest rate swap market was highly liquid, the central counterparty consumed only about one-third of the margin and, within a few weeks, completed the hedging, auction, and transfer of the positions without inflicting any loss on other clearing members [21]. That success, however, depended on a special set of conditions: the interest rate swap market itself had ample liquidity; Lehman's portfolio was relatively balanced, which lowered the difficulty of hedging the net exposure; and the world's central banks injected large-scale liquidity into the market synchronously during the same period. These conditions do not hold in crypto perpetual futures markets: on-chain insurance funds are limited in size, liquidity can evaporate in an instant under extreme conditions, and no lender-of-last-resort role exists. The central counterparty's performance in the Lehman default was called an "unlikely hero" by Bell and Holden (2018) [21], because even under the extreme condition of a near-total freeze in bilateral markets, central clearing kept functioning normally.

The default at Nasdaq Clearing in Sweden in 2018, however, produced an entirely different outcome. On September 10, 2018, the Norwegian trader Einar Aas failed to meet a margin call from the commodities clearing division of Nasdaq Clearing and was

promptly declared in default. Aas had bet that Nordic and German power prices would converge and held large power-futures positions on Nasdaq Oslo. Changes in weather forecasts and adjustments to German carbon-emissions policy drove the two regions' power prices sharply apart, causing the value of his positions to shrink dramatically. In the subsequent auction of the positions, the clearinghouse suffered losses exceeding the value of Aas's posted collateral by €114 million. Nasdaq Clearing's default waterfall was triggered in sequence: it exhausted all of Aas's margin, then drew on €7 million of the clearinghouse's own capital, and ultimately consumed most of the non-defaulting members' €166 million default fund [21]. Post-event analysis revealed that Nasdaq Clearing had granted Aas a correlation offset of up to 50% in its margin calculation and had not charged a concentration add-on margin on his oversized positions in the Nordic power market, even though that market's trading volume had shrunk substantially over the preceding decade [21]. The core lesson of the Aas episode is not simply that "the offset ratio was too high" but that concentration risk was not adequately priced and that correlation assumptions are fragile in tail events. This forms a direct analogy to the concentration risk of whale accounts in a single trading pair in crypto perpetual futures markets and to the fragility of cross-asset correlation assumptions under a unified-margin model.

These two default events, a decade apart, reveal the inherent tension of the central counterparty model: it can effectively manage dispersed, routine defaults, but when confronted with a highly concentrated, illiquid, and highly correlated portfolio, the very feature of risk concentration can turn it into a transmission node for systemic risk rather than a risk-isolation mechanism. The central counterparty has been given a larger systemic role in the post-crisis era, but this also means that its own failure would produce more severe systemic consequences—a paradox that remains one of the core issues in global financial regulation to this day.

5.2.4 The efficiency cost of functional separation

While the functional-separation architecture provides institutional safety, it inevitably sacrifices the efficiency of capital turnover. Because trade execution, clearing, and custody are handled by different institutions, synchronizing the system's state takes time, which manifests directly as the existence of a settlement cycle. The longer the settlement cycle, the larger the time window during which the two sides are exposed to counterparty-default risk and market-price risk before delivery is complete, and the greater the systemic margin required.

The evolution of the settlement cycle in U.S. securities markets reflects traditional finance's continual search for a new balance between safety and efficiency. Before 1995, the standard settlement cycle for U.S. securities was T+5—delivery completed on the fifth business day after the trade date. In 1995, the U.S. Securities and Exchange Commission (SEC) shortened the cycle to T+3 through Rule 15c6-1. In 2017, the rule was amended again, compressing the cycle to T+2. On May 28, 2024, the U.S. market formally transitioned to a T+1 settlement cycle [22]. Each compression of the cycle was accompanied by large-scale infrastructure upgrades and industry coordination—from the digitization of paper certificates to the automation of trade-confirmation processes

to the standardization of cross-time-zone settlement instructions.

An after-action report jointly issued by the Investment Company Institute, the Securities Industry and Financial Markets Association, and the Depository Trust & Clearing Corporation shows that after T+1 took effect, the National Securities Clearing Corporation's clearing fund fell from an average of $12.8 billion under T+2 to $9.8 billion, a reduction of 23%, meaning that about $3 billion of capital was released from margin [22]. On settlement quality, the continuous net settlement (CNS) fail rate on the first day of T+1 was 1.90%, below the 2.01% average for May under the earlier T+2 regime; by July 2024, the average fail rate had stabilized at 2.12%, essentially unchanged from the T+2 period [22]. Trade-confirmation efficiency also improved markedly: nearly 95% of trades were confirmed before 9 p.m. on trade date, up from just 73% in January 2024 [22]. These data indicate that, under a multi-layered intermediary architecture, further compressing the time window did not significantly degrade settlement quality—but the scale of margin released also reveals just how large the loss of capital efficiency had been from funds previously locked up by settlement delay.

Although the T+1 reform achieved initial success, the advance of traditional financial infrastructure toward T+0 faces significant technical and institutional obstacles. The inherent contradiction of a multi-layered architecture is that each additional layer of independent intermediary, while enhancing risk isolation, multiplies the coordination complexity of information reconciliation and fund transfer. Under T+0, trade and settlement must occur almost simultaneously, which requires instantaneous data consistency and fund availability among brokers, exchanges, clearinghouses, and custodian banks. Industry reports make clear that pushing T+0 directly within the existing functional-separation framework is not a simple next step but requires a comprehensive, independent assessment of the entire market structure, because the disappearance of the time buffer could trigger large-scale liquidity shortfalls and settlement failures [22]. Traditional financial-infrastructure institutions have begun actively exploring blockchain technology for a breakthrough: the Depository Trust & Clearing Corporation's Project Ion completed a proof of concept in 2021 for T+0 settlement supported by distributed ledger technology (DLT) [23] and has since continued to advance pilots for the tokenization of digital assets, seeking to use atomic settlement to eliminate the risk created by the settlement cycle itself while preserving the advantages of functional separation.

Every increment of safety comes at the cost of some sacrifice in efficiency. When the institutional barrier of functional separation is viewed as a structural constraint impeding gains in capital efficiency, concentrating all authority in a single entity to achieve instantaneous settlement on an internal ledger becomes a commercially attractive alternative. This pursuit of efficiency is the core logic behind the rise of centralized cryptocurrency exchanges, and the structural consequences they bring form the starting point for the next section's analysis.

5.3 The functional bundling of centralized exchanges

The business model of centralized exchanges is rooted in bundling within a single entity the functions that traditional finance keeps separate. Section 5.2 analyzed how tradi-

tional finance manages conflicts of interest and systemic risk through functional separation; centralized exchanges represent the diametrically opposite architectural choice. This bundling brings substantial efficiency gains that have made centralized exchanges overwhelmingly dominant in the crypto-derivatives market, but it also creates an ineliminable hidden cost—the concept of the "bundling tax" introduced in this chapter. The bundling tax is the service-provider-level mirror of the "trust tax" from Chapter 1 (elaborated in Section 5.3.2) and is likewise the key to evaluating the relative merits of different service-provider architectures.

5.3.1 The CEX's bundling of five roles

The architecture of traditional financial markets rests on a strict separation of functions. As Section 5.2 described, the broker handles the client interface and owes the client a statutory fiduciary duty, the exchange provides a neutral matching venue and does not participate in trading, the market maker independently supplies liquidity and bears inventory risk, the clearinghouse manages systemic default risk as central counterparty, and the custodian stands apart from trading activity to ensure the safe segregation of client assets [3]. The separation of these five roles is not a waste of efficiency but the institutional accumulation of two centuries of financial-crisis lessons, and its core design principle is to disperse power so that the failure of any single institution is confined to a limited scope.

Centralized exchanges fundamentally alter this architecture by integrating the above five roles into a single, highly integrated technical and commercial entity. Under this model, a centralized exchange not only provides the trading interface and the matching engine but also directly controls the custody of client assets, performs clearing and settlement internally, and, in many cases, acts as the largest market maker on its own platform through a proprietary trading desk or affiliated parties. Dell'Erba (2024) [6] defines this architecture as a "full-service business model" and argues that it creates "unprecedented conflicts of interest." Such a high concentration of functions means that effective information barriers are difficult to establish in practice. An exchange operator can, in theory and in practice, exploit its advantage of seeing client order flow first to front-run, lend client funds from custody wallets to its proprietary market-making desk, adjust liquidation parameters to protect affiliated parties' positions, or handle liquidation orders in ways unfavorable to clients during market swings. These are not purely theoretical conjectures but facts repeatedly demonstrated in crypto markets.

The scope of this bundling can be grasped by comparing it, role by role, with the five independent roles of traditional finance. The broker's fiduciary duty is compressed into a user agreement for the account system; the exchange's neutral matching is folded into an engine the platform itself controls; the market maker's independence gives way to a proprietary trading desk; the clearinghouse's independent risk control becomes an internal module; and the custodian's asset segregation becomes internal bookkeeping within the platform's wallet system. Table 5-1 lays out these five pairings and the original institutional intent behind each.

Independent role in traditional finance	Counterpart in the CEX	Original intent of functional separation
Broker	Account system and user interface	Bears a fiduciary duty and acts in the client's interest
Exchange	Internal matching engine	Remains neutral and does not participate in trading
Market maker	Proprietary market making or affiliated parties	Independent of the exchange, preventing information abuse
Clearinghouse	Internal clearing and risk engine	Maintains independent risk management, free from commercial interests
Custodian	Hot/cold wallet system controlled by the exchange	Independent of trading activity, ensuring the safe segregation of assets

Table 5-1. Mapping between traditional finance's functional separation and the centralized exchange's role bundling (Data source: theoretical frameworks based on Harris 2003 and Dell'Erba 2024 [3][6])

This mixing of roles is not an accidental technical lapse but a structural choice through which centralized exchanges gain competitive advantage. Binance, for example, in September 2022 automatically converted competing stablecoins held by customers—USDC, USDP, and TUSD—into its own stablecoin, BUSD, a move that pushed BUSD's market capitalization to $20.5 billion [6]. This case clearly shows how a platform can use its dual identity as both custodian and exchange to reinforce its own commercial interests—behavior that would be impossible under traditional finance's functional-separation architecture. An industry that flies the banner of "decentralization" has in fact produced intermediaries more centralized than those of traditional finance, forming what Yadav (2023) [7] describes as the "centralization paradox." As of 2023, the vast majority of cryptocurrency users—an estimated 90%—chose custodial platforms over self-custody solutions [7]. This figure should not be read simply as users' rational preference for efficiency. A behavioral-finance perspective reveals more complex drivers: status quo bias inclines users who registered early on centralized platforms to stay put, complexity aversion leads nontechnical users to instinctively avoid the cognitive burden of private-key management and smart-contract interaction, and the brand marketing of leading exchanges manufactures a "sense of safety" that may lack substantive basis. Distinguishing a "rational preference for efficiency" from "path dependence driven by cognitive bias" is essential for assessing the true market potential of on-chain alternatives.

When five roles are combined in one entity, conflicts of interest are no longer occasional; they become an intrinsic property of the architecture. Between the custody function and proprietary trading lies a structural incentive to misappropriate funds; between the matching engine and the market-making desk lies a channel for abusing the informational advantage of order flow; and between the clearing function and affiliated parties lies the possibility of parameter manipulation. Multiple conflict paths criss-cross one another, forming a self-reinforcing network of bundled interests—precisely the core mechanism by which functional bundling turns occasional risk into structural risk (Figure 5-6).

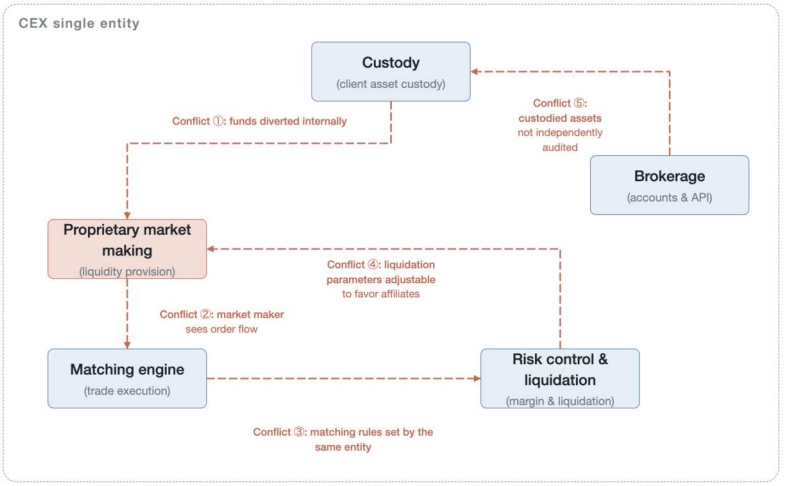


Figure 5-6. Internal functional bundling and conflict-of-interest mapping within a centralized exchange (conceptual conflict-of-interest diagram; edges denote mechanistic relationships, not measured magnitudes; not empirical data; Data source: based on Dell'Erba's 2024 conflict-of-interest analysis [6])

5.3.2 The definition and composition of the bundling tax

A direct consequence of the function-bundling architecture is a systematic hidden cost. Centralized exchanges achieve extremely high trading efficiency by eliminating the friction of cross-institutional coordination, but this efficiency is not without cost. Traders in fact pay a structural risk premium for this convenience, which this chapter defines as the "bundling tax."

The bundling tax is the structural risk cost that users pay for the efficiency and convenience of centralized exchanges, and it consists of three inseparable parts. The opacity cost is the first component; it arises because users cannot independently verify the true safety of their assets, the fairness of order execution, or the effectiveness of the risk-control mechanism. In a centralized exchange's closed system, users can only rely on trust in the exchange operator and cannot verify key parameters for themselves as they could when auditing on-chain smart contracts. FTX before its collapse was the extreme embodiment of this opacity, as revealed by the bankruptcy trustee's statement cited at the opening of this chapter—the complete failure of control and the absence of trustworthy information [2]. The conflict-of-interest cost is the second component; it arises because a single entity simultaneously controls custody, execution, and clearing—three core functions that should check one another. When the exchange is at once the rule-maker and the rule-enforcer, at once the custodian of client assets and the largest participant in the market, both the motive and the ability to favor itself are present. Traditional finance suppresses such conflicts of interest through statutory fiduciary duty and independent regulatory review [3], but in the bundled architecture of a centralized exchange these checks are structurally absent. The single-point-of-failure cost is the third component, and the most easily overlooked. When all core functions are concentrated in one entity, a technical fault, management error, or external attack at

any single link immediately spreads to and paralyzes the entire system. In traditional finance, even if one broker fails, client assets remain safely held at an independent custodian and the operation of the exchange and clearinghouse is unaffected. In a centralized exchange, this risk isolation is entirely absent.

Conceptually, the bundling tax can be expressed as a function of three components: $\mathrm{BT} = f\big(C_{\mathrm{opacity}}, C_{\mathrm{conflict}}, C_{\mathrm{spof}}\big)$, where C_{opacity}, C_{conflict}, and C_{spof} denote the opacity cost, the conflict-of-interest cost, and the single-point-of-failure cost, respectively. Each component can be approximated through proxy variables: proxies for C_{opacity} include the frequency with which proof of reserves is published and the share of on-chain-verifiable assets; proxies for C_{conflict} include the degree of disclosure of related-party transactions and the presence or absence of liquidation-exemption clauses; and proxies for C_{spof} include the scale of historical security incidents and the concentration of assets under custody. Notably, the relationship among these three cost components is not simply additive but multiplicatively interactive. C_{opacity} (opacity) is the enabling condition for C_{conflict} (conflict of interest): only in an environment that outsiders cannot observe can a conflict of interest be systematically exploited. This implies that the optimal strategy for lowering the bundling tax should target opacity first, because greater transparency simultaneously narrows the exploitable space for conflicts of interest. Although this framework has not yet developed into a precise quantitative model, it provides an analytical basis for comparing structural risk across different exchanges.

All three components have already produced serious consequences in reality. The opacity cost manifested in the FTX case as the absence of externally verifiable financial information before the collapse; the conflict-of-interest cost manifested in Alameda Research obtaining a liquidation-exemption privilege and misappropriating $8.7 billion in client funds; and the single-point-of-failure cost was concentrated in the $1.5 billion Bybit theft (Figure 5-7).

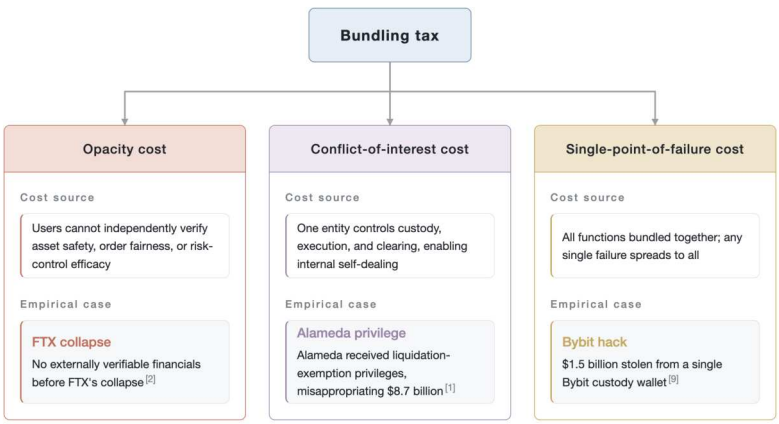


Figure 5-7. The three components of the bundling tax and their empirical cases (Data source: CoinDesk's report on FTX customer liabilities [1], FTX bankruptcy filings [2], and the Wilson Center security report [9])

The bundling tax forms a structural correspondence with the "trust tax" discussed in

Chapter 1. Through mandatory functional separation, traditional finance makes traders pay a "trust tax"—high intermediation fees, long settlement cycles, and strict barriers to access—in exchange for the safety of the system and the suppression of conflicts of interest. Centralized exchanges eliminate most of the "trust tax" through functional bundling, achieving a millisecond-level trading experience and near-zero barriers to market access; but the resulting "bundling tax" requires traders to bear three structural risks: opacity, conflict of interest, and single points of failure. The two are the two ends of the same trade-off, and no cost-free architectural choice exists. This tax cannot be fully eliminated through compliance reviews or technical upgrades alone, because as long as the exchange simultaneously controls the private keys to client assets and the matching engine for trade execution, the technical possibility of evading external audit always remains. On the verifiability ladder defined in Section 5.1, a centralized exchange can at most reach the L1 level of "attested verifiability" by publishing proof of reserves and similar attestations, but it is structurally unable to rise to L2 or above, the "partially verifiable" level.

5.3.3 The efficiency advantages of the bundling model

Despite the significant structural risk the bundling tax entails, centralized exchanges still dominate the crypto-derivatives market. In 2024, the top 10 centralized exchanges recorded $58.5 trillion in total perpetual futures volume, double the $28 trillion of 2023 [4]. The fourth quarter of 2024 alone reached $21.2 trillion in volume [4]. This market dominance is no accident; its economic roots lie in the three efficiency advantages created by functional bundling and the powerful network effects that follow.

The first efficiency advantage of centralized exchanges comes from internal-ledger settlement. In traditional finance, a futures trade from order to final settlement must pass in sequence through the broker, the exchange, the clearinghouse, and the custodian bank, and every cross-institutional handoff of information inevitably generates time delays and coordination costs. Centralized exchanges compress these stages within a single unified technology stack, turning settlement essentially into a numeric update in a centralized database and thereby reducing trading latency to the millisecond level, which supports high-frequency trading strategies and complex algorithmic execution. Traditional finance's multi-institution architecture physically cannot match this internalized coordination efficiency.

A unified liquidity pool constitutes the second efficiency advantage. The buy and sell intentions of all users converge in a single order book, maximizing liquidity depth. Deep liquidity directly lowers the bid-ask spread that traders face and the price impact of executing large orders, which has a direct bearing on the institutional traders and high-frequency market makers analyzed in Chapter 4. By contrast, on-chain protocols' liquidity is dispersed across different blockchains and protocols, producing a fragmented market. Yet a unified liquidity pool can also become a source of risk concentration under extreme conditions: when a single asset undergoes large-scale liquidation, the loss can consume the shared insurance fund and affect the safety margin of all trading pairs.

A minimalist user experience is the third efficiency advantage. Users need not manage complex private keys or seed phrases, need not understand the gas-fee mechanism and

transaction-confirmation process of the underlying blockchain, and can begin trading with only an email address. This low-barrier design significantly expands the pool of potential users, so that crypto-derivatives trading is no longer confined to technically capable early adopters.

Together, these three advantages form a powerful positive feedback loop of network effects. Abundant liquidity attracts more traders, the concentration of traders in turn attracts market makers to offer narrower bid-ask spreads, and better quotes further lower trading costs and attract still more volume. This self-reinforcing loop makes the market position of leading centralized exchanges hard to shake. Even though a market leader such as Binance saw its perpetual futures market share fall from 43% to 34% in 2024, the lost share was absorbed mainly by other centralized exchanges rather than flowing to on-chain protocols [4]. These data indicate that although competition among centralized exchanges is fierce, the dominance of the bundling model as an architectural paradigm has not been substantively challenged.

5.3.4 Comparing the CEX and CCP risk waterfalls

Clearing and risk management are among the core functions of a trading service provider. In designing their risk-defense mechanisms, centralized exchanges have borrowed heavily, in form, from the default-waterfall model of the central counterparty in traditional finance, but fundamental differences remain in the substance of legal constraint, transparency, and user protection. A precise comparison of the two waterfalls helps reveal the structural defects of a centralized exchange's risk-management system.

As Section 5.2.3 described, the central counterparty's default waterfall is a legally binding, multi-layered defense system. Data from Paddrik and Zhang (2020) [5] show that, in credit central counterparties, the defaulting member's initial margin accounts for about 77.9% of the waterfall's total resources and the central counterparty's own capital for about 2.0%. Every parameter setting, trigger condition, and loss-sharing rule of the waterfall is disclosed in advance and subject to mandatory review by financial regulators [20].

A centralized exchange's risk waterfall mimics this structure in form but differs fundamentally on three key dimensions. On transparency, a centralized exchange sets its margin parameters unilaterally and can adjust them at any time, and the adjustment history is usually not disclosed; the size, funding source, and specific use of the insurance fund often cannot be independently verified from the outside. On legal constraint, a central counterparty's capital adequacy, risk models, and default-handling procedures are bound by statutory regulatory requirements, whereas a centralized exchange in most jurisdictions lacks statutory obligations of comparable rigor. On user protection, a central counterparty's non-defaulting members enjoy loss-sharing rules disclosed in advance and statutory recovery procedures, whereas a centralized exchange's users, when the platform goes bankrupt, can usually participate in bankruptcy proceedings only as general unsecured creditors, ranking behind secured creditors. This generalization, however, is subject to significant cross-jurisdictional variation: in the FTX bankruptcy, convenience-class creditors (with claims below $50,000) ultimately received a recovery of about 119% on their dollar-denominated claims, but claims were fixed at their

U.S.-dollar value as of the November 2022 filing date, so customers did not capture the subsequent appreciation of crypto assets, and the portion above 100% came from principal plus roughly 9% statutory interest and recoveries from the bankruptcy estate. The Mt. Gox case in Japan, by contrast, achieved an in-kind (BTC) distribution through civil rehabilitation proceedings, giving customers procedural treatment different from a fixed dollar claim. A blanket "unsecured creditor" judgment may understate the differences in user protection across legal systems.

The risk waterfalls of the three architectures differ fundamentally in their enforcement guarantees: the central counterparty's waterfall is legally binding, and the trigger conditions and loss-sharing rules of each layer are subject to regulatory review and disclosed in advance; a centralized exchange's waterfall is merely asserted by the platform, and its parameter settings and adjustment history are invisible to outsiders; and an on-chain protocol's waterfall is enforced by smart-contract code, with rules hard-coded and auditable by anyone. The three are highly similar in the formal design of their waterfall layers, but the nature of their enforcement guarantees produces a sharp contrast among four modes—legal mandate, platform attestation, code enforcement, and code enforcement with governance override (Figure 5-8)—which is precisely the key variable determining whether a waterfall can be faithfully executed in a moment of crisis. The fourth mode, code enforcement with governance override, refers specifically to the emergency governance-intervention capability that on-chain protocols retain on top of automatic smart-contract execution. When an extreme event exceeds the handling range of the code's preset parameters (such as the JELLY incident analyzed in Section 5.4.4), validators or governance-token holders can override the code logic through a consensus vote. The conditions for such override capability include a validator set small enough for rapid coordination, governance power that is not fully decentralized, and the absence of an on-chain time-lock constraint on emergency intervention. Its trigger threshold is usually the critical situation in which the insurance fund faces depletion or the protocol's solvency is threatened.

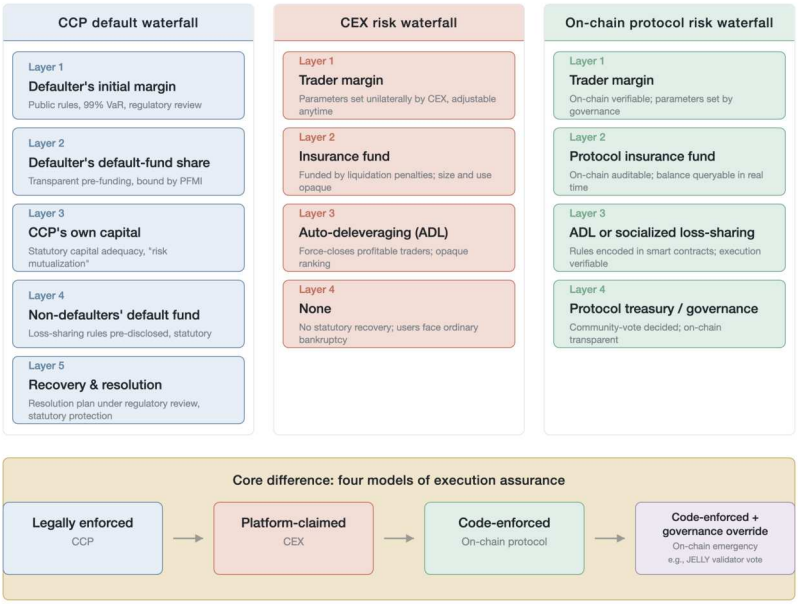


Figure 5-8. A comparison of the risk waterfalls of the central counterparty, the centralized exchange, and on-chain protocols (conceptual comparison of mechanism structures; the waterfall layers and the "99% VaR" follow the CPMI-IOSCO PFMI [20], not empirical data; the 77.9%/2.0% figures from [5] appear in the text, not the figure; Data source: Paddrik & Zhang 2020 [5] and the CPMI-IOSCO Principles [20])

Under normal market conditions, the differences among the three modes of legal mandate, self-attestation, and code enforcement may not be obvious, because the first few layers of the waterfall are usually enough to absorb everyday trading losses. In a moment of crisis, however, when extreme volatility causes systemic shortfalls, a centralized exchange that lacks statutory recovery procedures and transparent loss-sharing rules often leaves users facing an unpredictable allocation of losses. When the insurance fund is exhausted, a centralized exchange usually resorts to auto-deleveraging, forcibly closing the positions of profitable users to make up the loss gap—and the specific algorithm and execution details of this ranking mechanism likewise lack transparency.

5.3.5 The limits of proof of reserves

To ease users' concerns about the opacity cost, centralized exchanges widely introduced proof-of-reserves mechanisms after FTX's collapse. Both academic research and industry practice, however, indicate that proof of reserves has fundamental limitations in addressing the structural risk created by functional bundling.

The core defect of proof of reserves is that it can provide only a point-in-time snapshot of assets and cannot guarantee that assets have not been misappropriated or put to other uses during the audit interval. More critically, proof of reserves usually shows only the on-chain asset balances the exchange controls and cannot effectively reveal its hidden liabilities, complex related-party transactions, or off-balance-sheet commitments. Research by Vidal-Tomás (2025) [14] further points out that, to maintain solvency under

extreme market stress, centralized exchanges in fact need to hold additional reserves of 6% to 14% above their liabilities. The study also finds that, among all exchange-bankruptcy events, the collapses of FTX and Celsius had the most significant negative impact on the stability of the entire centralized-exchange market [14], indicating that the failure of a single large exchange has a systemic contagion effect. In the highly opaque system of a centralized exchange, the capacity for external audit faces an unbreakable ceiling—not because audit technology is insufficiently advanced but because the function-bundling architecture itself makes it inevitable. When the audited party simultaneously controls the assets being audited and the system that generates the audit data, the independence of the audit cannot be guaranteed as a matter of logic.

In the face of these structural risks, regulators worldwide have begun to adopt more systematic responses. The European Union's Markets in Crypto-Assets Regulation (MiCA), which took full effect on December 30, 2024, requires crypto-asset service providers to obtain a license and to comply strictly with requirements on client-asset segregation, conflict-of-interest disclosure, and capital adequacy [24]. Notably, MiCA excludes crypto derivatives such as perpetual futures from its direct scope, placing them under the Markets in Financial Instruments Directive (MiFID II); this means that perpetual futures trading in the EU is in fact subject to the regulatory framework for traditional derivatives rather than a dedicated framework for crypto assets. MiCA's Article 70, on the general safekeeping of clients' crypto-assets and funds, poses a direct challenge to the function-bundling model, requiring exchanges to separate client assets from their own assets at the legal and operational levels. This regulatory framework, however, runs into an enforcement dilemma when confronting on-chain protocols: when the "service provider" is a set of validator nodes distributed around the world rather than a registered legal entity, the licensing regime and compliance requirements lack a clear point of enforcement. The U.S. Securities and Exchange Commission has also launched enforcement actions against several leading centralized exchanges. These regulatory measures have, to a considerable degree, raised the legal cost of exchange misconduct, but Dell'Erba (2024) [6] warns that the effectiveness of traditional financial regulation depends heavily on the clear regulatory entry points that the functional-separation architecture provides. When all functions are concentrated inside one entity, regulators are in effect trying to supervise a complex, opaque, closed system from the outside. Merely adding compliance requirements without structurally unbundling the centralized exchange's functions may amount only to "adding more layers on an unstable foundation" [6]. A compliance framework can raise the legal cost of wrongdoing but cannot eliminate the technical possibility of wrongdoing at the base layer. On the verifiability ladder, this means that regulation can help a centralized exchange rise from L0 to L1 but cannot push it across to the verifiable levels of L2 and above.

5.3.6 The FTX and Bybit failure cases

The fragility of the centralized exchange's function-bundling model has already been validated in reality along two distinct paths: one the complete collapse of internal governance, the other the failure of external technical defenses. These two paths respectively represent the extreme realization of the conflict-of-interest cost and the single-point-of-failure cost within the bundling tax, and together they prove that the architecture's

structural risk is not an accidental, isolated case.

The extreme case of internal-governance collapse is FTX's bankruptcy. On November 11, 2022, FTX filed for bankruptcy protection, and subsequent investigation revealed that it had misappropriated about $8.7 billion in client funds [1]. These funds were used to cover trading losses at the affiliated market maker Alameda Research, to fund personal loans to executives, and to purchase luxury real estate. Using its dual identity as both matching-engine operator and risk manager, FTX granted Alameda Research the privilege of exemption from automatic liquidation at the level of code, allowing it to maintain enormous positions without sufficient collateral. The essence of this episode is that functional bundling turned all of the exchange's core functions—custody, clearing, and risk control—into tools serving the interests of an affiliated party. Functional bundling made misappropriation technically trivial, because the same entity that controlled the matching engine also controlled the private keys to client assets; the absence of independent external oversight kept the misappropriation undiscovered for a long time, because no independent third party could verify the true state of the assets; and the temptation of billions of dollars in commercial gain made the misappropriation almost inevitable in motive. The FTX case proves that relying on a centralized exchange's self-restraint to manage conflicts of interest is entirely unworkable.

The archetype of technical fragility is the hack Bybit suffered on February 21, 2025. Because the exchange used a centralized-custody model, a vast quantity of client assets was concentrated in a wallet system controlled by a single entity, an architecture that made it a target with high expected returns. The North Korea-linked Lazarus Group, through months of social-engineering infiltration and malware deployment, successfully breached the exchange's internal security defenses and stole about $1.5 billion in digital assets [9]. Specifically, the attackers used a supply-chain attack to infiltrate the signing interface of a multi-signature cold-wallet solution, so that what the signers saw visually was a normal transaction while what they actually signed was a malicious transfer instruction. This attack vector reveals that the real security boundary of a multisig scheme is far lower than its theoretical expectation: when the signing device or interface is compromised, the numeric threshold of the multisig becomes meaningless. Notably, the same attack vector applies equally to multisig governance wallets in DeFi protocols, and the claim that "self-custody eliminates custody risk" must be re-examined in the face of supply-chain attacks. The stolen funds were then dispersed within the first 48 hours across dozens of wallet addresses each holding about 10,000 ETH and were further laundered in layers through decentralized exchanges, cross-chain bridges, and mixers [9], using decentralized exchanges to swap assets and evade freezing. Addresses associated with the Lazarus Group appear on the sanctions list of the U.S. Treasury's Office of Foreign Assets Control (OFAC), which poses a compliance challenge for the entire on-chain ecosystem; the Tornado Cash sanctions precedent shows that decentralized protocols that interact with sanctioned addresses may themselves face regulatory risk. This was the largest crypto theft up to that time, and its scale further confirms the structural fragility of the centralized-custody model: under this architecture, the defender must maintain a perfect defense against every attack, while the attacker need succeed only once to inflict a major, irreversible loss. However much an exchange invests in internal security, the very fact that a vast quantity of assets is concentrated

under the control of a single entity constitutes an ineliminable systemic risk.

Whether FTX's internal corruption or Bybit's external breach, the root cause points to the same conclusion: the risk of the function-bundling model is not a matter of a particular operator's moral failure or technical negligence but a structural property of the architecture itself. FTX's collapse exposed the fragility of the bundled architecture along the governance dimension, and Bybit's theft exposed the fragility of the same architecture along the technical dimension. The existence of two failure paths shows that even if a centralized exchange's governance system is sound, it still faces the single-point-of-failure risk at the technical level; conversely, even if its technical defenses are sound, a failure of internal governance can still destroy the entire system at the root. This repeated disappointment with the "trust an institution" model is precisely the fundamental force driving the market to explore alternatives more seriously. Those on-chain trading models that seek to rebuild the foundation of trust through publicly verifiable code and cryptography are analyzed in detail in Section 5.4.

5.4 On-chain trading models

On-chain perpetual futures trading has expanded rapidly over the past several years. By 2025, decentralized exchanges handled $6.7 trillion in perpetual futures volume, up 346% from the prior year and accounting for about 10% of the overall perpetual futures market [25] [26]. This growth reflects traders' persistent demand for alternatives to centralized exchanges. Decentralized exchanges, however, have not eliminated the risks inherent in financial markets; through cryptography and consensus mechanisms, they transform risk from one form into another. This section examines in depth the core mechanisms of on-chain perpetual futures trading and analyzes the new risk exposures it introduces in self-custody, execution models, and app-chain architecture.

5.4.1 Self-custody and atomic settlement

The core promise of decentralized-finance infrastructure is to reshape asset control and the trade-settlement process through cryptography. On a traditional centralized exchange, users must deposit assets into a wallet the platform controls, the platform records users' account balances in an internal database, and trades are booked in the matching engine. This model introduces extremely high centralized-custody risk: once the platform is hacked or funds are misappropriated internally, users' assets face direct loss. The Bybit theft of about $1.5 billion discussed earlier (Section 5.3.6) further demonstrates the fragility of the centralized-custody model [27].

On-chain perpetual futures trading eliminates this platform-custody risk through a self-custody mechanism. Under self-custody, the user controls assets on the blockchain directly through private keys, and a smart contract can operate on those assets only under specific conditions authorized by the user's cryptographic signature. This mechanism ensures that the platform cannot unilaterally misappropriate user funds, eliminating traditional centralized-custody risk at the root. On-chain trading also uses an atomic-settlement mechanism, completing trade matching and fund transfer indivisibly within the same blockchain transaction—if any step fails, the whole transaction

reverts—thereby eliminating settlement risk (see Chapter 1 for the discussion of atomic settlement).

The complete flow of on-chain perpetual futures trading includes connecting a wallet, depositing margin, submitting an order, on-chain matching and settlement, and withdrawing funds (Figure 5-9); throughout, assets remain under the control of the user's private keys, a smart contract can operate on funds only under the specific conditions authorized by a cryptographic signature, and matching and fund transfer complete atomically within the same blockchain transaction.

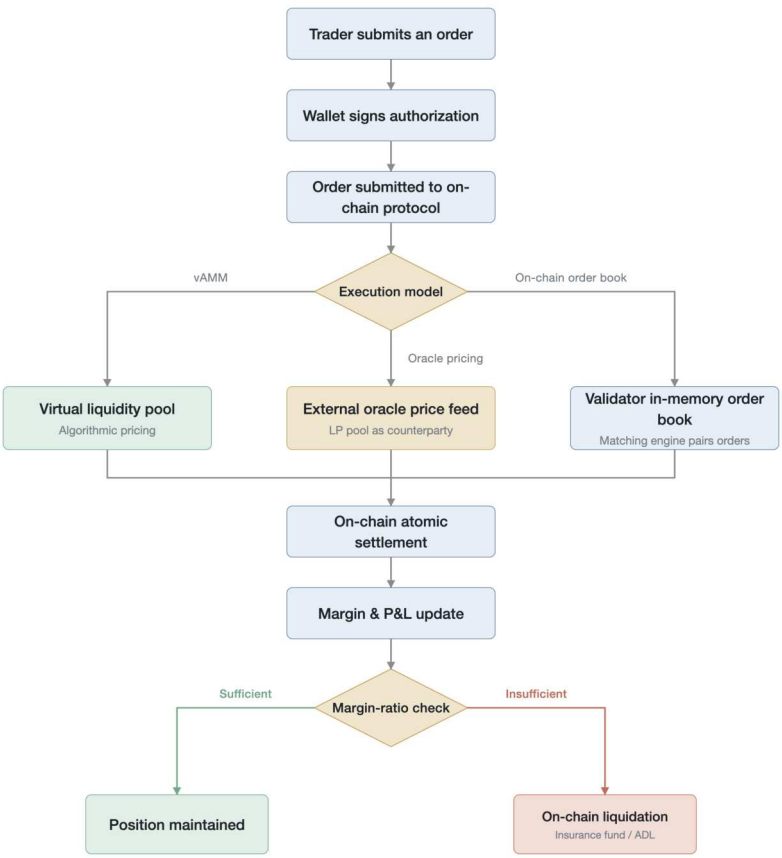


Figure 5-9. The on-chain perpetual futures trading flow (a schematic of the self-custody and atomic-settlement flow, not empirical data; Data source: protocol-architecture analysis based on the dYdX Foundation [28] and Nansen Research [29])

This elimination of risk, however, is not without cost; it essentially transforms institution-level systemic risk into user-level operational risk and protocol-level technical risk. Self-custody requires the user to assume full responsibility for protecting the private keys. Once a private key is lost or stolen, the assets face irreversible loss, and no centralized institution can provide a recovery service. Moreover, self-custody now faces tightening regulatory constraints: the EU's Transfer of Funds Regulation (TFR) requires crypto-asset service providers to apply additional identity verification

to transfers involving unhosted wallets, and the FATF Travel Rule likewise extends compliance requirements to self-custody users, which may in practice limit the convenience with which self-custody users interact with regulated services. In addition, when interacting with a smart contract, a user must authorize the contract to operate on the assets in the wallet. If the user inadvertently authorizes a malicious smart contract, or if a legitimate smart contract contains a vulnerability, an attacker can still drain the user's assets.

Against the backdrop of increasingly frequent cross-chain trading, cross-chain bridge contracts have become a new locus of concentrated systemic risk. To move assets between different blockchain networks, a user must lock the assets in a bridge contract on the source chain and mint an equivalent amount of wrapped assets on the destination chain. This mechanism makes bridge contracts high-value attack targets. As noted earlier, the cumulative losses from cross-chain bridge hacks have exceeded $2.5 billion, about 40% of the total value stolen across Web3 [11]; this shows that although on-chain trading eliminates the custody risk of a single exchange, the complexity of cross-chain infrastructure introduces smart-contract-vulnerability risk that is more insidious and far more destructive.

5.4.2 A comparison of three execution models

To achieve efficient derivatives trading in a decentralized environment, on-chain perpetual futures protocols have evolved three main execution models. Each makes a different trade-off among performance, decentralization, and transparency, and each exposes particular risk characteristics in actual operation.

The virtual automated market maker (vAMM) model is the archetypal design of early on-chain perpetual futures. This model borrows the constant-product formula of the automated market maker from spot markets, but the pool contains no real reserve assets; instead, it uses an algorithm to simulate liquidity depth and determine prices. After depositing collateral, users take long and short positions against one another in the virtual pool. The advantage of this design lies in its high degree of decentralization and its permissionless nature: anyone can create a new trading pair based on the algorithm. A vAMM, however, relies entirely on internal-algorithm pricing, and when the market experiences an extreme one-sided move, the virtual pool's price readily depegs severely from the external spot price. The first version of Perpetual Protocol once generated large amounts of bad debt because of price depegging, ultimately facing bankruptcy risk [30]. The specific consequences of depegging for traders include immediate unrealized losses for position holders at the moment of depegging, unreasonable forced liquidations triggered by a mark price that deviates from the true price, and funding rates that spike to abnormal levels because of an extreme imbalance in the long-short ratio.

The oracle-pricing model takes an entirely different approach. It abandons the internal price-discovery mechanism and instead relies on the spot prices of mainstream exchanges provided by an external oracle as the execution price. In this model, liquidity providers (LPs) pool their funds into a global liquidity pool that serves as the unified counterparty for all traders. Traders can enjoy a zero-slippage trading experience because the price is determined entirely by the external oracle. This design, however,

introduces fundamental oracle risk and LP-loss risk. Oracle risk can be systematically divided into four categories: manipulation risk, in which an attacker manipulates the price of the underlying asset in an illiquid external spot market and exploits the feed lag to perform risk-free arbitrage on the on-chain protocol; feed-latency risk, in which the oracle network's update frequency cannot fully match instantaneous market movements, creating a pricing window under extreme volatility; upstream data-source dependency risk, in which the oracle's aggregated price sources are concentrated in a few centralized exchanges, so that an interruption or error in the upstream data source transmits directly to on-chain pricing; and operator conflict-of-interest risk, in which, when the protocol's validators also serve as oracle nodes, the combination of pricing power and settlement power can create an incentive for self-dealing. In September 2022 and July 2025, leading oracle-pricing protocols such as GMX suffered repeated price-manipulation and reentrancy attacks of this kind, causing losses ranging from several million to tens of millions of dollars [31] [32].

The on-chain order-book model represents the latest direction in on-chain derivatives trading. This model seeks to rebuild on the blockchain the time-tested central limit order book (CLOB) mechanism of traditional financial markets. Early hybrid architectures placed order matching on off-chain centralized servers and submitted only the final fund settlement to the chain, trading this for higher throughput. But this compromise sacrificed the transparency and censorship resistance of the matching process. A new generation of on-chain order-book protocols instead builds a dedicated app-chain that embeds the matching engine, margin checks, and liquidation logic entirely within the blockchain's consensus layer. This model can deliver performance close to that of a centralized exchange while preserving deterministic trade ordering and full on-chain transparency.

Notably, the three models differ structurally in the design of their funding-rate mechanisms: the vAMM's funding rate is implicit in the curve shift of the virtual pool and is not calculated independently; the oracle-pricing model's funding rate is usually calculated explicitly on the basis of the long-short position ratio; and the on-chain order-book model's funding-rate calculation is closer to that of a centralized exchange, based on the deviation between the contract price and the mark price. These differences directly affect each model's contribution of the funding rate to the efficiency of price anchoring. The relative positions of the three models within the trilemma are shown in Figure 5-10.



Figure 5-10. A comparison of the positions of three DEX execution models within the trilemma (conceptual schematic; qualitative positions, not measured coordinates; placements reflect the author's judgment, not empirical data; Data source: Nansen Research [29])

As the figure shows, the chosen points of the three models form an evolutionary path extending from the decentralization vertex toward the performance vertex: the vAMM is closest to the decentralization and transparency vertices but farthest from the performance vertex; the oracle-pricing model occupies the middle ground, trading dependence on external infrastructure for higher capital efficiency; and the on-chain order-book model advances performance substantially through an app-chain architecture but moves correspondingly away from the decentralization vertex. This path reveals the core evolutionary logic of on-chain derivatives-trading execution models: while preserving on-chain transparency, they trade a gradual concession in decentralization for a performance experience approaching that of a centralized exchange.

5.4.3 App-chain architecture and validator centralization

As on-chain trading's demands for low latency and high throughput keep rising, the performance bottleneck of general-purpose blockchains has become increasingly apparent. To break through this limit, leading decentralized exchanges have begun migrating to an app-chain architecture. An app-chain is an independent blockchain tailored to a specific decentralized application, and its core idea is to bind the underlying consensus mechanism tightly to the upper-layer application logic.

In a hybrid architecture, a decentralized exchange is usually deployed on a general-purpose blockchain such as Ethereum or its layer-2 network. Although this architecture can share the underlying network's vast validator set and high degree of security, a general-purpose network must accommodate the execution of all kinds of smart contracts and cannot provide the deterministic low latency required for high-frequency trading. Moreover, a hybrid architecture relies on centralized infrastructure such as Amazon Web Services for off-chain matching, which not only introduces single-point-of-failure risk but also leaves the trade-ordering process without cryptographic verifiability

[28].

By contrast, an app-chain architecture achieves full-stack integrated control from the consensus layer to the application layer. In the case of a decentralized exchange that adopts an app-chain architecture, it usually customizes a Byzantine fault-tolerant (BFT) consensus algorithm optimized specifically for the throughput of financial messages. Under this architecture, order submission, cancellation, and matching, along with margin liquidation, all take place under the protection of on-chain consensus. Validator nodes are responsible not only for packaging blocks but also for running the order-book state machine directly in memory, ensuring that the execution order of every trade is network-wide consistent and tamper-proof. This deep customization allows an app-chain to achieve sub-second block-confirmation times and a throughput capacity of hundreds of thousands of orders per second (here, throughput refers to message throughput—the processing rate of messages such as order submissions and cancellations—and actual fill throughput is typically far lower than message throughput) [33] [34].

The core difference between the hybrid architecture and the app-chain architecture lies in where the components sit and in the corresponding trust assumptions. In a hybrid architecture, the matching engine runs on an off-chain centralized server and only final settlement is submitted to the chain, so users must trust the fairness of the off-chain components; in an app-chain architecture, the matching, margin checks, and liquidation logic are all embedded in the blockchain's consensus layer, and trade ordering is jointly determined by the validator set. This difference gives the two architectures different performance boundaries in latency, throughput, transparency, and censorship resistance (Figure 5-11), and the fundamental difference in their trust assumptions is that the former's trust rests on "whether the off-chain platform is fair," while the latter's shifts to "whether the validator set is honest."

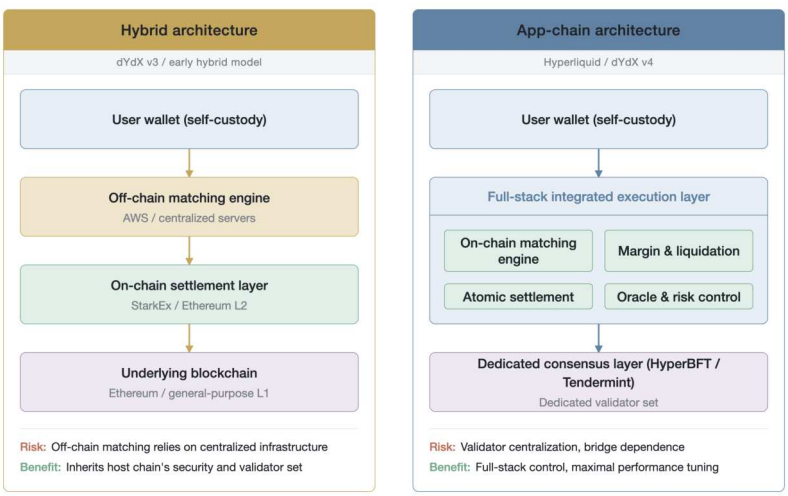


Figure 5-11. An architectural comparison of a hybrid DEX and an app-chain (conceptual architectural comparison, not empirical data; Data source: dYdX Foundation [28])

While gaining high performance, however, the app-chain architecture faces the significant cost of validator centralization. To maintain an extremely high network-synchronization speed, an app-chain can usually accommodate only a limited number of validator nodes. For example, in early 2025 Hyperliquid maintained network consensus with only about 16 validator nodes (later expanded to about 24), and the staked tokens were highly concentrated in the hands of early investors and the team [12] [35]. Beyond centralization risk at the governance level, the infrastructure concentration of validators also constitutes a significant operational risk: if most validators run on the same cloud provider or in data centers in the same geographic region, a single data-center failure or network partition can cause the network to lose liveness. Solana's multiple network-wide outages between 2021 and 2023 provide a precedent: during a network outage, all unsettled perpetual futures positions face the risk of not being liquidated, and the combination of extreme market conditions and network downtime can cause large-scale depletion of the insurance fund. This highly concentrated validator structure significantly weakens the blockchain's censorship resistance and its decentralized trust assumptions. When a few validators hold absolute control of the network, they can not only determine the ordering of trades but even change the protocol's core rules through a consensus upgrade. This concentration of power can cause an app-chain, under extreme conditions, to exhibit the arbitrary characteristics of a centralized institution, thereby departing from the original intent of decentralized finance.

5.4.4 Architecture audit and verifiability rating

Objectively assessing the security and degree of decentralization of an app-chain architecture requires a fact-based verifiability-rating framework. This framework does not rely on subjective scoring; instead, it performs a structured analysis by examining, for each dimension of the system, whether it is on-chain queryable, auditable, and under user control. Taking Hyperliquid, the current volume leader, as an example (other app-chain architectures such as dYdX v4 face similar trade-offs), one can conduct a deep audit of its architecture along four core dimensions.

On the dimension of custody and asset safety, this app-chain exhibits relatively high verifiability. Users' account balances are fully recorded in the on-chain state, and anyone can independently verify the authenticity of their assets by running a full node, without relying on financial statements the platform provides [36]. The platform does not hold users' private keys, and asset transfers require authorization by the user's cryptographic signature. On the withdrawal path, however, the system carries a partially restricted risk. Because this app-chain is an independent network, cross-chain asset transfers depend heavily on a cross-chain bridge contract controlled by a validator multisig. Although the bridge-contract code is publicly auditable, its security is ultimately anchored to the honesty of the validator set.

On the dimension of trade execution and matching, this system achieves fully on-chain execution of order matching. Unlike a hybrid model that relies on centralized servers, its consensus mechanism ensures deterministic trade ordering, and all matching logic runs openly in the validator nodes. The system's low latency and high throughput have been validated in actual operation. Hyperliquid's core node code long remained closed-

source and was only gradually opened after sustained scrutiny from the community [37]. In addition, on preventing MEV, the system currently lacks an independent third-party audit, and the fairness of trade ordering still largely depends on trust in the validators.

On the dimension of clearing and risk management, the system's margin checks and liquidation-engine logic both execute on-chain, and their trigger conditions and execution processes are fully public and queryable. The insurance fund's balance is likewise displayed transparently on-chain, and the rules of the auto-deleveraging mechanism are hard-coded into the protocol. The system's oracle price-feed mechanism, however, involves a significant compromise. In pursuit of the highest update speed, the system does not use a decentralized third-party oracle network but instead has its own validator set directly maintain the price oracle. Although this design reduces latency, it also gives validators both pricing power and settlement power, which can trigger serious conflicts of interest and manipulation risk under extreme conditions.

Combining the audit results across these four dimensions, this app-chain achieves relatively high on-chain verifiability in custody and trade execution, carries latent conflict-of-interest concerns in clearing and risk management because of the compromise of a self-built oracle, and, in governance and emergency response, exposes discretionary risk arising from validator centralization (Figure 5-12).

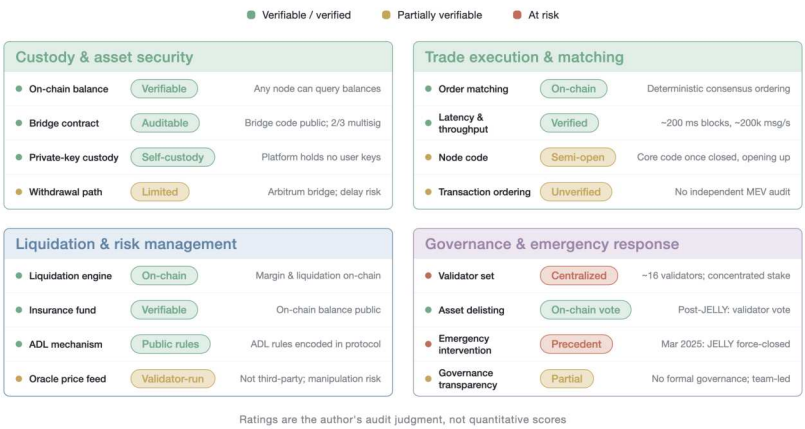


Figure 5-12. The verifiability-rating matrix from a four-dimensional architecture audit of the app-chain (the author's four-dimensional structured rating; ratings reflect the author's audit judgment and are ordinal, not quantitative scores, and not empirical data; the figure of about 16 validators reflects the genesis count as of early 2025, later expanded to about 24 [12][35]; the JELLY incident resolution [38][39]; the roughly 200,000 per second figure is message throughput—orders, cancellations, and similar messages, not fills—on a vendor basis [33][34]; architecture [28]; OneKey Blog is a secondary review [37])

On the dimension of governance and emergency response, this app-chain exposes the most obvious centralization risk. The validator set is limited in number and its stake is highly concentrated, so the system's governance lacks broad community participation. In responding to sudden events, this centralized structure grants validators extensive discretion.

The JELLY incident of March 2025 is an in-depth case of this governance risk [38] [39].

A trader opened a large short position in the low-liquidity token JELLY on Hyperliquid and then sharply pushed up the token's price in the external spot market. Because JELLY's spot liquidity was extremely thin, a small amount of capital could significantly move the oracle price feed, causing the Hyperliquid liquidity provider vault (HLP), as it passively absorbed the liquidated short, to face a peak unrealized loss of about $12 million (with exposure at one point reaching about $230 million); because the position was ultimately liquidated at the very low price of $0.0095, the HLP actually booked a small profit. This incident first exposed the fundamental risk of listing perpetual futures on low-liquidity tokens: when the oracle price feed can be manipulated with a small amount of capital, the leverage effect of perpetual futures amplifies the manipulation gains many times over.

The validator set's response to this crisis raised deeper questions of governance legitimacy. The validators rapidly reached consensus, decided to forcibly delist JELLY, and forcibly closed all related positions at a price of $0.0095, while the oracle quote at the time was about $0.50 (the on-chain delisting-vote mechanism was introduced only after this incident, and the handling at the time was in fact a rapid coordination among validators). The exercise of this price-setting power was in essence a product of off-chain coordination, and its basis of legitimacy was neither the automatic execution of code nor sufficient community deliberation. The Hyper Foundation subsequently promised to compensate affected users, but this compensation mechanism was itself an off-chain promise outside the protocol rather than a deterministic guarantee at the smart-contract level.

Viewed from the dimension of legal risk, the validators' unilateral forced liquidation at a price far below the market price could constitute market manipulation or a breach of fiduciary duty. The CFTC's enforcement precedent against Ooki DAO has shown that decentralized governance does not exempt participants from legal liability, and the narrative that "code is law" is fragile in the face of regulatory enforcement.

The implications of the JELLY incident for this chapter's theoretical framework are multidimensional. On the verifiability ladder, Hyperliquid has L3-level verifiability in normal operation, but the validators' governance override in a moment of crisis effectively downgraded it to L1, with the final enforcement of the rules depending on the discretion of a few rather than the determinism of code. Within the trilemma, the incident revealed a latent third dimension between decentralization and performance: when a crisis arrives, a limited validator set is forced to tilt toward centralized decision-making, sacrificing the predictability of the rules for the sake of protecting the system's solvency. The community's reaction to the incident also displayed a marked framing-effect split: some users viewed the intervention as a "responsible emergency response," while others viewed it as a "fundamental betrayal of the commitment to decentralization," and this narrative contest directly affected the market's repricing of trust in the protocol.

5.4.5 The on-chain risk waterfall

Sections 5.2.3 and 5.3.4 analyzed, respectively, the legally mandated default waterfall of the traditional central counterparty and the self-attested risk waterfall of the central-

ized exchange. The former manages systemic risk through a statutory, multi-layered loss-absorption mechanism whose parameters are public and subject to regulatory review; the latter mimics this structure in form but differs fundamentally along three key dimensions—transparency, legal constraint, and user protection.

Building on the experience of both models, on-chain perpetual futures protocols have constructed a distinctive on-chain risk waterfall, but the waterfall structure differs markedly across protocols. In the case of Hyperliquid, its risk waterfall in fact comprises four layers: the first layer is the trader's initial and maintenance margin; the second layer is the backstop liquidation of the HLP—that is, when routine liquidators fail to take over in time, the protocol-operated market-making vault passively absorbs the liquidated position; the third layer is the on-chain insurance fund (about $300 million to $500 million in size as of early 2025); and the fourth layer is auto-deleveraging. dYdX v4's waterfall differs: after margin is exhausted, the position is taken over by an external liquidator network, and if liquidation produces a shortfall loss, it is absorbed by the insurance fund (on the order of tens of millions of dollars in size), with ADL initiated after the insurance fund is exhausted.

The absolute sizes of these insurance funds differ from those of the traditional central counterparty by orders of magnitude. CME's total default resources (including the default fund and the statutory power to levy additional assessments on members) exceed $10 billion, of which the default fund itself is about $3 billion, whereas an on-chain protocol's insurance fund is usually only tens of millions to hundreds of millions of dollars and lacks a mandatory replenishment mechanism. Because of the blockchain's transparency, all margin balances and liquidation thresholds are publicly queryable, and any market participant can monitor the system's overall leverage level. When a user account's margin ratio falls below the maintenance level, the smart contract automatically triggers the liquidation logic. Unlike a centralized exchange, an on-chain protocol usually allows external liquidators to participate in the liquidation process and incentivizes them, through a liquidation discount, to take over high-risk positions in time, thereby using market forces to defuse risk. The reliability of the external liquidator network, however, has its own operational fragility. During "Black Thursday" in March 2020, MakerDAO's liquidation auctions saw the extreme situation of a liquidator winning bids at a price of zero: when a spike in Ethereum gas fees made the cost of a liquidation transaction higher than the gain from the liquidation discount, liquidators chose to withdraw, forming a "liquidator strike" that left the system exposed and unprotected.

The funding source of an on-chain insurance fund is usually a share of the trading fees the protocol collects, and its balance is fully visible on the blockchain [40]. This transparency lets the market assess the protocol's risk-bearing capacity in real time. When a protocol faces an extreme black-swan event and the insurance fund is breached, the auto-deleveraging rules are fully hard-coded into the smart contract. The system ranks accounts according to a preset algorithm, by leverage ratio and profit ratio, and executes deleveraging transparently and deterministically. In addition, in some app-chain architectures, as the JELLY incident in Section 5.4.4 shows, the validator set can also exercise emergency-intervention power through a consensus vote—for example, forcibly halting a market or modifying the settlement price.

The distinctive role the funding rate plays in the risk waterfall deserves separate discussion. The funding rate is in essence a continuously running risk-redistribution mechanism with no counterpart in the traditional central counterparty's default waterfall. Under normal market conditions, the funding rate maintains the anchoring of the contract price to the spot price by transferring payments between longs and shorts. Under extreme conditions, however, the funding rate can produce a positive-feedback effect: when the market rises sharply, the positive funding rate spikes, attracting short sellers to enter to earn the rate; if the uptrend persists, these short sellers are liquidated, further pushing up the price and the funding rate, forming a self-reinforcing loop of "high rate → attracts shorts → trend continues → shorts are liquidated → higher rate." This mechanism means that under extreme conditions the funding rate can turn from a stabilizer into an amplifier, constituting a distinctive risk-transmission channel in the on-chain risk waterfall that differs from traditional finance.

The three market structures display systematic differences across five key dimensions of the risk-management mechanism: the frequency of margin monitoring rises progressively from daily settlement to real-time on-chain checks; liquidation execution evolves from a closed internal engine toward an open external liquidator network; the transparency of the insurance fund moves from periodic audit reports to a real-time-queryable on-chain balance; the ultimate backstop mechanism shifts from statutory member assessments to algorithmically determined auto-deleveraging; and the guarantee of rule enforcement moves from legal constraint to cryptography and consensus. Table 5-2 lays out the comparison across these five dimensions. One evolutionary throughline runs across all dimensions: from the traditional central counterparty to on-chain protocols, margin monitoring moves from discrete periodic settlement to continuous real-time checking, liquidation execution moves from a closed membership system to an open liquidator market, and the insurance fund moves from periodic audit to real-time on-chain querying. This evolution takes gains in transparency and determinism as its benefit and takes declines in the depth of financial resources and the flexibility of crisis response as its cost.

Dimension of comparison	Traditional CCP	Centralized exchange	On-chain derivatives protocol
Margin monitoring	Daily or intraday settlement	Real-time risk engine	Real-time checks by on-chain smart contracts
Liquidation executor	Clearing members	Platform's internal engine	External arbitrageurs or protocol liquidators
Insurance fund transparency	Periodic audit reports	Opaque or partially disclosed	On-chain balance queryable in real time
Ultimate backstop mechanism	Member assessments	Auto-deleveraging (opaque)	Auto-deleveraging (algorithmically determined)
Rule-enforcement guarantee	Legal and regulatory constraints	Platform reputation and internal controls	Cryptography and consensus mechanisms

Table 5-2. A comparison of risk-management mechanisms across the three structures of derivatives markets (Data source: Harris [3], Paddrik & Zhang [5], CPMI-IOSCO [20])

On the whole, the on-chain risk-waterfall mechanism is markedly superior to the centralized exchange in transparency and the determinism of its rules, but it still falls short of the strictly regulated traditional central counterparty in the depth of financial resources and the flexibility of responding to extreme crises. The on-chain trading model has not eliminated the leverage risk and liquidity risk inherent in derivatives markets; instead, through cryptographically verifiable smart contracts and decentralized consensus, it transforms the management of risk from unverifiable internal operations into publicly auditable algorithmic execution. This transformation grants users greater rights to information and control, but it also demands of them a higher capacity for risk identification, in order to cope with new on-chain systemic risks such as smart-contract vulnerabilities, oracle failure, and validator consensus collusion.

5.5 Cross-architecture risk isomorphism

Many of the new problems of on-chain finance are, at a deeper level, the recurrence of old problems from traditional finance in a new technical environment. Recognizing this isomorphism helps strip away the technical shell to see the essence of a problem and to draw on traditional finance's two centuries of experience in solving it. At the same time, the on-chain versions of these new forms exhibit fundamental differences in speed, scale, and transmission mechanism, requiring market participants to reassess the boundaries of risk. This section analyzes, in turn, four core isomorphic mappings and characterizes the complete path of cross-modal risk transmission.

5.5.1 Adverse selection and impermanent loss

Adverse selection is the core dilemma facing market makers; in essence, a market maker is always trading against counterparties who hold more information than it does, and so it continuously incurs losses. Glosten and Milgrom (1985) [41] proved that under information asymmetry, even a risk-neutral market maker still posts a positive bid-ask spread, because the spread is the rational compensation for adverse-selection risk. In a traditional order-book market, the market maker manages this risk by adjusting the width of its quotes: when the share of informed traders rises, the spread widens to cover the higher adverse-selection cost; when market information is relatively uniform, competitive pressure forces the spread to narrow. Harris (2003) [3] treats adverse selection as the core issue of market-maker economics, noting that a market maker's profit comes from the liquidity demand of uninformed traders while its losses come from the informational advantage of informed traders.

In an automated market maker, the impermanent loss that liquidity providers face is essentially the on-chain version of the adverse-selection cost. When an asset's true price moves, arbitrageurs—acting as informed traders—use the pricing formula to trade against the liquidity pool until the price inside the pool aligns with the external market. In this process, arbitrageurs extract value while liquidity providers bear the loss. Milionis et al. (2022) [42] proposed the loss-versus-rebalancing (LVR) model, which rigorously proves that impermanent loss can be quantified as the cost to liquidity providers of being exploited by arbitrageurs for providing stale prices. The model's central conclusion is that the instantaneous LVR—the loss rate per unit time, with dimensions of

value per unit time—is proportional to the square of the underlying asset's price volatility, σ^2, and to the marginal liquidity at the price, $L(P)$; this form closely resembles the gamma-risk structure in Black-Scholes option pricing and applies to all constant function market makers (CFMMs). It should be noted that Black-Scholes assumes continuous hedging and constant volatility, whereas AMM rebalancing is constrained by the discreteness of block time, and under extreme volatility, price jumps within a block interval cause actual LVR losses to deviate substantially from the predictions of the continuous-time model. Whether in a traditional order book or an automated market maker, the underlying mechanism by which liquidity providers are systematically drained of value by the better-informed party is exactly the same; only the outward form changes, from passive fills against active quotes to active arbitrage against passive liquidity. The strategy spaces of the two structures do differ importantly: a traditional order-book market maker can adjust the width and depth of its quotes in real time in response to information events, whereas an AMM liquidity provider is bound by the mathematical structure of the pool and can only passively accept the new asset ratio when the price moves. This asymmetry implies that the adverse-selection cost in an AMM may be systematically higher in magnitude than in an active market-making environment. This isomorphism means that the theoretical tools and practical strategies developed in traditional finance to address adverse selection can, with appropriate adjustment, be applied directly to the risk management of on-chain liquidity provision. Chapter 19 will systematically analyze this continuum from the perspective of the "toxic flow spectrum."

5.5.2 Margin cascades and liquidation cascades

The positive feedback loop of leveraged systems is known in traditional finance as the *price accelerator*. Harris (2003) [3] notes that under a margin regime, a sharp market decline triggers margin calls, and if a trader cannot top up in time, the position is forcibly liquidated, generating fresh selling pressure that causes further declines and more forced liquidations, forming a self-reinforcing positive feedback loop. The 1929 stock market crash is the classic case of this mechanism: the widely used margin loans of the time triggered a wave of large-scale forced liquidations as the market fell, turning a market correction into a systemic collapse. In the nearly one hundred years since, traditional finance has developed a series of buffering mechanisms—including price limits, circuit breakers, and the central counterparty clearing system—to break the positive feedback loop.

The liquidation cascade of perpetual futures markets is the extreme amplification of this old problem in an on-chain environment. Because the market broadly offers extremely high leverage and trades around the clock, chained liquidations can transmit rapidly across platforms via oracles. Ali (2025) [43] conducted a quantitative analysis of the October 2025 crypto market flash crash, finding that under a macroeconomic shock, $19 billion in open interest was wiped out within just 36 hours, the volatility-persistence coefficient was close to 0.90, and the cross-asset contagion intensity was about 20% higher than the spillover effect during the 2018 trade war. Compared with the day-scale transmission of traditional finance, the on-chain version of the liquidation cascade is two orders of magnitude faster ("orders of magnitude" here refers to the cross-market

spillover time window: in traditional finance it is measured in days, whereas on-chain liquidation completes cross-platform transmission at the minute scale through oracles and arbitrage mechanisms) and has a distinctive cross-platform transmission mechanism. Moreover, the transparency of the on-chain environment gives the liquidation cascade a behavioral-contagion dimension that does not exist in traditional finance: large liquidation events are fully visible on-chain, propagate in real time through block explorers and social media, and turn a mechanical liquidation process into a "panic signal" for market participants. Social media compresses the timescale of panic propagation from the hour scale of traditional finance further to the minute scale, and the behavioral-level panic selling and the mechanism-level forced liquidation reinforce each other, potentially making the risk structure of an on-chain liquidation cascade structurally higher than that of a traditional margin cascade. After a large-scale liquidation on one decentralized exchange causes the price to deviate, arbitrageurs quickly transmit the shock to other platforms, triggering broader liquidations. The time buffer that settlement cycles and margin-call notices provide in traditional finance barely exists in an on-chain environment, so the risk-control system must respond within a second-scale time window. Chapter 20 will provide a formal analysis of the full dynamics of the liquidation cascade through the reflexivity triangle model.

5.5.3 Market fragmentation and cross-chain fragmentation

Trading the same asset on multiple exchanges disperses liquidity, increases transaction costs, and makes the price-discovery process more complex. In traditional finance, market fragmentation has long been a core concern of regulators and scholars, its main driver being competition among different trading venues. After the 2005 Regulation National Market System took effect, U.S. equity trading volume was dispersed across more than ten exchanges and dozens of dark pools, and the consolidation of the best bid and offer relied on complex market-data infrastructure.

In on-chain finance, the coexistence of multiple public blockchains has produced cross-chain fragmentation. Liquidity is dispersed across different decentralized exchanges on different chains such as Ethereum, Solana, and Arbitrum, and a user seeking the best price must resort to cross-chain bridges and aggregators, which not only raises transaction costs but also introduces an entirely new security dimension. In a Bank for International Settlements working paper, Shin (2026) [44] points out that the congestion and high-fee structure that public chains create in order to incentivize validators structurally drive user dispersion, making fragmentation an endogenous feature of decentralized economies rather than the result of external competition. An empirical study of Uniswap by Lehar et al. (2023) [45] finds that high-fee pools attract 58% of liquidity supply but execute only 21% of trading volume, indicating that fixed transaction costs produce a systematic stratification of liquidity providers.

Cross-chain fragmentation exposes the system to extremely high security risk. Data from Chainalysis (2022) [46] show that in just the first eight months of 2022, about $2 billion in crypto assets was stolen in 13 separate cross-chain bridge hacks, 69% of the total stolen that year. The cumulative losses from cross-chain bridges have reached the scale of billions of dollars (the DefiLlama data compiled by Chainlink put the figure

at more than $2.5 billion) [11]. The failure of a cross-chain bridge not only causes direct fund losses but can also destroy all downstream derivatives positions collateralized by bridged assets: when bridged WETH or WBTC depegs because the bridge contract is breached, all perpetual futures positions using these assets as margin face simultaneous liquidation, forming "collateral-level contagion." The on-chain version of fragmentation is not merely an economic problem of reduced price-discovery efficiency but a technical problem that threatens the security of the entire ecosystem. Traditional finance's fragmentation can be mitigated through regulatory coordination and market-data consolidation, whereas the solution to on-chain fragmentation must address the dual challenge of economic efficiency and technical security at once.

5.5.4 Closing gaps and low-liquidity amplification

Information accumulated during a traditional financial market's closure is usually released at the open in the form of a price gap. To digest overnight information and smooth the price shock, traditional exchanges introduced mechanisms such as the call auction to provide a price-discovery window before formal continuous trading begins. This discrete closure-open rhythm creates a discontinuity in information release, but it also provides market participants a time window to reassess positions and adjust risk-control parameters.

On-chain finance's around-the-clock trading eliminates the discrete closure period, but liquidity displays marked periodic variation across different time windows. Hansen et al. (2024) [47] find that cryptocurrency volatility and volume exhibit systematic intraday and intraweek patterns, and that these patterns strengthen over time, possibly driven by algorithmic trading and futures-market settlement schedules. During Asian hours or on weekends, market liquidity falls significantly, and the impact of large orders is extremely amplified. Makarov and Schoar (2020) [48] document persistent, sizable arbitrage opportunities across cryptocurrency exchanges, with price deviations especially pronounced across national exchanges and widening further during periods of rapid price increase. Research by Mourey et al. (2025) [49] further confirms a "weekend effect" in crypto markets, noting that negative weekend returns and high volatility can even predict weak Monday equity-market performance, and that this effect became more pronounced after the 2022 LUNA collapse.

Because they lack the closure-period shock absorber of traditional finance, the continuous impact of information during low-liquidity windows replaces the concentrated release at the open, raising the market's intrinsic volatility. This poses distinctive challenges for service providers. The risk-control system must respond in real time while the market is running and cannot, like a traditional futures exchange, perform system maintenance and parameter adjustments after the close. Market makers must bear higher inventory risk and widen their spreads during low-liquidity windows, and any system vulnerability can cause an irreversible loss before the operations team can intervene. Chapter 22 will analyze the implications of 24/7 volatility in depth, and Chapter 23 will examine the distinctive mechanism of volatility clustering in a market without shock absorbers.

5.5.5 The cross-modal risk transmission path

The risk transmission between centralized and decentralized architectures forms a tightly coupled cyclical system whose transmission speed far exceeds the cross-market contagion of traditional finance. This transmission path can be characterized precisely in four steps. In the first step, large-scale liquidation on the centralized-exchange side causes a sharp price drop, and highly leveraged positions are forcibly liquidated, generating concentrated selling pressure. In the second step, the oracle network collects the centralized exchange's price data and passes it to the mark-price system of the on-chain decentralized exchange, so that the on-chain price reflects the centralized market's decline almost synchronously. In the third step, the update of the decentralized exchange's mark price triggers on-chain liquidations, and the liquidation flow hits the liquidity pool or order book, further depressing the on-chain price. In the fourth step, arbitrageurs buy low on the decentralized exchange and sell on the centralized exchange, adding to the centralized exchange's selling pressure and causing the price to fall further, thereby completing one positive-feedback cycle. This theoretical arbitrage loop, however, can be blocked in extreme conditions by a key execution friction: a centralized exchange may freeze withdrawals during a crisis. As an "emergency circuit breaker," a withdrawal freeze has a double-edged effect: while it stops risk from transmitting from the centralized market to the chain, it also stops users from moving assets to the safe option of self-custody, and it can cause a persistent cross-venue price divergence, because arbitrageurs cannot execute the cross-platform convergence trade.

It should be noted that the four-step linear cycle above is a simplified model. Actual cross-modal transmission is a nonlinear network structure containing multiple feedback loops operating in parallel (such as the funding-rate feedback, the liquidity-withdrawal feedback, and the collateral-value feedback). These loops may behave as negative feedback (self-stabilizing) under normal market conditions, but under certain conditions—such as when multiple loops switch to positive feedback simultaneously—they can produce a resonance effect that amplifies systemic risk far beyond the predictions of a linear model.

Moreover, cross-modal transmission is not confined to an internal loop within crypto markets. Macro-financial factors such as the U.S. dollar liquidity cycle constitute an important external trigger for systemic shocks in crypto markets. The macro backdrop of FTX's collapse was the repricing of global risk assets caused by the Federal Reserve's aggressive rate hikes in 2022; a tightening of dollar liquidity transmitted to the crypto-derivatives market through common holders (institutional investors holding both traditional risk assets and crypto assets), triggering deleveraging and selling. This bidirectional macro-crypto transmission means that a systemic-risk assessment of perpetual futures markets cannot be confined to the crypto ecosystem alone but must also incorporate global macro-liquidity conditions.

This transmission process is amplified by three key factors. The immediacy of the oracle keeps prices in the centralized and decentralized markets nearly synchronous, eliminating the cross-market risk buffer that settlement delay provides in traditional finance. Around-the-clock trading means transmission can occur at any time, including the low-liquidity hours of the early morning, making the shock effect more severe. High

leverage ensures that every price drop can trigger fresh liquidations, accelerating the positive feedback loop.

Naifar and Makni (2025) [50] provide empirical evidence of cross-modal contagion in their analysis of the FTX collapse. The study finds that after the extreme event, Bitcoin's outward directional connectedness jumped from 67.00 to 82.21, and the intra-group spillover among DeFi tokens likewise strengthened significantly, with Uniswap's risk reception doubling from 14.89 to 30.92. These data indicate that shocks in the centralized market transmit efficiently to the decentralized ecosystem through oracles and arbitrage mechanisms, and that the two modes are tightly linked at the macro-risk level by the same underlying price. Each platform's risk-control system can set parameters only from a local perspective and cannot assess the systemic risk level of the whole market, and this information asymmetry constitutes a systemic-risk hazard in the current perpetual futures market.

The four isomorphic mappings above share a common feature: the underlying economic mechanism remains unchanged, but the on-chain environment introduces fundamental differences in transmission speed, scale amplification, and security dimension. Adverse selection changes from passive fills against active quotes to active arbitrage against passive liquidity; the margin cascade accelerates from day-scale transmission to the minute scale; market fragmentation expands from an economic-efficiency problem to a technical-security problem; and the closing gap changes from a discrete release at the open to a continuous impact during low-liquidity windows. Table 5-3 lays out these four mappings and their shared underlying mechanisms, and the fourth column marks where each is analyzed in depth in later chapters. As the table shows, the third column of each mapping—the "core isomorphic mechanism"—points to the same underlying logic: whether adverse selection or the margin cascade, the driving mechanisms of information asymmetry and leverage amplification are essentially the same across different technical environments; only the transmission speed and scope of impact change by orders of magnitude because of the characteristics of the on-chain environment.

Old problem in traditional finance	New form in on-chain finance	Core isomorphic mechanism	Deeper analysis in later chapters
Adverse selection: market makers face informed traders	Impermanent loss: AMM liquidity providers face arbitrageurs	Liquidity providers are systematically drained of value by the better-informed party	Chapter 19
Margin cascade: margin calls trigger forced liquidations	Liquidation cascade: high leverage triggers chained liquidations that propagate across platforms via oracles	Self-reinforcing positive feedback in leveraged systems; two orders of magnitude faster on-chain	Chapters 20 and 22
Market fragmentation: the same asset trades on multiple exchanges	Cross-chain fragmentation: liquidity is dispersed across chains and cross-chain bridges are weakly secured	Dispersed liquidity reduces price-discovery efficiency; on-chain adds a security dimension	Chapter 14
Closing gap: information accumulated during the closure is released at the open	Low-liquidity amplification: in 24/7 continuous trading, shocks are amplified in the early morning and on weekends	A timing mismatch between information release and liquidity supply	Chapters 22 and 23

Table 5-3. The isomorphic mapping between traditional finance's old problems and on-chain finance's new forms (Data source: Harris [3], Milionis et al. [42], Ali [43])

Figure 5-13 presents these four isomorphic mappings in a three-column comparison.

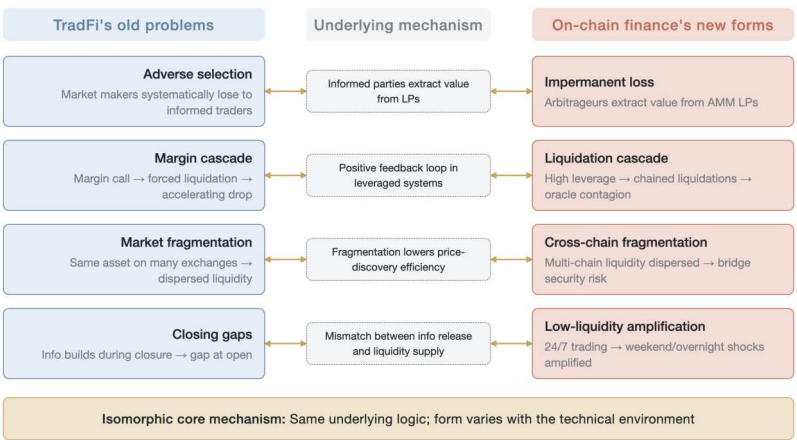


Figure 5-13. The isomorphic mapping between the old problems of traditional finance and their new forms in on-chain finance (conceptual comparison synthesized by the author, not empirical data; Data source: Harris [3])

The four steps of this transmission cycle—liquidation on the centralized exchange triggering a price drop, the oracle transmitting the price on-chain, on-chain liquidation generating fresh selling pressure, and arbitrageurs transmitting the shock back to the

centralized market—form a closed positive-feedback loop, and three factors, the immediacy of the oracle, around-the-clock trading, and high leverage, together amplify the transmission efficiency of this loop (Figure 5-14).

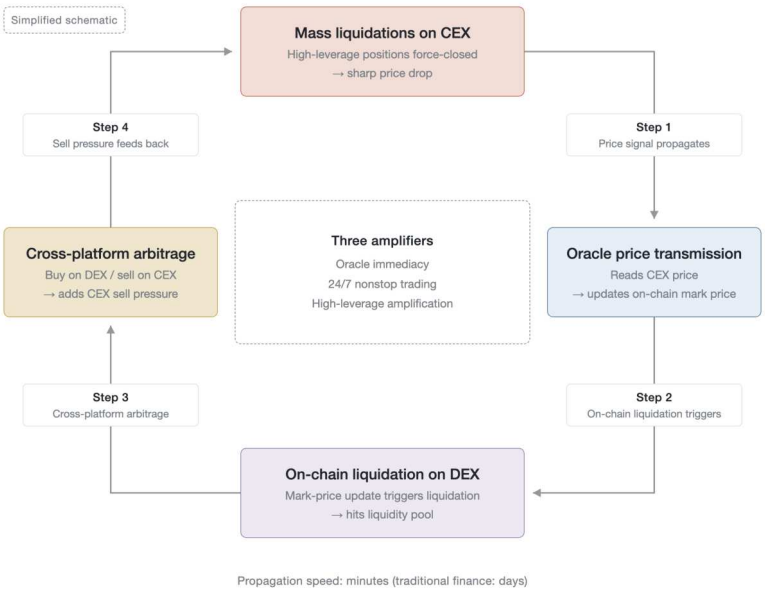


Figure 5-14. The four-step cyclical path of cross-modal risk transmission (a schematic of the cross-modal transmission mechanism, not empirical data; a simplified model—the actual structure is a nonlinear multi-loop network, as detailed in the text; the connectedness values from [50] appear in the text; Data source: Naifar and Makni [50])

5.6 Effects on market microstructure

Differences in service-provider architecture ultimately transmit systematically to the core dimensions of market microstructure. Four dimensions—price-discovery efficiency, the mode of liquidity supply, the structure of transaction costs, and the distribution of systemic risk—together determine the operating quality of a market. This section builds, for each dimension, a causal chain of "architecture choice affects microstructure," identifying the core variables and transmission paths; detailed quantitative analysis will follow in the corresponding later chapters.

5.6.1 The reconstruction of price discovery

Price discovery is the process by which a market aggregates dispersed information about an asset's fundamental value into a single price. In centralized exchanges and on-chain central limit order books, price forms through the bidding contest between buyers and sellers, following the traditional supply-and-demand-driven pattern. Market makers' quoting behavior, the distribution of limit orders, and the impact of market orders together determine the path of price formation. In a decentralized exchange that relies

on oracle pricing, by contrast, the platform outsources pricing entirely, becoming a price taker rather than a price discoverer. Research by Chen et al. (2024) [51] shows that in oracle-based decentralized exchanges, traders tend to react to external price sources rather than contribute new price information through their own trading. This design may cause price-discovery lag, but it also avoids the price distortion that would arise from insufficient in-house liquidity. A comparative study by Barbon and Ranaldo (2024) [52] finds that decentralized exchanges may offer more competitive execution costs on large trades, but the presence of gas fees creates persistent arbitrage opportunities that interfere with the efficiency of price discovery.

In perpetual futures markets, the funding rate constitutes a distinctive dimension of price discovery. The funding rate is not only the mechanism that maintains the anchoring of the contract price to the spot price; it is itself a key indicator of market sentiment and of the balance of long and short forces. Differences in settlement frequency across service providers directly affect the speed of price convergence: centralized exchanges usually settle every 8 hours, whereas some app-chains achieve hourly or even more frequent settlement, and a higher settlement frequency means faster price convergence and smaller basis fluctuations. Specifically, under an 8-hour settlement cycle, the basis can accumulate to a significant level within the settlement interval, offering basis arbitrageurs an ample window to operate; hourly settlement compresses this window, reducing basis fluctuation but also raising the sensitivity of arbitrage strategies to latency; and continuous settlement (such as the per-block settlement achieved by some app-chains) compresses, in theory, the basis-arbitrage space to the block-interval level, rendering traditional funding-rate arbitrage strategies economically more marginal. Differences in the funding rate across platforms create cross-platform arbitrage opportunities that objectively promote the convergence of the global price, but they also expose arbitrageurs to several entirely different risk frameworks simultaneously—the counterparty risk of the centralized exchange, the smart-contract risk of the on-chain protocol, and the cross-platform settlement-timing risk all act on the arbitrageur's position portfolio at once. Chapters 13 through 15 will systematically develop the analysis of price discovery.

5.6.2 The evolution of liquidity supply

The mode of liquidity supply under different architectures reflects a trade-off between barriers to access and professional risk-management capability. A centralized exchange's liquidity is provided mainly by professional market makers, who use high-frequency algorithms and complex inventory-management models to control risk and can provide extremely deep liquidity, but from highly concentrated sources. Harris (2003) [3] defines a market maker as "a trader who provides immediacy to the market," whose profit comes from the bid-ask spread as compensation for the liquidity demand of uninformed traders. In a centralized exchange, a market maker's barriers to entry include technical infrastructure, capital scale, and a cooperative relationship with the exchange, and these barriers concentrate liquidity supply in the hands of a few professional institutions.

Automated market makers lower the barrier to liquidity supply to the minimum, al-

lowing any asset holder to become a liquidity provider. A passive liquidity provider, however, must bear the full adverse-selection cost, and capital efficiency under the constant-product model is generally low, with a large amount of capital allocated to price ranges that will in fact never be used. Although the concentrated-liquidity design raises capital utilization by several orders of magnitude, it essentially turns passive supply into a position that requires active management, placing higher demands on the professional ability of ordinary users. Research by Milionis et al. (2022) [42] shows that in a concentrated-liquidity environment, the behavior pattern of liquidity providers is closer to that of a traditional market maker than to that of a passive asset holder, and their returns depend on the quality of range selection and rebalancing strategy. On-chain order books, in turn, seek a balance between low barriers and professional risk management, usually combining professional market makers with protocol-level market-making vaults to maintain liquidity, so that professional institutions and ordinary users can participate in liquidity supply in different ways within the same market. Chapter 19 will develop the full analysis of market-making economics.

5.6.3 Structural differences in transaction costs

The structure of transaction costs differs fundamentally across architectures and can be analyzed at three levels: explicit costs, implicit costs, and architecture-specific costs. Explicit costs mainly comprise trading fees and network gas fees. Centralized exchanges usually offer the lowest fees and grant tiered discounts to large traders, whereas a decentralized exchange's explicit costs are affected by the congestion of the underlying network, and gas fees on Ethereum mainnet can make small trades uneconomic, although the development of layer-2 networks and app-chains is significantly lowering this cost.

Implicit costs are often more important than explicit costs. Because of concentrated liquidity, a centralized exchange has relatively small slippage and price impact, and large orders can be executed within a narrower price range. In an automated market maker, slippage constitutes the main implicit cost, its size depending on the ratio of trade size to the depth of the liquidity pool. In addition, different architectures have their own distinctive cost types. A centralized exchange operator has a global view of all order flow and could in theory abuse this information; this cost is hard to observe and quantify but can have a systematic effect on users' actual execution quality. On-chain markets, by contrast, face MEV costs. Adams et al. (2024) [53] find that when trading volatile assets, the probability of encountering adversarial slippage is 80% higher than when trading mature assets, and MEV-extraction behavior such as sandwich attacks constitutes a significant implicit cost for on-chain traders. Wu et al. (2025) [54] tracked 19 months of CEX-DEX arbitrage activity and found that just 19 major searchers extracted $233.8 million in value through more than 7.2 million arbitrage events, with the top three searchers accounting for three-quarters of the total value extracted, indicating a highly concentrated trend in MEV extraction. An app-chain architecture, by imposing order-ordering rules at the consensus level, can reduce certain types of MEV from the base layer to some extent—a structural advantage that general-purpose-chain decentralized exchanges struggle to achieve. Although a precise cross-venue cost comparison is limited by data availability and the continual change

in market conditions, order-of-magnitude estimates offer a reference: for a $1 million BTC-PERP trade, the total transaction cost (fees plus slippage) at a leading centralized exchange such as Binance is about 2 to 4 basis points; at an on-chain order book such as Hyperliquid, about 3 to 6 basis points; and the explicit cost (fees plus price-impact fee) of an oracle-pricing model such as GMX, about 6 to 10 basis points, though its zero-slippage feature may make it competitive on very large trades. These values fluctuate with market liquidity and volatility, but the order-of-magnitude differences reflect the structural differences in execution efficiency across architectures.

5.6.4 The distribution of systemic risk

Systemic risk takes markedly different distributional forms across architectures, and this difference directly affects the design of regulatory strategy and risk-management frameworks. The risk of the centralized-exchange model is highly concentrated in a few entities. Dell'Erba (2025) [55] points out that the crypto-derivatives market is replicating the "too big to fail" (TBTF) structure of traditional finance, with a few leading exchanges concentrating the vast majority of trading volume and user assets. In traditional finance, however, the TBTF problem is accompanied by an implicit government guarantee: systemically important institutions can obtain central bank liquidity support and fiscal bailouts in a crisis, whereas centralized entities in crypto markets enjoy no safety net, and FTX's actual collapse proves that they are in essence "too big and will fail." This distinction means that the systemic consequences of concentration in crypto markets may be more severe than in traditional finance. Similar concentration risk is also forming at the on-chain-protocol level: if a leading protocol such as Hyperliquid dominates most of the trading volume of the on-chain perpetual futures market, its failure would not only cause direct user losses but could also destroy the market's trust in the entire on-chain derivatives model, producing an ecosystem-contagion effect far beyond the scope of a single protocol. FTX's collapse is an extreme expression of concentration risk: the failure of a single institution directly caused enormous fund losses and triggered a broad crisis of industry trust. Research by Jalan and Matkovskyy (2023) [56] argues that the FTX episode was in essence a failure of corporate governance and regulation rather than a defect of crypto assets themselves, and that the systemic risk of crypto markets actually peaked in 2021 because of China's regulatory tightening.

The on-chain model disperses risk across multiple independent protocols, and the failure of a single protocol usually does not directly cause the system to collapse. The composability of decentralized finance, however, constitutes an efficient channel for risk contagion. A report by the Financial Stability Board (2023) [40] emphasizes that the mutual nesting and dependence among protocols allow the risk of one protocol to transmit rapidly to the entire ecosystem, its main vulnerabilities being operational fragility, liquidity and maturity mismatch, leverage, and interconnectedness. In the Terra/LUNA episode, the depegging of the algorithmic stablecoin triggered a downward spiral that caused chained liquidations across multiple lending protocols, with a lasting negative impact on the entire decentralized-finance ecosystem. The cross-modal contagion mechanism (as described in Section 5.5.5) further links the centralized and decentralized markets tightly, together forming a complex and fragile risk network. The Financial Stability Board (2023) [40] also notes that although the 2022 crypto-market

turmoil and FTX's collapse had a limited impact on the traditional financial system, spillover effects could strengthen significantly in the future as the ties between the two systems deepen. Chapter 20 will provide a formal analysis of cross-dimensional contagion through the reflexivity triangle model, and Chapter 25 will build a comprehensive evaluation framework for market quality.

5.7 Chapter summary

From traditional finance to centralized exchanges and then to on-chain protocols, the evolution of the service-provider ecosystem is not a linear process of progress but a continuous transformation of the risk surface. Each new architecture, while dissolving the core risks of its predecessor, necessarily introduces a set of new risks of a different nature. Understanding this dynamic is the prerequisite for making a rational service-provider choice.

The five theoretical tools proposed in this chapter provide a systematic analytical framework for this understanding. The risk substitution theorem shows that no zero-risk architecture exists, only a choice among types of risk, and that the criterion for the choice is not which architecture is absolutely safer but which risk is more manageable and more hedgeable for a particular user. The service provider's trilemma reveals the inherent conflict among performance, transparency, and decentralization that cannot all be had at once; any service provider can only choose its own position within the triangle, and although frontier technologies such as zero-knowledge proofs are trying to expand the area of the triangle, they cannot eliminate the constraint of the triangle itself. The concept of the bundling tax reveals the hidden cost of a centralized exchange's efficiency: while enjoying ultra-fast trading and a convenient experience, users pay the price of opacity, conflict of interest, and single points of failure—which forms, with the trust tax that traditional finance pays to achieve functional separation (discussed in Chapter 1), the two ends of the same trade-off. The verifiability ladder makes trust cost concrete: the five-level standard from fully unverifiable to fully verifiable shows that every increase in verifiability comes at the cost of a higher cost of technical understanding. And the old-problems-in-new-forms perspective reminds us that many of on-chain finance's challenges are essentially the technical recurrence of traditional finance's old problems: adverse selection, the margin cascade, and market fragmentation appear in the on-chain environment as impermanent loss, the liquidation cascade, and cross-chain fragmentation, and although their transmission speed and scale change fundamentally, the underlying mechanisms remain isomorphic.

The purpose of these theoretical tools is not to judge which service provider is superior but to give market participants a framework for rational choice grounded in understanding rather than faith. A mature trader should combine multiple service providers according to personal risk preference, technical ability, and trading needs, so as to achieve the best balance of risk and return. For users with lower risk tolerance and limited technical ability, a regulated centralized exchange may be the more reasonable starting point; for users who pursue asset self-sovereignty and possess the ability to understand the technology, on-chain protocols offer a transparency and verifiability that a traditional architecture cannot achieve.

This chapter has conducted a static analysis of the current service-provider ecosystem. One trend worth watching is that the boundary between the centralized and decentralized models is blurring: some centralized exchanges are beginning to migrate the settlement layer on-chain to improve transparency, while some on-chain protocols are moving closer to centralized platforms at the user-experience layer to lower the barrier to use. Chapter 6 will shift from static analysis to dynamic analysis, examining how the three paradigms evolve through competition, how price-discovery power shifts among different architectures, and whether this convergence trend will produce a hybrid model combining the strengths of both or instead create new systemic risks.

References

[1] CoinDesk. (2023, June 26). FTX bankruptcy team says the exchange owed customers $8.7B. CoinDesk. https://www.coindesk.com/policy/2023/06/26/ftx-bankruptcy-team-says-the-exchange-owed-customers-87b

[2] Reuters. (2022, November 17). Bankrupt FTX's new CEO outlines fund abuses, 'complete failure' of corporate control. Reuters. https://www.reuters.com/technology/new-ftx-ceo-slams-complete-failure-corporate-control-2022-11-17/

[3] Harris, L. (2003). Trading and exchanges: Market microstructure for practitioners. Oxford University Press.

[4] CoinGecko. (2025). State of crypto perpetuals market 2024. https://assets.coingecko.com/reports/2025/CoinGecko-State-of-Crypto-Perpetuals-Market.pdf

[5] Paddrik, M., & Zhang, S. (2020). Central counterparty default waterfalls and systemic loss (OFR Working Paper No. 20-04). Office of Financial Research. https://www.financialresearch.gov/working-papers/files/OFRwp-20-04_central-counterparty-default-waterfalls-and-systemic-loss.pdf

[6] Dell'Erba, M. (2024). Crypto-trading platforms as exchanges. Michigan State Law Review, 2024(1). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4405361 https://doi.org/10.2139/ssrn.4405361

[7] Yadav, Y. (2023). The centralization paradox in cryptocurrency markets. Washington University Law Review, 100, 1725. https://wustllawreview.org/2023/08/09/the-centralization-paradox-in-cryptocurrency-markets/

[8] Three Sigma. (2025, January 1). 2024 most exploited DeFi vulnerabilities. Three Sigma. https://threesigma.xyz/blog/exploit/2024-defi-exploits-top-vulnerabilities

[9] Wilson Center. (2025, March 31). The Bybit heist: What happened & what now? Wilson Center. https://www.wilsoncenter.org/article/bybit-heist-what-happened-what-now

[10] European Securities and Markets Authority. (2025, July 1). Maximal extractable value: Implications for crypto markets (ESMA50-481369926-29744). https://www.esma.europa.eu/sites/default/files/2025-07/ESMA50-481369926-29744_Maximal_Extractable_Value_Implications_for_crypto_markets.pdf

[11] Chainlink. (2025, June 6). Seven key cross-chain bridge vulnerabilities explained.

Chainlink Education Hub. https://chain.link/education-hub/cross-chain-bridge-vulnerabilities

[12] The Block. (2025, January 8). Hyperliquid responds to community concerns over validator issues. The Block. https://www.theblock.co/post/333559/hyperliquid-responds-to-community-concerns-over-validator-issues

[13] Buterin, V. (2017, December 31). On sharding blockchains (the scalability trilemma). Sharding FAQ. https://vitalik.eth.limo/general/2017/12/31/sharding_faq.html

[14] Vidal-Tomas, D. (2025). Centralized exchanges & proof-of-solvency: The guardians of trust. Journal of International Financial Markets, Institutions and Money, 103. https://www.sciencedirect.com/science/article/pii/S1042443125000733

[15] CME Group. (2025). CME Group reports record average daily volume for 2024. https://www.cmegroup.com/media-room/press-releases/2026/1/08/cme_group_internationalaveragedailyvolumereachesrecord84millionc.html

[16] Federal Reserve History. (2013). Banking Act of 1933. https://www.federalreservehistory.org/essays/glass-steagall-act

[17] Brown, S. J., & Steenbeek, O. W. (2001). Doubling: Nick Leeson's trading strategy. Pacific-Basin Finance Journal, 9(2), 83-99. https://doi.org/10.1016/s0927-538x(01)00004-x

[18] Congressional Research Service. (2013). The MF Global bankruptcy, missing customer funds, and proposals for reform (Report R42091). https://www.congress.gov/crs-product/R42091

[19] Financial Stability Board. (2022). Derivatives markets and central counterparties. https://www.fsb.org/work-of-the-fsb/market-and-institutional-resilience/derivatives-markets-and-central-counterparties/

[20] Committee on Payment and Settlement Systems & Technical Committee of the International Organization of Securities Commissions. (2012). Principles for financial market infrastructures. Bank for International Settlements. https://www.bis.org/cpmi/info_pfmi.htm

[21] Bell, S., & Holden, H. (2018). Two defaults at CCPs, 10 years apart. BIS Quarterly Review, December 2018, 75-76. https://www.bis.org/publ/qtrpdf/r_qt1812x.htm

[22] Investment Company Institute, Securities Industry and Financial Markets Association, & The Depository Trust & Clearing Corporation. (2024). T+1 after action report. https://www.ici.org/news-release/24-t1-after-action-report-release

[23] The Depository Trust & Clearing Corporation. (2021). DTCC's Project ION platform moves to development phase following successful pilot with industry. https://www.dtcc.com/news/2021/september/15/dtccs-project-ion-platform-moves-to-development-phase-following-successful-pilot-with-industry

[24] European Parliament. (2023). Regulation 2023/1114 on markets in crypto-assets. Official Journal of the European Union.

[25] CoinGecko. (2026, February 26). Is the future of crypto perpetual? The meteoric

rise of perp DEXs. https://www.coingecko.com/learn/rise-of-perpetuals-and-perp-dexs

[26] Binance Square. (2026, March 4). Perpetual DEX 2025-2026: From "CEX alternative" to core infrastructure. https://www.binance.com/en/square/post/297934520960097

[27] Cointelegraph. (2025, February 24). Bybit suffers $1.5 billion hack - the biggest crypto heist in history. https://cointelegraph.com/news/bybit-suffers-1-5-billion-hack-biggest-crypto-heist-history

[28] dYdX Foundation. (2023, November 1). v4 technical architecture overview. https://www.dydx.xyz/blog/v4-technical-architecture-overview

[29] Nansen Research. (2023, March 15). Examination of decentralized perpetual designs. https://research.nansen.ai/articles/examination-of-decentralized-perpetual-designs

[30] Optimism Governance. (2022, July 19). Perpetual Protocol: Optimism governance's first failure. https://gov.optimism.io/t/perpetual-protocol-optimism-governances-first-failure/3029

[31] Cointelegraph. (2022, September 18). Decentralized exchange GMX suffers $565K price manipulation exploit. https://cointelegraph.com/news/decentralized-exchange-gmx-suffers-565k-price-manipulation-exploit

[32] Resonance Security. (2025, July 9). How the GMX exchange was hacked for $42 million [full analysis]. https://www.resonance.security/blog-posts/how-the-gmx-exchange-was-hacked-for-42-million-full-analysis

[33] Presto Research. (2024, November 7). Hyperliquid: The HYPE Begins. https://www.prestolabs.io/research/hyperliquid-the-hype-begins

[34] Yin, M., Malkhi, D., Reiter, M. K., Gueta, G. G., & Abraham, I. (2019). HotStuff: BFT Consensus with Linearity and Responsiveness. Proceedings of the 2019 ACM Symposium on Principles of Distributed Computing, 347-356.

[35] CoinDesk. (2025, January 8). HYPE Slumps 15%, HyperLiquid Responds to Scrutiny Over Lack of Decentralization. CoinDesk. https://www.coindesk.com/business/2025/01/08/hyper-liquid-responds-to-scrutiny-over-lack-of-decentralization-hype-slumps-15

[36] CryptoPotato. (2025, December 22). Hyperliquid Denies $362M Risk Claims, Says Platform Is Fully Solvent. https://cryptopotato.com/hyperliquid-denies-362m-risk-claims-says-platform-is-fully-solvent/

[37] OneKey Blog. (2026, January 26). Hyperliquid Architecture Explained: Security & Decentralization. https://onekey.so/blog/ecosystem/hyperliquid-architecture-explained-security-decentralization-19137c/

[38] CoinDesk. (2025, March 26). HyperLiquid Delists JELLY After Vault Squeezed in $13M Tussle. https://www.coindesk.com/markets/2025/03/26/hyperliquid-delists-jellyjelly-after-vault-squeezed-in-usd13m-tussle

[39] Binance Square. (2025, March 27). Hyperliquid Loses $13.5M in JELLY Exploit:

Key Details and AMM Risks. https://www.binance.com/en/square/post/22139574223130

[40] Financial Stability Board. (2023). The financial stability risks of decentralised finance. https://www.fsb.org/2023/02/the-financial-stability-risks-of-decentralised-finance/

[41] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics, 14(1), 71-100. https://doi.org/10.1016/0304-405x(85)90044-3

[42] Milionis, J., Moallemi, C. C., Roughgarden, T., & Zhang, A. L. (2022). Automated market making and loss-versus-rebalancing. arXiv preprint arXiv:2208.06046. https://doi.org/10.48550/arXiv.2208.06046

[43] Ali, Z. (2025). Anatomy of the Oct 10-11, 2025 crypto liquidation cascade: Macroeconomic triggers, market microstructure, and systemic risk lessons. SSRN. https://ssrn.com/abstract=5611392

[44] Bank for International Settlements. (2026, March 5). Tokenomics and blockchain fragmentation (BIS Working Papers No. 1335). https://www.bis.org/publ/work1335.pdf

[45] Lehar, A., Parlour, C. A., & Zoican, M. (2023). Fragmentation and optimal liquidity supply on decentralized exchanges. arXiv preprint arXiv:2307.13772. https://doi.org/10.48550/arXiv.2307.13772

[46] Chainalysis. (2022). Vulnerabilities in cross-chain bridge protocols emerge as top security risk. https://www.chainalysis.com/blog/cross-chain-bridge-hacks-2022/

[47] Hansen, P. R., Kim, C., & Kimbrough, W. (2024). Periodicity in cryptocurrency volatility and liquidity. Journal of Financial Econometrics, 22(1), 224-251. https://doi.org/10.1093/jjfinec/nbac034

[48] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. Journal of Financial Economics, 135(2), 293-319. https://doi.org/10.1016/j.jfineco.2019.07.001

[49] Mourey, M., Shahrour, M. H., & Soiman, F. (2025). A crypto-stock weekend effect: Predicting Monday stock returns using weekend cryptocurrency returns. SSRN. https://ssrn.com/abstract=5382090 https://doi.org/10.2139/ssrn.5382090

[50] Naifar, N., & Makni, M. S. (2025). Dynamics of cryptocurrencies, defi tokens, and tech stocks: Lessons from the FTX collapse. International Journal of Financial Studies, 13(3), 169. https://doi.org/10.3390/ijfs13030169

[51] Chen, E., Ma, M., & Nie, Z. (2024). Exploring the impact: How decentralized exchange designs shape traders' behavior on perpetual future contracts. arXiv preprint arXiv:2402.03953. https://doi.org/10.48550/arXiv.2402.03953

[52] Barbon, A., & Ranaldo, A. (2024). On the quality of cryptocurrency markets: Centralized versus decentralized exchanges. arXiv preprint arXiv:2112.07386. https://doi.org/10.48550/arXiv.2112.07386

[53] Adams, A., Chan, B. Y., Markovich, S., & Wan, X. (2024). Don't let MEV

slip: The costs of swapping on the Uniswap protocol. arXiv preprint arXiv:2309.13648. https://doi.org/10.48550/arXiv.2309.13648

[54] Wu, F., Sui, D., Thiery, T., & Pai, M. (2025). Measuring CEX-DEX extracted value and searcher profitability: The darkest of the MEV dark forest. arXiv preprint arXiv:2507.13023. https://doi.org/10.48550/arXiv.2507.13023

[55] Dell'Erba, M. (2025). Crypto-derivatives. University of Illinois Law Review, 2025(5), 1613.

[56] Jalan, A., & Matkovskyy, R. (2023). Systemic risks in the cryptocurrency market: Evidence from the FTX collapse. Finance Research Letters, 53, 103670. https://doi.org/10.2139/ssrn.4364121

The Evolution and Trade-offs of Perpetual Futures Market Structure

A single 10x-leveraged long position in ETH perpetual futures traverses markedly different execution paths on a centralized exchange (CEX), a general-purpose public-chain decentralized exchange (DEX), and an application-chain DEX. On a centralized exchange, the order is matched within 500 milliseconds at fees as low as 0.02%, yet the margin is held in the exchange's centralized wallet; what the user holds is merely a claim recorded in a database, the order-sequencing process is entirely opaque to outsiders, and the 2022 collapse of FTX—which exposed a customer-fund shortfall of roughly $8.7 billion—revealed the systemic risk inherent in this custodial structure [1]. On a general-purpose public-chain DEX built on Ethereum, the user retains sovereignty over assets through a self-custodial wallet, but trading intent is exposed in the public mempool, where a sandwich attack can convert 0.3% of slippage into profit for maximal extractable value (MEV) searchers [2], so the price of asset sovereignty is an implicit *transparency tax*. On an application-chain DEX, the user obtains sub-second matching and near-zero slippage while retaining on-chain self-custody, yet the underlying network may be maintained by only a dozen or so validator nodes, and the risk of validator collusion becomes a new trust assumption.

These three execution paths epitomize the structural transformation now under way in crypto derivatives markets, which grew rapidly in 2025. CEX perpetual futures volume reached approximately $86.2 trillion, up 47.4% year over year [3]. The more striking change, however, occurred in decentralized markets: DEX perpetual futures volume grew 346% year over year in 2025, reaching approximately $6.7 trillion [3]. The ratio of DEX to CEX perpetual futures volume climbed from 2.1% in early 2023 to 11.7% by November 2025—roughly one DEX trade for every nine CEX trades at that ratio [4]. Within this shift, application-chain DEXs led by Hyperliquid at one point captured roughly 60% of the on-chain perpetual futures market, with annualized volume rivaling that of established CEXs such as Coinbase [5].

These figures reflect a deep transformation in the market's fundamental structure. The migration of volume from CEXs to DEXs is not merely a reallocation of market share; it signals that the market's basic rules are being redefined. How prices form, how participants compete, how risk is distributed, and how fairness is defined are all undergoing

fundamental change. Drawing on the market-microstructure methodology of Harris (2003) [6], this chapter builds an analytical framework that dissects how each structure shapes the economic activity and power relations within it, rather than simply judging which is superior.

This chapter argues that the evolution of market structure is not a linear replacement of CEXs by DEXs but a paradigm shift in the mode of competition. The object of competition migrates from *latency* to *information* and then to *mechanism*, and each migration redefines who wins and who loses. Beginning with the micro-level trade lifecycle, the chapter advances a competition-mode migration thesis, examines the rise of strategic MEV in depth, explores the price-discovery migration hypothesis, analyzes the evolution of risk distribution, and finally considers the convergence and endgame of the three paradigms.

6.1 The trade lifecycle

The execution path of a trade—from order submission through matching to final settlement—is far more than a technical distinction. It fundamentally determines how information flows, how power is allocated, and how costs arise. Chapter 5 performed a *static dissection* of the internal components and risks of the three architectures from the service provider's perspective; this section shifts the vantage point, tracing a trade's complete journey through the three architectures from the trade's own point of view. This complementary perspective reveals the sharply different trade-offs that centralized exchanges, general-purpose public-chain DEXs, and application-chain DEXs make among performance, transparency, security, and cost, providing a solid factual basis for the competition-mode analysis in Section 6.2.

Figure 6-1 compares the essential differences among the three paradigms across four stages: order submission, matching, settlement, and asset custody.

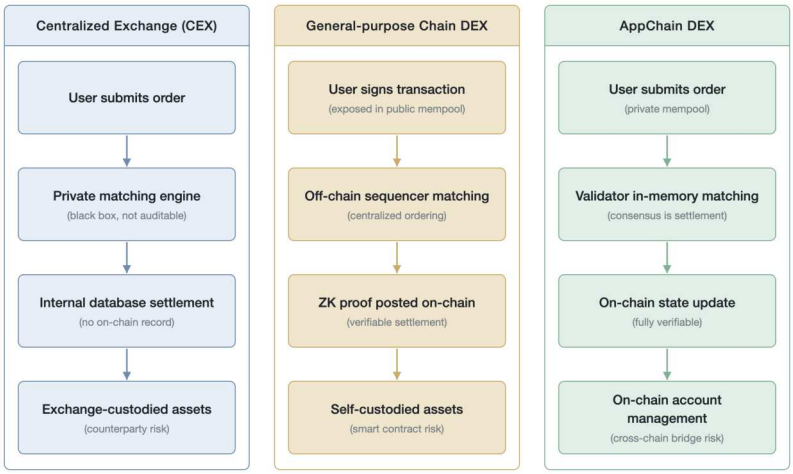


Figure 6-1. Comparison of the trade execution flow across the three perpetual futures market-structure paradigms (a conceptual process comparison, not empirical data; drawn by the author, based on Harris [6])

6.1.1 Centralized exchanges

A centralized exchange is operated by a single entity whose private, centralized servers maintain the order book, match trades, and custody user assets. A user's trading records are merely entries in an internal database, interacting with the blockchain only at deposit and withdrawal. The core advantage of this architecture is high performance: the internal latency of top-tier matching engines has reached the sub-millisecond and even sub-100-microsecond range, while end-to-end response latency from API submission to confirmation is typically between 1 and 10 milliseconds, with throughput exceeding a million orders per second—a level no on-chain solution can currently match.

Order submission stage. A user submits a limit order or market order to the exchange's private servers through an API or web front end. From that moment, the trade enters an execution environment entirely opaque to the outside world. The user cannot verify which internal systems the order passes through before matching, cannot confirm whether sequencing strictly follows price-time priority, and cannot know whether the exchange itself or its affiliates enjoy some privilege in the order flow. This opacity is the structural root of information asymmetry in CEX markets. The exchange holds real-time order-flow data for all users, data of extremely high value for anticipating short-term price movements. This informational advantage creates structural room for potential unfairness. The flow of benefits between FTX and its affiliated trading firm, Alameda Research, is an extreme manifestation of this structural risk [1].

Matching stage. The high-performance matching engine scans the order book at microsecond speed, pairing buy orders with market makers' sell orders under price-time priority. It is here that the core game of the CEX paradigm—the adverse-selection game—unfolds. Market makers sustain market liquidity by quoting on both the bid and ask sides, earning their profit from the bid-ask spread. Their greatest risk, however, is being exploited by better-informed traders: when the market is about to move directionally, informed traders quickly hit the stale quotes that market makers have not yet updated, inflicting losses on the market makers.

Facing adverse-selection risk, market makers have two basic responses. The first is to widen the bid-ask spread, compensating for expected adverse-selection losses with a higher margin, though this reduces the competitiveness of their quotes. The second strategy has come to dominate in practice: increase speed, updating or withdrawing stale quotes faster than informed traders can act. This logic of *speed as survival* has driven a continuous arms race around infrastructure latency. Institutions with colocation services place their servers in the same physical space as the exchange's data center to minimize network latency; some deploy custom hardware such as field-programmable gate arrays (FPGAs) to accelerate trading logic at the physical layer, or use microwave communication networks in pursuit of a microsecond-level speed edge. This latency race, which has run for decades in traditional financial markets, is now replicated in full in crypto CEXs. As a result, market-making profits are highly concentrated among a handful of institutions with top-tier infrastructure, such as Wintermute, Jump Trading, and Cumberland.

The CEX fee structure further reinforces this competitive landscape. The mainstream maker-taker model rebates fees to makers who provide liquidity and charges higher

fees to takers who consume it. Institutional market makers, as the primary makers, enjoy fee discounts or even negative fees, whereas retail traders, as takers, pay higher transaction costs. Seen this way, the fee structure is itself an institutionalized cost-transfer mechanism, shifting the implicit cost of sustaining liquidity from market makers onto ordinary traders.

Settlement stage. Once matching succeeds, settlement completes instantly on the exchange's internal ledger. This is essentially just an update of database account balances, involving no blockchain operation. Although extremely fast and cheap, what the user holds is not a genuine on-chain asset but a liability the exchange owes the user. The safety of assets depends entirely on the exchange's solvency and operational integrity. The 2022 collapse of FTX exposed a customer-fund shortfall of roughly $8.7 billion (per the debtors' estimate and the August 2024 court restitution order) [1][53], and the $1.5 billion security incident suffered by Bybit in February 2025 [7] both plainly reveal the systemic risk embedded in this centralized custodial structure. Notably, these two events belong to entirely different categories of operational risk: FTX was a governance and fraud failure, whereas the Bybit incident was a supply-chain attack in which the attacker—attributed to the Lazarus Group—manipulated the signing-ceremony interface of the Safe{Wallet} front end to induce multisig holders to sign a malicious transaction. This UI-layer attack vector is not unique to CEXs; any system relying on multisig approval—including DEX treasury management, cross-chain-bridge operator multisigs, and application-chain governance wallets—faces the same class of risk. In enjoying peak performance, the user in effect entrusts the entire security of their assets to an unauditable third party.

Risk management. The liquidation engine monitors the margin level of every account in real time. When the mark price (typically computed as the index price across multiple exchanges plus a funding-rate basis smoothed by an exponential moving average, or EMA) drives an account's margin ratio below the maintenance-margin requirement, the system automatically triggers forced liquidation. This mark-price methodology differs fundamentally from the oracle-based mark price used in DEXs, a design difference analyzed further in the discussion of oracle manipulation in Section 6.3.2. If a shortfall arises during liquidation—that is, if the liquidation proceeds are insufficient to cover the loss—the loss is borne by the exchange's insurance fund. Should the insurance fund be exhausted in extreme conditions, the remaining loss is allocated by rule to traders on the profitable side through an auto-deleveraging mechanism. The entire risk-control process, from margin calculation through liquidation execution to the use of the insurance fund, runs inside a *black box* that outsiders cannot independently audit. Users can neither verify the fairness of liquidation prices nor confirm the true size and use of the insurance fund.

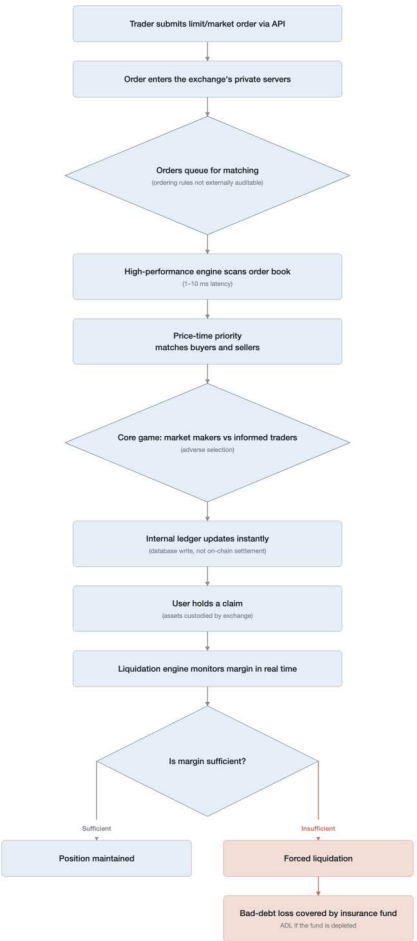


Figure 6-2. The trade lifecycle on a centralized exchange (from order submission to settlement and risk control, completed entirely within the exchange's private servers; a mechanism-process schematic in which the 1-to-10-millisecond figure refers to end-to-end response latency, not empirical data; drawn by the author, based on Harris [6])

6.1.2 General-purpose public-chain DEXs

A general-purpose public-chain DEX deploys its core settlement logic on a public blockchain while executing some functions—such as order-book maintenance and matching—off-chain. This hybrid model of *off-chain matching, on-chain settlement* seeks a balance between the security of the blockchain and the performance of trading. Representative general-purpose public-chain DEXs today include GMX v2 and Vertex on Arbitrum. Notably, the early architecture typified by dYdX v3 (built on StarkEx) has undergone a major transformation: dYdX v4 migrated to a standalone application chain based on the Cosmos SDK at the end of 2023 and, under this chapter's taxonomy, now falls within the application-chain DEX category—a migration that is itself strong evidence for the competition-mode migration thesis in Section 6.2. Its typical

architecture is based on ZK-Rollup technology, with an off-chain sequencer handling high-speed matching and on-chain smart contracts handling final state verification and asset custody.

Order submission stage. A user cryptographically signs the trade with a self-custodial wallet. Before the trading intent is finally confirmed and packed into a block, it is typically exposed in the public mempool. The public mempool is the cornerstone of blockchain transparency: it ensures that all pending transactions are visible to every participant in the network, safeguarding the system's auditability. Yet this same transparency makes pending transactions highly vulnerable to MEV extraction. MEV searchers continuously scan the mempool with specialized algorithms, and once they spot a large order, they extract value from it through a variety of strategies.

The most common forms of MEV attack include the following. Front-running: after detecting a large buy order, the searcher buys ahead of it at a lower price and sells for a profit once the large order pushes the price up. The sandwich attack: the searcher inserts a buy and a sell transaction, respectively, before and after the target transaction, capturing a risk-free profit from the target's price impact. And liquidation sniping: the searcher monitors the margin status of on-chain positions and, when liquidation conditions are about to trigger, front-runs the liquidation to capture the reward the protocol offers. The common feature of these attacks is that they all rely on the trading-intent information exposed by the public mempool.

The MEV value chain and the multi-tier auction. The competitive relationship in a general-purpose public-chain DEX is fundamentally different from that in a CEX. In a CEX, the core game plays out between market makers and informed traders and centers on the asymmetry of price information. In a general-purpose public-chain DEX, the core game plays out between active MEV searchers and passive ordinary traders and centers on the speed and strategic depth with which public information is exploited.

MEV extraction forms a layered value chain. Ordinary traders are the source of value; their trading intent is exposed in the mempool. MEV searchers capture arbitrage opportunities through algorithmic scanning and send carefully constructed *arbitrage bundles*, together with a bribe fee, to block builders. Block builders select the profit-maximizing combination from the bundles submitted by many searchers, assemble a complete block, and submit it to validators. Validators select the highest-bidding block to propose and confirm. The total cost of this multi-tier auction is ultimately borne by ordinary traders, while the profit is distributed among searchers, builders, and validators according to the equilibrium of the game.

In a thematic report released in July 2025, the European Securities and Markets Authority (ESMA) noted that MEV searchers typically pay the vast majority of their revenue—roughly 85% to 93%—to block proposers, making the right to order transactions an extremely scarce resource [2]. ESMA further noted that MEV poses a potential threat to market fairness and investor protection, marking the problem's formal entry into the field of view of global financial regulators. This judgment bears a deep structural resemblance to traditional financial markets' regulatory concerns about high-frequency trading; both involve the systematic exploitation of an informational advantage and the fairness issues that result.

Matching and settlement stage. Orders are sent to an off-chain sequencer for matching. The sequencer packages a batch of matched trades, generates a zero-knowledge proof (such as a ZK-STARK), and submits it to a smart contract on the Ethereum mainnet for verification. The core advantage of this design is that it guarantees the correctness of settlement at the mathematical level: even if the sequencer tries to forge trade results or tamper with account balances, the on-chain zero-knowledge-proof verification mechanism can detect and reject the invalid state transition. Users need not trust the sequencer's honesty—only the correctness of the mathematics.

Nevertheless, the centralization of the sequencer itself remains a risk worth attention. A sequencer outage means the entire exchange halts, and users cannot execute any trades or adjust positions during that time. On top of this, the sequencer could in theory selectively delay or reject transactions, though such behavior can be detected after the fact. To mitigate this risk, most Layer-2 designs include a *forced-exit* mechanism that lets users withdraw assets directly through the mainnet contract when the sequencer fails, ensuring assets cannot be permanently locked.

Risk management. In sharp contrast to the *black-box* risk control of a CEX, the liquidation logic of a general-purpose public-chain DEX is hard-coded into on-chain smart contracts. Anyone can monitor the margin status of on-chain accounts and, when liquidation conditions trigger, call the contract function to perform the liquidation and capture the reward the protocol offers. This permissionless, open liquidation market greatly enhances the system's overall stability, because liquidation no longer depends on the risk-control capacity of a single entity but is maintained jointly by participants across the entire network. Yet this mechanism brings both benefit and risk: in extreme conditions, all participants can see in real time which positions are about to be liquidated, which can produce a *liquidation cascade*—a large number of liquidations triggering at once, further depressing the price, triggering still more liquidations, and forming a vicious cycle. In addition, code vulnerabilities in the smart contracts themselves are a risk source that cannot be ignored; once a contract is attacked, losses can be catastrophic and hard to reverse.

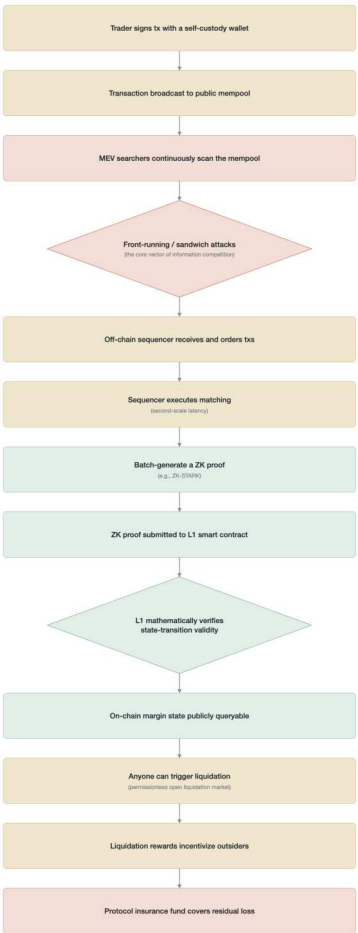


Figure 6-3. The hybrid execution path of a general-purpose public-chain DEX (after the trade is exposed in the public mempool, it is matched by an off-chain sequencer and finally settled on the mainnet via a ZK proof; a conceptual process schematic, not empirical data; drawn by the author, based on ESMA [2] and Harris [6])

6.1.3 Application-chain DEXs

The application-chain DEX represents a third direction in the evolution of market structure. It builds the entire exchange—order book, matching engine, settlement, and risk management—as an independent, sovereign blockchain deeply optimized for trading. For application-chain protocols typified by dYdX Chain (built on the Cosmos SDK) and Hyperliquid, the core idea is this: the architecture of a general-purpose public chain cannot provide sufficient performance or room for customization for high-frequency trading, so a dedicated chain must be purpose-built for the exchange [5].

Full-stack control and extreme performance optimization. Unlike a general-purpose public-chain DEX, which deploys smart contracts only at the application layer, an application-chain team holds full-stack control, spanning the application layer (ex-

change logic), the consensus layer (the validator network), the network layer (node communication), and the data layer (state storage). This full-stack control creates room for high-performance optimization. In a traditional blockchain architecture, a transaction must pass through several steps—mempool broadcast, block packing, and consensus confirmation—before it can finalize. In an application chain, these steps are deeply integrated and optimized: the order book is maintained directly in validators' memory, and matching logic is embedded in the consensus process itself. Take Hyperliquid's HyperBFT consensus protocol: matching and confirmation complete in the same step, achieving sub-second confirmation latency and giving users an experience close to that of a CEX. The operational risk of Byzantine fault-tolerant (BFT) consensus, however, cannot be ignored: when the leader node goes offline, the protocol must perform a view change (leader rotation), during which no blocks are produced and no transactions are confirmed. For a perpetual futures exchange, a consensus interruption of several seconds amid volatile conditions can prevent users from adjusting margin or closing positions, thereby triggering passive liquidation.

Order submission and matching stage. A user's signed order is sent directly to validator nodes and enters a private mempool. In sharp contrast to the public mempool of a general-purpose public-chain DEX, trading intent in a private mempool is not broadcast network-wide and is visible only to the validator currently responsible for producing the block. This design structurally and substantially weakens external searchers' sandwich attacks and front-running, because searchers can hardly scan pending transactions, making these traditional forms of MEV technically very difficult. The private mempool does not eliminate MEV, however; it concentrates MEV-extraction capability, moving it from a group of openly competing searchers to the leader validator currently responsible for producing the block, creating a monopolistic window for MEV extraction. In the absence of verifiable ordering rules—such as first-in-first-out (FIFO) ordering enforced by a ZK proof—a validator can freely order transactions during its leader round, and the covert nature of this MEV behavior makes it harder to detect and quantify than MEV in a public mempool.

In an application chain, validators play a dual role: they are participants in consensus (voting to confirm blocks) and operators of the matching engine (maintaining the order book in memory and executing matching). After orders are matched at high speed in a validator's memory, the result is proposed to other validators as part of a block for consensus confirmation. Here, *consensus is settlement*: once a block receives confirming signatures from enough validators, the trade completes matching and settlement simultaneously, with no need to wait for the additional on-chain verification step required by a general-purpose public-chain DEX.

New trust assumptions and validator MEV. Even as the application-chain architecture eliminates one class of risk, it introduces a new trust assumption. Because validators control the right to order transactions, they can in theory perform self-preferential ordering—for example, inserting their own buy order ahead of a user's large buy order, or delaying the processing of certain transactions. This *validator MEV* is more covert than MEV in a public mempool, because it occurs inside a validator's internal systems and is hard for outsiders to detect and quantify in real time.

The degree of decentralization of the validator network and the cost of misbehaving therefore become key variables for the system's fairness. If validators are numerous enough and broadly distributed geographically and in their interests, the payoff to a single validator from misbehaving will be far smaller than the cost of having its staked tokens slashed, forming an effective economic deterrent. But some application chains had very few validators in their early stages. Hyperliquid had only 16 validators early on [5]. Although Ethereum has roughly 1 million validator nodes, large staking pools such as Lido control about 28% of the staked share, so the number of effective independent operators is far below the nominal node count. A more appropriate comparative metric is the Nakamoto coefficient—the minimum number of independent entities needed to control 33% or 51% of the staking weight—rather than the raw node count. Even so, Hyperliquid's validator set remains far inferior to Ethereum's on this metric. This limited validator set makes the possibility of node collusion a hidden danger that cannot be ignored. Although the staking-slashing mechanism provides economic-level security, if a small number of validators control the majority of the network's staking weight, the potential payoff from collusion may exceed the cost of slashing, weakening the system's security assumptions.

Risk management and on-chain governance. An application chain's core risk-control mechanisms—liquidation parameters (such as the maintenance-margin ratio and the liquidation-penalty rate) and the rules governing use of the insurance fund—are determined by on-chain governance. This means all rules are public and transparent, and any participant can know in advance under what conditions which risk-control action will trigger. This predictability is a major improvement over the *black-box* risk control of a CEX.

Governance itself, however, is also a complex arena of play. Large token holders can, through their votes, profoundly influence protocol parameters that bear on the interests of all traders. For example, adjusting the liquidation-penalty rate directly affects the incentives of liquidators and the losses of those being liquidated; adjusting the insurance fund's replenishment mechanism affects the profit distribution of all LPs. Where governance power is highly concentrated, the interests of a few large holders may take priority over those of the community as a whole. In addition, depositing assets into an application chain usually requires a cross-chain bridge, and the security of cross-chain bridges has historically been among the more fragile links in DeFi. According to public tallies from the DeFiLlama and Rekt databases (data as of 2025, with figures varying by how cross-chain bridges are defined and how incidents are classified), cross-chain-bridge attacks caused more than $2.5 billion in asset losses between 2021 and 2025 (the larger individual incidents include Ronin at roughly $625 million, Wormhole at roughly $320 million, and Nomad at roughly $190 million) [8]; and for an application-chain DEX, nearly all user funds must be deposited through the bridge, making the bridge the single largest attack surface. Before a user's assets enter the application chain, they must bear the smart-contract risk and relay-node risk of the bridging process.

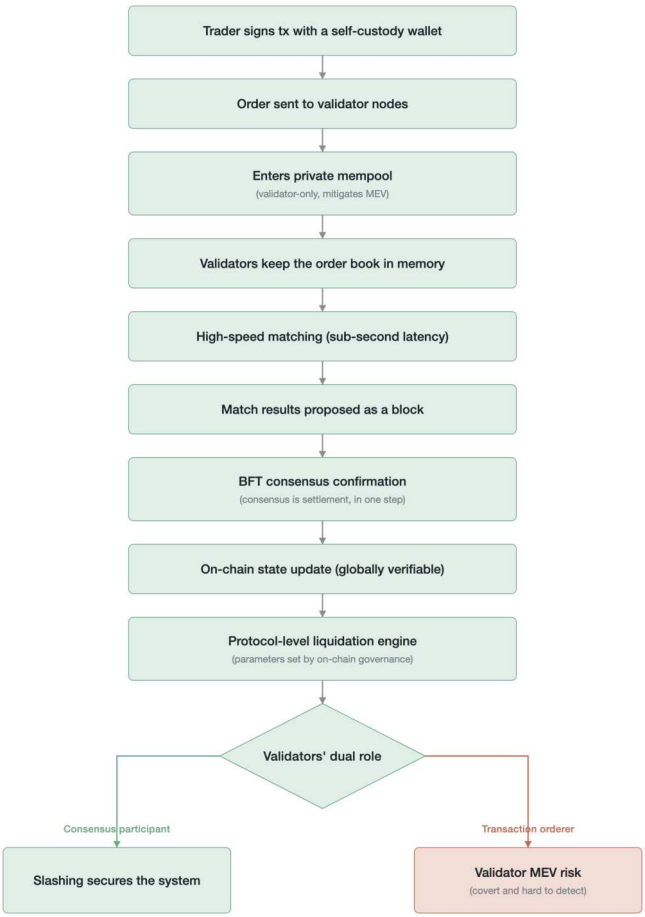


Figure 6-4. The integrated trade flow of an application-chain DEX (orders enter validator memory for matching through a private mempool, and consensus and settlement complete in the same step; a conceptual process schematic, not empirical data; drawn by the author, based on Coinage Media [5] and Eco.com [9])

6.1.4 Structural comparison

Having traced the complete trade lifecycle, the table below systematically compares the essential differences of the three paradigms at their key junctures along eight dimensions: order visibility, matching mechanism, settlement path, asset custody, the core game, opacity risk, risk-control transparency, and the liquidation mechanism. These differences are not a simple ranking of *which is better* but structural trade-offs made under different constraints.

Dimension	Centralized exchange	General-purpose public-chain DEX	Application-chain DEX
Order visibility	Visible only to the exchange; inside a *black box*	Present in the public mempool; globally visible	Present in the private mempool; visible only to validators
Matching mechanism	Centralized high-performance engine (internal sub-millisecond; end-to-end 1–10 ms)	Off-chain sequencer matching (sub-second); settlement finality depends on L1 (seconds to minutes)	In-memory matching by validators (sub-second latency)
Settlement path	Internal database-ledger update; no on-chain operation	ZK proof submitted to the mainnet for verification	Consensus and settlement merged into one
Asset custody	Centralized exchange custody; the user holds a claim	Self-custodial wallet; managed by smart contracts	On-chain self-custody; depositing relies on a cross-chain bridge
Core game	Market makers vs. informed traders (adverse selection)	MEV searchers vs. ordinary users (information exploitation)	Mechanism exploiters vs. protocol LPs (the mechanism game)
Primary opacity risk	The exchange uses its informational advantage for self-preferential treatment	MEV extraction is public, but ordinary users cannot escape it	Validator MEV is covert and hard for outsiders to detect
Risk-control transparency	Fully opaque; not externally auditable	Smart contracts are public; anyone can verify	On-chain governance sets parameters; rules are transparent
Liquidation mechanism	Executed internally by the exchange; the insurance fund is opaque	Permissionless, open liquidation market	Protocol-level liquidation engine; parameters are adjustable

Table 6-1. Structural comparison of the trade lifecycle across the three paradigms (Data source: compiled by the author)

The comparison in Table 6-1 reveals a deep structural insight: each of the three paradigms selects a different resting point within the *impossible triangle* of performance, transparency, and decentralization. The centralized exchange chooses high performance, sacrificing transparency and control over assets, and its core game centers on latency. The general-purpose public-chain DEX chooses absolute transparency and decentralization, accepting a compromise in performance and the affliction of MEV, and its core game centers on information exploitation. The application-chain DEX tries to find an optimum between performance and transparency through a customized architecture, but it introduces new trust assumptions, and its core game centers on understanding and exploiting protocol mechanisms.

Market structure also systematically shapes trader behavior at the micro level. For example, on oracle-priced DEXs, a mechanically lower slippage and adverse-selection risk on the long side [11] has given rise to a systematic long bias [10], showing that different pricing mechanisms, fee structures, and risk-management rules regularly incentivize or suppress specific types of trading behavior—a shaping effect detailed in Section 6.4.4.

Each structural choice creates a different form of competition and gives rise to a different

dimension of *unfairness*. In a CEX, unfairness manifests in the capital threshold of hardware and in the exchange's informational privilege; in a general-purpose public-chain DEX, it manifests in the systematic extraction of value from ordinary users by MEV searchers; in an application-chain DEX, it manifests in validators' implicit informational advantage and the possibility that protocol mechanisms are strategically exploited. Section 6.2 systematically analyzes how these forms of competition migrate as market structure evolves, and what that migration means for market participants and regulators.

6.2 The competition-mode migration thesis

In Chapter 5, we performed a static dissection of the three service-provider architectures, analyzing how their internal components operate and how risk is distributed in each. Markets, however, are not static structures but arenas of intense, real-time competition. Different market structures create not only different risks but also fundamentally different modes of competition. Across this evolution, the focus of competition, its weapons, its battlegrounds, and the very definition of *fairness* all undergo a profound migration.

This section advances the *competition-mode migration thesis*, arguing that the evolution of market structure is not a simple, linear replacement of centralized exchanges by decentralized exchanges but a paradigm shift in the focus of competition. In this shift, the object of the game migrates from traditional *latency* (whose hardware is faster) to *information* in general-purpose public-chain DEXs (who can exploit transparent data faster) and finally evolves, in application-chain DEXs, into a contest over *mechanism* (who understands the protocol's rules more deeply). Each migration redefines who wins, who loses, and what fairness means.

6.2.1 Latency competition in CEXs

As described in Section 6.1.1, the core game market makers face in a CEX is adverse selection [11][12]; their main response is to increase speed rather than widen spreads, which drives an arms race around microsecond-level latency waged with colocation and custom hardware as weapons [6].

Budish, Cramton, and Shim (2015) argued at the theoretical level that the design of the continuous limit order book inevitably produces mechanical arbitrage opportunities; competition does not eliminate these opportunities but merely raises the speed threshold needed to capture them [13]. The empirical data of Aquilina, Budish, and O'Neill (2022) further show that the most common latency-arbitrage races in traditional markets last only 5 to 10 microseconds [14]. In digital-asset markets, high-frequency trading firms likewise pursue double-digit-microsecond *tick-to-trade* performance [15]. The pattern of latency competition that ran for decades in traditional finance is replicated in full in crypto CEXs [16].

The consequences of latency competition are multidimensional. First, it greatly raises the barrier to market entry, and market-making profits are highly concentrated among a handful of institutions with top-tier infrastructure. Second, as described in Sec-

tion 6.1.1, the maker-taker fee model, as an institutionalized cost-transfer mechanism, further entrenches the advantage of leading institutions [17]. Finally, the CEX architecture inherently contains structural information asymmetry: the exchange holds the order-flow data of all users, while the entire matching process runs in a *black box* that outsiders cannot audit; the flow of benefits between FTX and Alameda Research is precisely an extreme manifestation of this structural unfairness [1]. Under the CEX competition paradigm, the exchange's own informational advantage is an absolute structural unfairness that transcends hardware capital.

6.2.2 Information competition in general-purpose public-chain DEXs

As DeFi matured, DEXs on general-purpose public chains set out to eliminate the CEX *black-box* problem through thoroughgoing transparency. On DEXs built on public chains such as Ethereum, every trading intent is exposed in the public mempool before it is finally confirmed and packed into a block. This fundamental architectural change migrated the focus of competition from *latency* to *information exploitation*. The transparency paradox discussed in Chapter 1 is fully displayed here: transparency removes the privilege of centralized institutions, but it also enables technically capable participants to extract value systematically from ordinary users.

At this stage, MEV became the primary vehicle of information competition. The public mempool makes all pending transactions visible to everyone, so competition is no longer about *whose hardware is faster* but about *who can analyze and exploit this public information faster and more precisely*. The three basic forms of MEV (front-running, the sandwich attack, and liquidation sniping) and their layered value chain and multi-tier auction structure were detailed in Section 6.1.2 and are not repeated here; the total cost of this multi-tier auction is ultimately borne by ordinary traders, while the profit is distributed among searchers, builders, and validators.

MEV has a broad impact on the integrity of DeFi markets, and its negative externalities have drawn close attention from regulators [2]. The actual data show that the scale of this information competition is enormous. For example, daily MEV extraction on the Ethereum mainnet has long stayed on the order of hundreds of thousands of dollars; and over the 30 days from late 2025 to early 2026, MEV profit extracted on Ethereum reached nearly $24 million [18]. In addition, sandwich attacks targeting private transaction pools occur frequently and cause significant user losses [19].

Under the information-competition mode, the market's definition of *fairness* changed. The threshold of competition shifted from hardware-capital investment to algorithmic and technical capability. MEV searchers' profit comes from their ability to process public information faster and more intelligently than ordinary users. Although technically anyone can run a searcher node and compete, in practice this capability is likewise highly concentrated among a handful of institutions with top-tier algorithm teams.

Everyone can see, but only the fastest and smartest can exploit—this is precisely the economic embodiment of the transparency paradox. While enjoying self-custody of assets and on-chain verifiability, ordinary users are forced to pay an implicit *transparency tax*. The MEV figures above (over $500,000 per day on Ethereum) reflect total MEV extraction across the entire Ethereum ecosystem, spanning spot DEX swaps, lending

liquidations, and many other scenarios. The effective *transparency tax* that perpetual futures traders actually bear must be assessed independently and varies significantly across different Layer-2 environments.

6.2.3 Mechanism competition in application-chain DEXs

To solve the performance bottleneck and severe MEV problems of general-purpose public chains, the application-chain architecture emerged. Application-chain DEXs typified by Hyperliquid deeply integrate order-book maintenance, the matching engine, and the consensus mechanism by building an independent blockchain purpose-optimized for trading. More importantly, they typically adopt a private-mempool design, so that a user's trading intent is not exposed on the public network before it is formally packed into a block.

This architectural change effectively suppresses traditional atomic MEV (such as front-running and sandwich attacks), because searchers lose the ability to observe pending transactions in advance. Competition, however, does not disappear; it evolves toward a deeper and more covert level. Notably, although the mempool is private, an application chain's API exposes real-time order-book state at low latency, so sophisticated traders who colocate servers near validators can conduct *statistical MEV*—informed trading based on observable order-flow patterns. This to some extent reintroduces, at the application-chain level, a shadow of latency competition. On this basis, a deeper *mechanism competition* emerges. In this mode, the focus of competition is no longer hardware latency, nor is it simply scanning the mempool, but rather a deep understanding and strategic exploitation of the protocol's own economic mechanisms. This is precisely the concretization, at the market-structure level, of the *mechanism parasitism* phenomenon defined in Chapter 4.

In mechanism competition, attackers or advanced traders no longer seek to extract tiny value at the millisecond scale or within a single block; instead, they turn to complex, cross-cycle, cross-market games. This section formally defines this as *strategic MEV*. Unlike atomic MEV, which relies on reordering transactions, the core features of strategic MEV are its non-atomicity, its cross-cycle nature, and its compound exploitation of the protocol's inherent economic vulnerabilities (such as liquidation thresholds, insurance-fund design, and leverage limits).

A series of events on Hyperliquid in March 2025 provides a textbook empirical case of this new mode of competition. On March 12, a whale trader known as *0xf3F* built a long position of more than $300 million in ETH with 50x leverage on Hyperliquid. When the position generated a large unrealized profit, the trader did not choose to close it on the open market but instead began strategically withdrawing margin from the account. This maneuver artificially raised the position's liquidation price until it finally triggered the protocol's forced liquidation [20].

The logic behind this *strategic self-liquidation* is a precise economic calculation. Fifty-times leverage corresponds to a 2% initial-margin ratio, meaning a $300 million notional position requires only about $6 million of initial margin. The trader realized that closing such a huge position directly on the order book would incur enormous slippage losses; estimated from the position's size and the order-book depth at the time, the market-

impact cost of a normal close was far higher than the liquidation penalty the protocol specified. In this way, the trader successfully locked in more than $1.86 million in profit while transferring the enormous market-impact cost of closing the position onto the protocol's liquidity providers. In Hyperliquid's liquidation cascade, a liquidated position is first taken on by backstop liquidators attempting to close it on the open market; if that fails within the time window, the position is transferred to the Hyperliquidity Provider (HLP) vault as the taker of last resort. In this incident, because the position was too large, the intermediate liquidation mechanisms failed to absorb it effectively, and the HLP vault was forced to take on the entire remaining position as the de facto central counterparty, ultimately bearing roughly $4 million in slippage losses [20].

Just two weeks later, on March 26, Hyperliquid suffered the mechanically more complex *JELLY attack*—three coordinated attackers used mutually hedged, huge long and short positions to induce forced liquidation, forcing the HLP vault to take on a low-liquidity position and ultimately eliciting the team's highly controversial centralized delisting and forced-liquidation intervention (the case is detailed in Section 6.3.2) [20].

These events profoundly reveal the essence of mechanism competition: the *weapon* of competition has shifted from engineering (code and hardware) to economics and game theory. Every on-chain operation by the attacker complies with the protocol's rules, and the profit comes from understanding the economic implications of those rules far more deeply than the protocol's designers do. This is a paradigm shift from *who is faster* to *who thinks more deeply.* In the application-chain era, the definition of *fairness* evolves further: the cognitive threshold becomes the new barrier to entry, and protocol liquidity providers and ordinary users who do not understand the deeper implications of the mechanisms become the ultimate bearers of this *mechanism tax.*

6.2.4 A unified comparison of the competition modes

To understand more clearly the essential differences among these three competition modes, we can construct a systematic comparison matrix. As described in Section 6.1.4 (and in Chapter 5), the three paradigms each select a different trade-off within the *impossible triangle.* The table below systematically compares the three competition modes along six dimensions: the focus of competition, core weapons, principal participants, the battleground, the definition of fairness, and the manifestation of unfairness.

Dimension	CEX latency competition	General-purpose public-chain DEX information competition	Application-chain DEX mechanism competition
Focus of competition	Hardware latency (microsecond matching speed)	Information exploitation (trading intent in the public mempool)	Mechanism understanding (strategic exploitation of the protocol's economic rules)
Core weapons	Colocation, custom FPGA hardware, microwave communication networks	MEV search algorithms, block-building optimization, gas-bidding strategies	Cross-cycle position building, exploitation of the liquidation mechanism, cross-market manipulation
Principal participants	High-frequency market makers (Wintermute, Jump Trading, etc.)	MEV searchers, block builders	Strategic traders, protocol-mechanism researchers
Battleground	Physical distance inside the exchange's data center	The public mempool and block-space auctions	The protocol's liquidation engine, insurance fund, and oracle mechanism
Definition of fairness	Capital threshold: who can invest more hardware capital	Technical threshold: who has superior algorithmic capability	Cognitive threshold: who understands protocol economics more deeply
Manifestation of unfairness	Hardware-capital barriers plus the exchange's informational privilege	Systematic value extraction from ordinary users by MEV searchers	Validators' implicit ordering power plus the transfer of losses to protocol LPs by strategic MEV

Table 6-2. A unified comparison matrix of the three competition modes (Data source: compiled by the author)

As Table 6-2 shows, the focus of competition—from physical latency in the CEX, to information exploitation in the general-purpose public-chain DEX, to mechanism understanding in the application-chain DEX—displays a progressive abstraction from hardware engineering to algorithmic technology and then to economic cognition. Corresponding to this is a migration in the definition of fairness: unfairness in the CEX is determined mainly by hardware-capital investment, in the general-purpose public-chain DEX by algorithmic capability, and in the application-chain DEX by the depth of one's understanding of protocol economics. The migration of the competition mode not only changes the rules of the game but also profoundly reshapes the role and situation of each class of market participant. A role that occupies a central position in the CEX paradigm may be marginalized in the on-chain paradigm, and vice versa. From the perspective of six core participants—retail traders, institutional market makers, high-frequency traders, MEV searchers, validators/nodes, and protocol governors—Table 6-3 analyzes the changing opportunities, challenges, and game positions each role faces across the three paradigms.

Participant role	CEX paradigm (latency competition)	General-purpose public-chain DEX paradigm (information competition)	Application-chain DEX paradigm (mechanism competition)
Retail traders	Enjoy the best experience and lowest latency but bear very high custodial risk and the information-asymmetry risk of the black box, passively accepting market makers' spreads.	Gain self-custody of assets and global transparency but must bear the cost of MEV extraction (the transparency tax), gas-fee volatility, and smart-contract risk.	Gain a near-CEX experience and the advantage of self-custody but face cross-chain-bridge risk. MEV extraction is partly mitigated by protocol design, but they may bear an implicit mechanism tax.
Institutional market makers	Hold a central position, earning spreads by providing liquidity. Must invest heavily in low-latency infrastructure to counter adverse selection and enjoy fee subsidies.	Role is weakened (replaced by algorithms under the AMM model). Under the order-book model, they face intense competition from MEV searchers and must adapt to high on-chain latency.	Role revives. They return to the center under a high-performance order book, may gain an informational advantage by running validator nodes, and face competition from protocol-native market-making mechanisms.
High-frequency traders	A traditional area of advantage; they gain a millisecond edge through hardware and network investment and execute latency-arbitrage and statistical-arbitrage strategies.	Advantage is greatly diminished; on-chain second-level latency renders traditional HFT strategies ineffective, but some can transform into professional MEV searchers.	Becomes a new arena. Sub-second latency makes some strategies viable again, and running a validator node may become the *colocation* of the application-chain era.
MEV searchers	Do not exist; information asymmetry and ordering privilege are sealed inside the exchange's black box.	A core role; they continuously scan the public mempool and drive a complex multi-tier auction market.	Role is constrained; the private mempool limits traditional scanning strategies. But strategic MEV targeting protocol-mechanism vulnerabilities becomes a new, higher-order form of play.
Validators/nodes	Not applicable (centralized servers).	An indirect role; as consensus maintainers of the underlying chain, they capture most of the risk-free MEV revenue through mechanisms such as proposer-builder separation.	A core role, simultaneously bearing the triple duties of consensus confirmation, order matching, and transaction ordering. They hold the ultimate informational advantage but face the constraint of staking slashing.
Protocol governors	Not applicable; decision-making power rests entirely with corporate management.	Participate through the DAO and governance tokens, usually facing the dilemma of low retail participation and decision-making dominated by large holders.	On-chain governance is decisive. Validators and large token holders wield significant influence over core mechanisms such as liquidation parameters, determining the protocol's economic security.

Table 6-3. The game matrix of six core participants across the three paradigms (Data source: constructed by the author)

Table 6-3 reveals a noteworthy pattern in the evolution of roles: institutional market

makers hold a central position in the CEX, are weakened in the general-purpose public-chain DEX, and regain a central position in the application-chain DEX; MEV searchers, in turn, go from nonexistent in the CEX to a core role in the general-purpose public-chain DEX, and then to being structurally constrained in the application-chain DEX while pivoting to higher-order strategic MEV. Accompanying this, the transmission path of information asymmetry also evolves from the closed form of the CEX, through the transparent form of the general-purpose public-chain DEX, to the semi-transparent form of the application-chain DEX (see Figure 6-5).

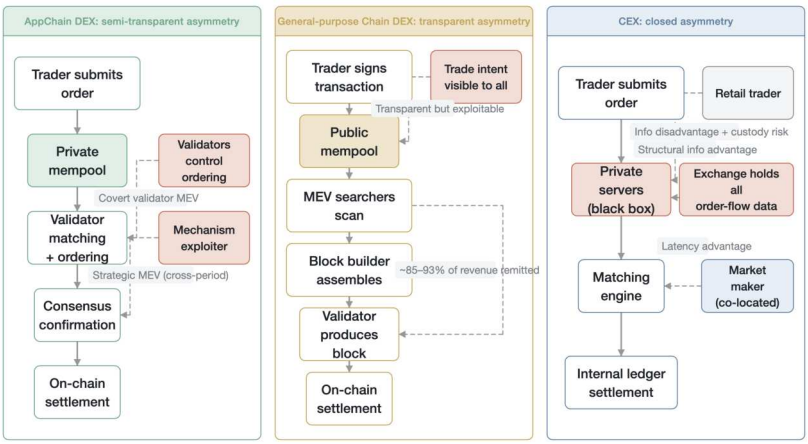


Figure 6-5. The transmission path of information asymmetry across the three paradigms (from the CEX *black box* to the general-purpose public chain's *public mempool* and then to the application chain's *private mempool plus validator matching*, the way an informational advantage is obtained changes fundamentally; a conceptual mechanism schematic, not empirical data; the roughly 85%–93% of revenue remitted follows ESMA [2]; drawn by the author, additionally based on Harris [6])

6.2.5 The drivers of competition-mode migration

Figure 6-6 decomposes the three migrations of the competition mode into three stages and labels the core driver of each: the diminishing returns of latency competition, the revaluation of transparency after the FTX collapse, and the application chain's recombination of performance and transparency within the *impossible triangle*.

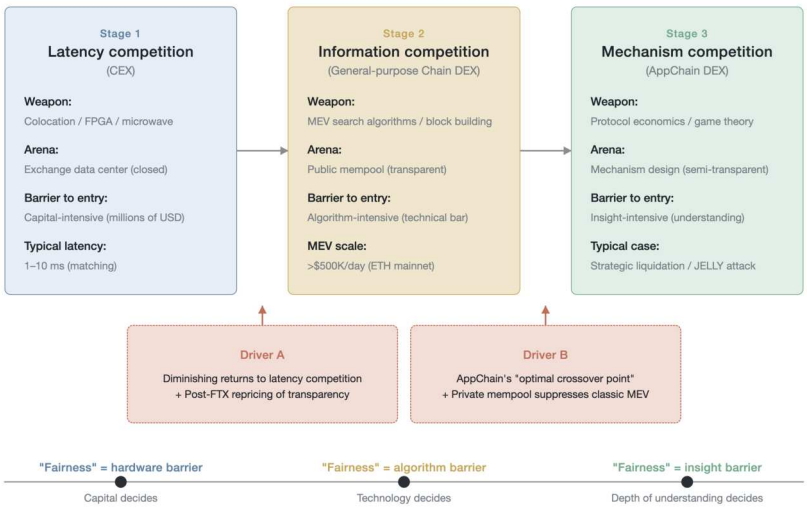


Figure 6-6. The three-stage migration of the competition mode (a mechanism-evolution schematic, not empirical data; the daily MEV figure of over $500,000 follows EigenPhi [18], on the late-2025-to-early-2026 basis; drawn by the author, based on Budish, Cramton, and Shim [13] and ESMA [2])

As Figure 6-6 shows, each migration corresponds to the economic return of the prior competition mode approaching zero: the diminishing marginal returns of latency competition in CEXs pushed capital toward on-chain information competition, while the high cost and poor user experience of MEV extraction on general-purpose public chains in turn drove the market toward the mechanism competition of application chains.

The starting point of the migration is the diminishing returns of latency competition. In CEXs, the arms race around microsecond latency has approached physical limits. To gain even a one-microsecond speed advantage, an institution must invest exponentially growing capital (for example, laying a straighter microwave link or developing more advanced FPGA chips), yet the excess profit such marginal investment can bring is rapidly approaching zero. This has become a zero-sum arms race in which participants keep investing merely to maintain their existing competitive position rather than to create new economic value [13]. When traditional hardware competition ceases to be economically attractive, capital and technical talent naturally turn to areas of competition with higher returns.

Even as the returns to latency competition diminished, the 2022 collapse of FTX triggered a fundamental reappraisal of transparency. After this event, market participants' valuation of *verifiability* rose sharply. Institutional investors and advanced retail traders came to realize that the high efficiency of CEXs was built on one-sided trust and an extremely high single-point-of-failure risk. They became increasingly willing to pay a certain efficiency cost for the ability to *independently verify asset safety and trade execution*. This fundamental shift in risk preference became a powerful underlying driver pushing capital and volume to migrate from CEXs to on-chain models. Behavioral finance, however, offers a competing explanation: availability bias and narrative contagion may lead participants to systematically overestimate the probability of sim-

ilar events after a salient catastrophe, causing an excessive flight from all custodial structures. Lo's (2004) adaptive markets hypothesis [21] predicts that such behavioral overreaction may partially reverse as the FTX event fades from collective memory. This means that in assessing the migration trend, one must distinguish its structural component (a genuine rise in demand for verifiability) from its behavioral amplification component (a panic-driven narrative shift), the durability of the latter being uncertain.

Nevertheless, although general-purpose public-chain DEXs offer transparency, their high gas fees, second-level confirmation latency, and large-scale MEV extraction make them ill-suited to carry high-frequency perpetual futures trading. It is against this backdrop that application chains try to find a more commercially viable optimal intersection within the *impossible triangle*. Through a customized consensus mechanism and in-memory matching, application chains achieve sub-second latency, meeting the performance needs of the vast majority (roughly 95%) of trading strategies; through a private mempool, they structurally suppress the worst atomic MEV; and at the same time they still preserve the baseline of self-custody of assets and on-chain verifiability of core logic. This *performance plus transparency* combination explains why an application chain like Hyperliquid could grow rapidly in 2025 and at one point capture roughly 60% of the on-chain perpetual futures market.

In sum, the migration of the competition mode from latency to information and then to mechanism is an evolutionary process driven jointly by technical bottlenecks, risk events, and economic incentives. In the sections that follow, we examine the highest expression of this mechanism competition—strategic MEV—and analyze how the dominance of price discovery shifts over the course of this evolution.

6.3 Strategic MEV

MEV has been a core issue throughout the development of decentralized finance. Early MEV mainly took the form of technical arbitrage, such as sandwich attacks based on transaction ordering or simple cross-DEX arbitrage. As complex financial derivatives such as perpetual futures have become widespread on-chain, however, the form of MEV is undergoing a profound evolution. This section elevates *strategic MEV* from an observation in the competition-mode discussion of Section 6.2 to a formally defined conceptual category; it systematically argues how strategic MEV differs in nature from traditional atomic MEV, analyzes its three typical forms in the perpetual futures setting, and explores paths for defending against it at the mechanism-design level and for institutionalizing MEV. This concept is the concretization, at the market-structure level, of the *mechanism parasitism* theory of Chapter 4.

6.3.1 Atomic MEV and strategic MEV

To understand the essence of strategic MEV, one must be clear about how it differs from traditional atomic MEV. This difference lies not only in the mode of execution but, more deeply, in the economic logic and risk characteristics behind it. The difference between the two is a leap in kind, not a gradual change in degree.

Traditional atomic MEV relies mainly on a structural feature of blockchains' decen-

tralized architecture: the absence of a unified, tamper-proof transaction timestamp. "Decentralization means there is no pipeline that creates a single queue, nor any natural time-based ordering of transactions as in traditional finance" [2]. On smart-contract platforms such as Ethereum, miners or validators (and, with the introduction of MEV-Boost, block builders) hold the privilege of deciding the order of transactions within a block. Searchers monitor the public mempool, identify profitable trading intent, and compete for block space by submitting transaction bundles that include a higher priority fee. The core feature of this MEV is its *atomicity*: the entire operation (such as the front transaction, the target transaction, and the back transaction) is executed within a single atomic transaction in the same block. If any link fails, the entire transaction is rolled back, and the extractor bears no execution risk, paying only a small gas fee [2][22].

By contrast, strategic MEV is no longer confined to micro-level transaction ordering within a single block; it extends to the macro level of economic gamesmanship. Extractors of strategic MEV exploit not only information in the mempool but also, more deeply, the structural design features of DeFi protocols (especially perpetual futures protocols) in their margin models, liquidation engines, or oracle mechanisms. This MEV is often a continuous operation spanning multiple blocks or even several days, requiring the extractor to commit substantial capital and bear real market-execution risk [23]. This book gives the following formal definition of strategic MEV:

> Strategic MEV: Unlike atomic MEV, which relies on reordering transactions within a single block, the core features of strategic MEV are its non-atomic, cross-cycle, cross-domain, compound attacks. It involves an attacker, over a longer time horizon, exploiting a protocol's inherent economic mechanisms (leverage, liquidation thresholds, insurance-fund design) and market structure (order-book transparency, cross-market liquidity imbalances) to construct, through a series of seemingly independent operations, a final high-value profit or loss-avoidance scenario. The goal of strategic MEV is typically not to extract value from a single transaction but to actively plan and trigger a larger market event, or to externalize a personally incurred loss onto the entire protocol system.

The two differ in nature across several key dimensions. In terms of the information source and profit basis, the profit of atomic MEV comes from the pending transactions of other users in the public mempool; in essence, it exploits users' slippage tolerance. The profit of strategic MEV, however, comes from the protocol's systemic design—especially the insurance fund or the liquidity providers' capital pool—as the extractor, by actively planning a market event (such as manipulating the price of a low-liquidity asset), forces the protocol to execute operations under rules unfavorable to it, thereby achieving a value transfer. In the conceptual framework of Chapter 4, atomic MEV corresponds to the behavioral pattern of the *economic bot*, whereas strategic MEV is the market-structure version of *mechanism parasitism*: the attacker does not violate the rules but understands the economic implications of the rules more deeply than their designers do.

This difference extends further to the dimensions of risk characteristics and capital requirements. Atomic MEV is risk-free arbitrage; the extractor uses the atomicity of smart contracts to guarantee an *all-or-nothing* execution result. Strategic MEV, by

contrast, is a risky investment with directional exposure; the extractor must build a real position, faces the market risk of price movements, and often needs to deploy millions or even hundreds of millions of dollars to drive the market price or exhaust the protocol's liquidity. Correspondingly, the party that bears the loss changes fundamentally: the victim of atomic MEV is usually a single attacked trader, whose loss is local and quantifiable; the loss caused by strategic MEV is often socialized, with the protocol's insurance-fund loss ultimately borne by all liquidity providers and possibly even spreading, through the auto-deleveraging mechanism, to all profitable traders. This socialization of losses makes strategic MEV far more destructive to the protocol ecosystem than atomic MEV.

A direct consequence of these differences is that the suite of technical defensive tools already developed against atomic MEV—such as private transaction pools (Flashbots Protect), the proposer-builder-separation (PBS) mechanism, and order-flow auctions—is nearly useless against strategic MEV [2]. These mechanisms protect ordinary users by hiding trading intent or redistributing MEV profit, but strategic MEV exploits the protocol's economic rules themselves rather than information asymmetry or transaction ordering. Defending against strategic MEV must return to the starting point of mechanism design—achieved by refining margin requirements, optimizing liquidation logic, and strengthening system resilience.

Dimension	Atomic MEV (technical arbitrage)	Strategic MEV (economic game)
Information source	Public mempool pending transactions	Protocol mechanism design structural flaws
Execution horizon	Within one block (seconds)	Across blocks / days (hours to days)
Capital requirement	No large capital needed (atomic arbitrage)	Millions to hundreds of millions in real capital
Risk profile	No execution risk (atomicity guarantee)	Bears execution risk (non-atomic)
Who bears the loss	The single targeted trader	Protocol insurance fund LPs / all participants
Defense	Private order flow PBS / MEV-Share	Mechanism redesign makes exploit infeasible

Qualitative comparison, not quantitative (cells encode categorical differences, not magnitudes)

Figure 6-7. A multidimensional comparison of atomic MEV and strategic MEV (a conceptual comparison matrix along qualitative dimensions, with no empirical data; Data source: ESMA [2]; Daian et al. [22]; Gramlich et al. [23])

As Figure 6-7 shows, the two types of MEV display a leap in kind across six dimensions—information source, execution cycle, capital requirement, risk characteristics, the party bearing the loss, and defensive means (the six-dimensional comparison is detailed in the figure). Between the two, *multi-block MEV* is drawing researchers' attention as an intermediate form: block builders may coordinate strategies across consecutive blocks, and under the PBS framework, where the top two or three builders produce over 90% of

Ethereum's blocks, this concentration may give rise to new strategic vectors unforeseen by traditional PBS design. In short, this leap from *risk-free technical arbitrage* to *risky economic gamesmanship* means that existing technical-layer defensive tools are nearly useless against strategic MEV, and the focus of defense must shift to the protocol's mechanism-design layer.

6.3.2 Typical forms

Because of their high-leverage nature and complex liquidation mechanisms, perpetual futures markets have become a high-risk domain for strategic MEV. The presence of leverage multiplies the precision and destructive power of an attack: under 50x leverage, the initial-margin ratio is only 2%, meaning an adverse price move on the order of 2% is enough to bring a position close to liquidation, which creates an extremely low cost threshold for a carefully planned attack. Several leading on-chain perpetual futures platforms have recently suffered such attacks, exposing deep vulnerabilities in protocol design. In the perpetual futures setting, strategic MEV mainly takes three typical forms. The three forms below each exploit, in different ways, a compound of the economic vulnerabilities described earlier (liquidation thresholds, insurance-fund design, and leverage limits); the relationship is many-to-many rather than one-to-one.

In *strategic self-liquidation*, the attacker exploits the passive-takeover mechanism of the protocol's liquidation engine, transferring the market-impact cost of closing its own huge position onto the protocol's insurance fund or liquidity providers. Its economic logic can be stated succinctly: when the slippage cost of a normal close exceeds the penalty incurred from being liquidated, a rational attacker will choose to *be liquidated* rather than *close the position itself*, thereby externalizing the difference onto the protocol system.

In March 2025, the *JELLY token attack* suffered by the decentralized derivatives exchange Hyperliquid was a textbook case of this technique. In this incident, the attacker used three coordinated accounts to carry out a complex market manipulation. One account (0xde9) built a short position of roughly $4.1 million in the JELLY token on Hyperliquid, while two other accounts (0x20e and 0x67f) built a total of roughly $4.05 million in long positions [20]. Through this *hedged* operation, the attacker controlled both the long and short sides while keeping a market-neutral exposure. Subsequently, as the price of the JELLY token swung violently (partly triggered by the attacker's own spot operations), the margin ratio of the short position fell rapidly, triggering forced liquidation.

Because of the design of Hyperliquid's liquidation mechanism, when a user's position is liquidated, the protocol's HLP vault is forced to take on these huge positions as the central counterparty. JELLY, however, is a low-liquidity meme coin, and its market depth simply could not absorb such a huge closing order. When the HLP vault tried to close these short positions on the market, it produced extremely severe slippage; the peak unrealized loss during the incident once exceeded $10 million, and the widely reported vault loss was roughly $13.5 million (a peak-unrealized-loss figure, not a realized loss; the HLP vault recorded a small profit of about $700,000 after the incident was finally settled) [37]. Facing potential insolvency, the HLP vault simultaneously encountered

panic withdrawals by LP participants, a chain of behavior that exhibited classic bank-run dynamics, further intensifying the liquidity crisis and very likely constituting the direct trigger for the team's forced resort to centralized intervention. The Hyperliquid team ultimately took the highly controversial step of centralized intervention, forcibly closing the position at $0.0095 and delisting the token, containing the loss within a certain range; but the incident starkly exposed the systemic risk of the liquidation engine acting as a passive taker [24]. "The same group of people build and control the protocol, operate the HLP pool, issue the HYPE token, and run the validator nodes; for the sake of the network, they decided to delist JELLY and make everyone whole except the attacker"—this incident reveals the structural contradiction that a decentralized exchange, in extreme circumstances, must resort to centralized intervention [24].

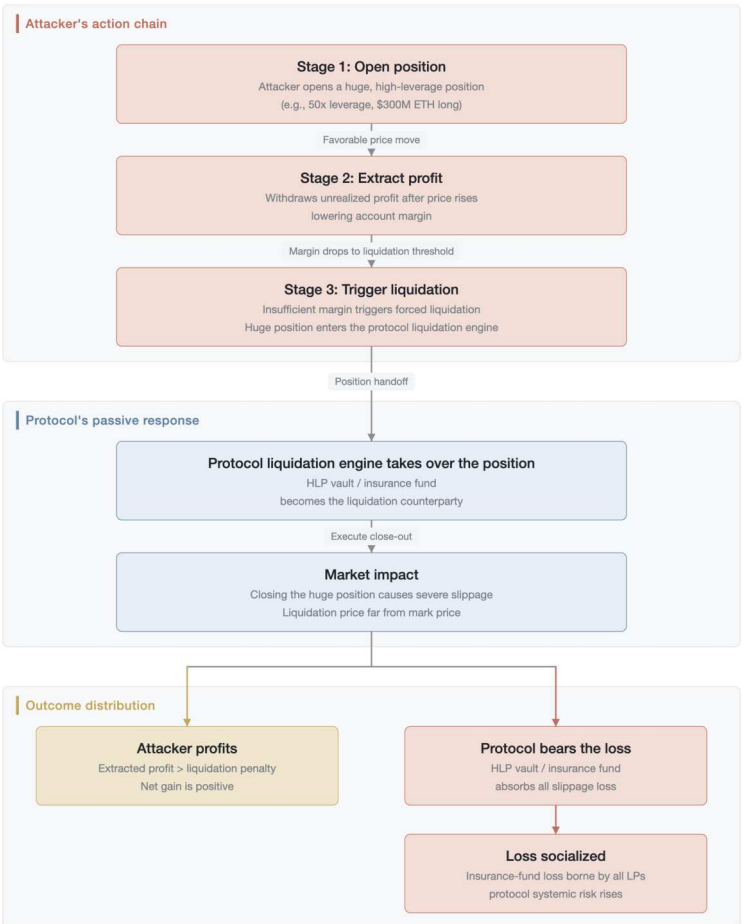


Figure 6-8. The self-liquidation attack flow and value-transfer mechanism in strategic MEV (a mechanism-process schematic, not empirical data; the illustrative $300 million ETH position at 50x leverage refers to the March 2025 whale self-liquidation incident [20]; drawn by the author, based on Arkham Intelligence [20] and Talos [24])

As Figure 6-8 shows, the core logic of the self-liquidation attack is this: the attacker withdraws unrealized profit when the price moves favorably, deliberately renders the ac-

count under-margined, and forces the protocol's insurance fund to take on the position under extremely unfavorable conditions, thereby socializing private risk.

The second form of strategic MEV is *cross-market liquidation manipulation*, which can be seen as an upgraded version of the self-liquidation attack, with the attacker extending the scope of operations from a single platform to multiple markets. The core strategy is this: build a huge position on a perpetual futures platform, then manipulate the underlying asset's price on the spot market (usually a low-liquidity DEX), thereby influencing the oracle price feed on which the perpetual futures platform depends, and finally triggering the liquidation of one's own or others' positions for profit.

The economic essence of this attack is cross-market arbitrage: as long as the cost of manipulating the spot market is lower than the profit obtained from perpetual futures liquidation, the attack is profitable. Because a perpetual futures protocol's oracle typically relies on price data from on-chain DEXs, whose liquidity is often far lower than that of centralized exchanges, the manipulation cost can be kept relatively low. A case in point is the market-manipulation attack Hyperliquid suffered on the POPCAT and HYPE tokens in November 2025: the attacker used only about 3 million USDC to build a buy wall on the spot market, which triggered a chain of liquidations on the perpetual futures market and cost the protocol vault nearly $4.9 million [25]. This kind of cross-market attack means the risk-control measures of a single platform cannot cover the entire attack surface, highlighting the severity of cross-protocol risk transmission in on-chain financial systems.

Compared with the first two forms, *insurance-fund arbitrage* is the most covert variant of strategic MEV. The attacker carefully constructs a fully hedged portfolio of positions and exploits logical flaws in the protocol's liquidation mechanism to trigger *pseudo-liquidations* that defraud the protocol of liquidation rewards, thereby systematically extracting value from the insurance fund. Unlike the first two forms, insurance-fund arbitrage often does not trigger violent market swings; instead, it siphons off protocol funds continuously and gradually. Reportedly, the same November 2025 manipulation of the POPCAT and HYPE tokens on Hyperliquid also exhibited the characteristics of insurance-fund arbitrage: through the interplay of hedged positions and the liquidation mechanism, the attacker caused the protocol a loss of roughly $4.9 million [25] (this is a different facet of the same incident described above under cross-market liquidation manipulation, not a second, independent loss), and the protocol became aware of the anomaly only after the incident had persisted for some time.

The three forms above share a profound common feature: every on-chain operation by the attacker complies with the protocol's rules, and the profit comes from the depth of understanding of the rules' economic implications rather than from violating them. *On-chain protocol compliance*, however, must be clearly distinguished from *legality under the law*: the CFTC's anti-manipulation framework (Commodity Exchange Act §9(a)(2)) and Article 91 of the EU's Markets in Crypto-Assets (MiCA) Regulation may both apply to coordinated cross-market manipulation strategies, regardless of whether the on-chain transactions are technically valid—a legal dimension explored in Chapter 26. This means that defending against strategic MEV cannot rely on *punishing the actor* but only on *fixing the mechanism vulnerabilities that make the behavior profitable.* This

marks the transformation of MEV from passive opportunity discovery to active strategy construction, a qualitative change from technical arbitrage to economic gamesmanship.

6.3.3 Defending against strategic MEV

In the face of increasingly frequent strategic MEV, traditional defenses based on technology and cryptography (such as encrypted mempools and batch auctions) prove inadequate, because strategic MEV does not exploit information asymmetry or network latency but the public, transparent economic rules themselves. Defending against strategic MEV must therefore begin with the protocol's mechanism design rather than a mere speed race. The table below systematically compares the defensive means and their effectiveness against the two types of MEV at four levels: the transaction layer, the ordering layer, the profit layer, and the mechanism layer.

Defense level	Defense against atomic MEV	Defense against strategic MEV
Transaction layer	Private transaction pools (hide trading intent)	Ineffective (strategic MEV does not depend on single-transaction information)
Ordering layer	PBS (separating builders and proposers)	Partly effective (reduces validator MEV but does not affect cross-cycle strategies)
Profit layer	MEV-Share (returns part of the profit to users)	Ineffective (strategic MEV profit comes from protocol mechanisms, not transaction ordering)
Mechanism layer	Not applicable	The core line of defense: redesigning the liquidation, insurance-fund, and margin mechanisms

Table 6-4. A comparison of the effectiveness of defenses against the two types of MEV at different levels (Data source: compiled by the author)

As Table 6-4 shows, the mature transaction-, ordering-, and profit-layer defensive tools for atomic MEV are either inapplicable or only partly effective against strategic MEV; only a redesign at the mechanism layer constitutes an effective line of defense, a judgment that directly determines the three core defensive directions below. The first defensive direction is to reshape the margin and dynamic-leverage model. To compete for market share, many perpetual futures platforms early on offered excessive high leverage and lowered initial-margin requirements, which gave strategic MEV ample leverage room. Introducing a dynamic, risk-based margin model (similar to the SPAN margin model of the Chicago Mercantile Exchange in traditional finance) is the core means of addressing this problem. But the SPAN analogy has clear limits for long-tail crypto assets: SPAN presupposes a well-behaved return distribution and deep liquidity for the underlying asset, whereas a meme coin like JELLY may experience volatility exceeding 1,000% within hours, fundamentally violating SPAN's statistical assumptions. Margin requirements must be dynamically linked to an asset's liquidity, historical volatility, and the concentration of a user's positions [24]. Hyperliquid's actual response after the JELLY incident (imposing open-interest caps and sharply reducing the maximum

leverage on long-tail assets) took a more direct path than dynamic margining, reflecting that under extreme liquidity conditions, coarse-grained position-size limits may be more operationally feasible than a fine-grained dynamic-margin model. This direction receives a deeper, engineering-oriented discussion in the liquidation-mechanism design of Chapter 11.

Above the margin layer, the design of the liquidation engine itself also needs fundamental optimization. Protocols need to introduce tiered liquidation and auto-deleveraging mechanisms: when market liquidity is insufficient to absorb a huge liquidated position, the system should not forcibly dump it on the market and cause a shortfall, but should instead trigger auto-deleveraging (ADL), forcibly closing counterparty positions against the most profitable traders on the opposite side of the liquidated position [26]. Although this harms the experience of some profitable traders, it protects the protocol's overall solvency. Paired with this is risk isolation of the insurance fund: high-risk assets (such as long-tail tokens) should have a separate isolated-margin pool, so that their liquidation losses are borne only by the liquidity providers participating in that pool, preventing the collapse of a single asset from triggering a systemic spread. Chapter 12 provides a detailed analysis of this from the perspective of insurance-fund governance.

The third level of defense concerns oracle resilience. To defend against oracle manipulation, a protocol must adopt a composite oracle mechanism that is multi-source and time-weighted or median-weighted, reducing the impact of an anomalous swing in any single spot market. At the same time, a *price circuit breaker* should be introduced: when the on-chain oracle price deviates significantly from the prices of multiple centralized exchanges, or when volatility exceeds a threshold within a short time, the opening and liquidation functions for that trading pair are automatically suspended to await a return to normal. But the circuit breaker itself can also become an attack surface: an attacker can deliberately trigger the circuit breaker during a genuine market crash to freeze the protocol and block the execution of legitimate liquidations, essentially constituting a denial-of-service vector disguised as a safety mechanism. The engineering implementation of these measures is given in concrete form in the clearing and settlement architecture design of Chapter 29.

6.3.4 The path toward institutionalizing MEV

The existence of MEV is an endogenous property of blockchains' decentralized architecture and cannot be eradicated. Although MEV causes wealth transfer and unfair treatment for ordinary users, some forms of MEV (such as cross-market arbitrage and timely liquidations) objectively improve the pricing efficiency and market liquidity of DeFi protocols [2]. The long-term direction of the industry should therefore not be a futile attempt to eradicate MEV but the institutionalization and standardization of MEV, moving it from disordered extraction toward an orderly economic mechanism.

The institutionalization of atomic MEV has begun to show results. In the Ethereum ecosystem, through the proposer-builder-separation mechanism promoted by Flashbots, the process of MEV extraction has been modularized and made market-based. As the most widely adopted implementation of PBS, MEV-Boost currently covers roughly 85% to 95% of Ethereum's block production [2]. Under this system,

searchers submit arbitrage bundles to block builders through sealed-bid auctions, and builders assemble them into complete blocks and pass them, via relays, to block proposers. Because of the fierce bidding competition among searchers, the vast majority of the profit MEV generates (roughly 85% to 93%) is ultimately forced to be paid to block proposers (that is, Ethereum's validator nodes) [2]. In addition, the MEV-Share protocol further returns part of the MEV profit to the original traders, while application chains' protocol-layer ordering rules (such as first-in-first-out, FIFO) reduce the room for certain forms of MEV to occur at the architectural level.

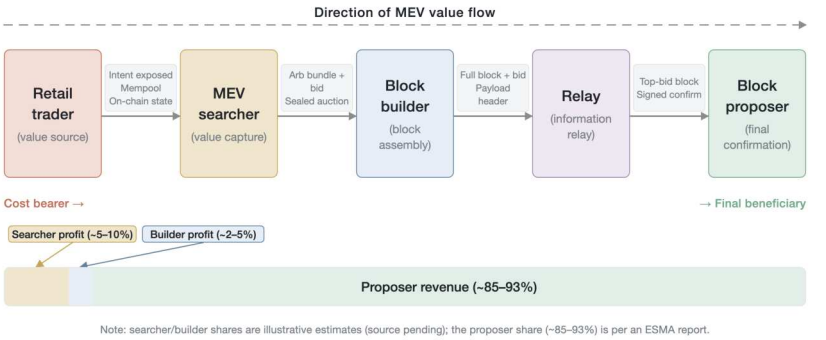


Figure 6-9. The value flow and profit-distribution architecture of the MEV supply chain (a value-distribution schematic, not empirical data; the 85%-to-93% figure for proposers follows ESMA [2], while the searcher and builder shares are the author's estimates; Data source: ESMA [2]; Flashbots Documentation)

As Figure 6-9 shows, value in the MEV supply chain flows out from ordinary traders, is captured by searchers and packaged by builders, and is finally passed, via relays, to block proposers, objectively subsidizing the security-consensus cost of the underlying blockchain. Yet even as the PBS mechanism reduces the risk of validator centralization, it also changes the incentive structure of the block-building market, driving it to evolve toward *generating and capturing more MEV*, which may increase total MEV and thus cause greater potential harm to ordinary users [27].

The institutionalization of strategic MEV is still in an early, exploratory stage. For application-layer protocols such as perpetual futures, institutionalizing MEV means *internalizing*, through mechanism design, the value that external extractors would otherwise seize for free, turning it into protocol revenue or user welfare. Concrete directions include: protocol-level position-size limits and a redesign of the liquidation mechanism (making strategic exploitation economically infeasible); governance-driven dynamic parameter adjustment and emergency-response mechanisms (as when Hyperliquid, after the JELLY incident, sharply tightened the leverage ceiling and position limits on long-tail assets); and cross-platform sharing of risk-control information (although this direction still faces enormous coordination challenges in a decentralized environment).

The structural resemblance between institutionalizing MEV and traditional financial regulation. The process of institutionalizing MEV bears a deep structural resemblance to the regulation of high-frequency trading in traditional finance. Both face

the same core problem: how, when information competition cannot be eliminated, to make it fairer and more transparent through institutional design. The market-fairness issues MEV raises correspond directly, at the conceptual level, to front-running prohibitions in traditional finance, but the decentralized nature of blockchains makes traditional regulatory enforcement hard to apply directly [2]. Regulators are gradually stepping into this area, exploring how to map traditional finance's principles of *best execution* and market-manipulation prevention onto the DeFi environment. In its 2023 DeFi policy recommendations, the International Organization of Securities Commissions (IOSCO) already explicitly required *responsible persons* to identify, disclose, and, where feasible, manage and mitigate the impact of MEV strategies [2]. Yet in a sufficiently decentralized protocol, identifying the *responsible person* is itself an unresolved enforcement problem, and IOSCO itself acknowledged in its 2023 report that when a protocol is governed by a token-based DAO, its recommended framework may lack a clear enforcement target.

In the long run, the evolution of strategic MEV will force the maturation of on-chain financial infrastructure. Only when protocol design can find a balance between openness and security, and when the value distribution of MEV can reconcile efficiency and fairness, can decentralized derivatives markets truly build institutional-grade trust and achieve larger-scale growth.

The evolution of MEV's forms and the process of its institutionalization directly affect the market's informational efficiency and pricing quality. Large-scale extraction of atomic MEV raises the implicit cost of on-chain trading and lowers informed traders' willingness to express information on-chain, which may suppress the price-discovery function of on-chain markets. Conversely, the effective institutionalization of MEV (reducing searchers' rent-seeking room through mechanisms such as PBS and MEV-Share) helps improve on-chain execution quality and creates conditions for the migration of price discovery from CEXs to on-chain venues. This logic forms an important premise for the discussion of the price-discovery migration hypothesis in the next section.

6.4 The price-discovery migration hypothesis

Market structure determines not only the distribution of risk and the mode of competition; its more far-reaching influence lies in reshaping the pricing power over assets. In the perpetual futures ecosystem, *price discovery*—the process by which new information is incorporated into asset prices—is a core topic of market-microstructure research. This section advances the *price-discovery migration hypothesis*: there is a structural trend indicating that the dominance of price discovery is gradually shifting from centralized exchanges to application-chain decentralized exchanges. This trend is not an incidental fluctuation in liquidity but a systematic change driven by deep economic forces. In theory, once this migration reaches a *critical mass*, the network effects of liquidity may render it irreversible. As Section 6.4.3 will argue, however, the current market evidence is not yet sufficient to confirm that the migration has crossed an irreversible critical point, and the analysis in this section should be regarded as a research hypothesis awaiting empirical testing.

6.4.1 Structural drivers

The trend of price discovery migrating from CEXs to on-chain venues does not stem from an ideological call but is driven by three mutually reinforcing structural forces. Before analyzing these drivers, one important confounding variable deserves attention: the 346% year-over-year growth in DEX perpetual futures volume must be read against the CEX growth rate of 29% and against the macro environment of the same period (the Federal Reserve's rate-cutting cycle, the approval of spot Bitcoin ETFs, and a general recovery in risk appetite). DEX volume may have a higher beta to risk-appetite sentiment, because in a bull market new retail traders disproportionately enter through DeFi channels. Part of the growth may therefore reflect a cyclical rising-tide effect rather than a purely structural advantage. Acknowledging this confounding variable, the following three structural forces remain identifiable.

The most fundamental of these forces is the trust premium of verifiable liquidity. In the black-box architecture of a CEX, the depth of the order book cannot be independently verified by outsiders. This opacity not only masks potential wash trading but, in extreme market conditions, also deprives participants of confidence in the authenticity of liquidity. The collapse of FTX (see Section 6.1.1) became a landmark event, and market participants came to grasp the fundamental gap between *paper liquidity* and *real liquidity* [28]. By contrast, every resting order on an on-chain central limit order book is a genuine on-chain commitment that anyone can independently audit cryptographically. As institutional investors and high-frequency trading firms increasingly value the verifiability of liquidity, this transparent liquidity acquires a significant trust premium. This premium is continuously pushing capital to migrate from CEXs to on-chain architectures. As the left panel of Figure 6-10 shows, the ratio of DEX to CEX perpetual futures volume climbed from 2.1% in the first quarter of 2023 to 11.7% in November 2025, with the trust crisis after the FTX collapse markedly accelerating this trend.

On top of the trust premium, the efficiency gains of protocol-level market making constitute the second driver of the migration. Traditional CEX market making relies heavily on a few professional institutions with top-tier infrastructure (such as Wintermute, Jump Trading, and Cumberland), and this high entry threshold leads to market-making profits being monopolized by a handful of institutions. Application-chain DEXs, however, are breaking this monopoly through protocol-level innovation. Take Hyperliquid's HIP-3 mechanism and its HLP vault: market making shifts from the privilege of a few institutions to an open, protocol-driven public service. Any capital holder can participate in market making by injecting funds into the HLP vault, and the protocol automatically manages quoting strategies and risk exposure. This architecture attracts broader capital participation in liquidity provision and significantly increases order-book depth. Data made public by the Hyperliquid team in early 2026 show that the cumulative resting-order depth of its Bitcoin perpetual futures at certain price levels has approached twice that of Binance, with bid-ask spreads comparable to those of top-tier CEXs [29]. Note, however, that these data come from a self-report by an interested party; the potential bias and the need for independent cross-validation are detailed in Section 6.4.3. More fundamentally, assessing liquidity quality requires distinguishing *resilient depth* from *fragile depth*: orders provided by independent, profit-

driven market makers are usually actively replenished after being filled, constituting *resilient depth*; whereas the liquidity provided by the HLP vault—a single, concentrated, protocol-driven LP—may freeze or be withdrawn under stress as participants panic-withdraw, constituting *fragile depth*. The JELLY incident already demonstrated the structural fragility of HLP's concentrated liquidity provision. This efficiency gain directly strengthens the on-chain market's ability to absorb large orders and reflect new information, which is precisely the core function of price discovery.

The on-chain migration of traditional financial institutions has further accelerated this process. Professional trading firms are moving their technical capabilities and capital from CEXs to on-chain markets. Market makers and proprietary trading firms from Wall Street have already begun deploying strategies on-chain [30]. For example, Jump Trading invested in Firedancer, a high-performance validator client for Solana, while Cumberland became a core data provider for the Pyth oracle network. These moves reflect deep institutional conviction in the price-discovery potential of on-chain markets. The inflow of institutional capital brings not only deeper liquidity but, more importantly, more efficient information processing: professional market makers possess sophisticated pricing models and risk-management systems, and their participation directly improves the speed and accuracy with which on-chain markets incorporate new information into prices. As the right panel of Figure 6-10 shows, annual DEX perpetual futures volume grew from roughly $0.3 trillion in 2022 to roughly $6.7 trillion in 2025, a compound annual growth rate exceeding 180%, with accelerating institutional participation an important driver of this growth.

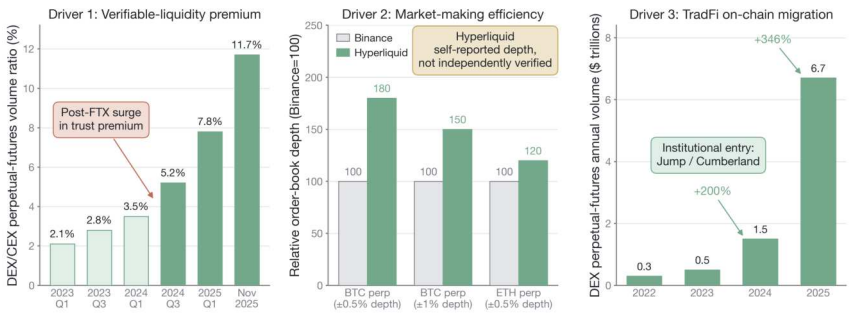


Figure 6-10. The three structural drivers of the price-discovery migration and the current evidence (the left panel shows the ratio of DEX to CEX perpetual futures volume, through November 2025 [4]; the middle panel shows Hyperliquid's self-reported resting-order depth [29], for a specific price level and time window, not cross-validated against independent data sources such as Kaiko and Amberdata; the right panel shows annual DEX perpetual futures volume [3], with +200% and +346% on a year-over-year basis; Data source: CoinGecko [4], DeFiLlama, The Block [3], and Hyperliquid team public data [29])

The three panels of Figure 6-10 together sketch the current progress of the price-discovery migration across three dimensions: market share, liquidity depth, and absolute scale. As the next subsection will argue, however, these data are not yet sufficient to prove that the migration has crossed an irreversible critical mass.

6.4.2 Critical mass and network effects

The liquidity of financial markets has powerful, self-reinforcing network effects. As classic theory in financial economics reveals, liquidity attracts liquidity. In perpetual futures markets, this network effect manifests as a clear positive feedback loop: deeper liquidity leads to narrower bid-ask spreads; narrower spreads attract more traders; the gathering of traders brings more order flow; abundant order flow further attracts more market makers; and the increase in market makers in turn provides deeper liquidity (see Figure 6-11).

This mechanism means that if the liquidity of application-chain DEXs can persistently surpass that of CEXs on certain core assets, this positive feedback loop may make the migration of price discovery irreversible. For once a certain *critical mass* is crossed, any effort to pull liquidity back to the traditional architecture must overcome the same enormous network-effect barrier. This closely resembles the competitive dynamics among exchanges in traditional finance: once an exchange gains an overwhelming liquidity advantage in a particular asset class, competitors find it nearly impossible to reclaim market share in that asset.

Assessing the current state of the migration, however, demands extreme rigor. Empirical research shows that the current crypto market still exhibits a clear two-tiered structure: CEXs remain dominant in price discovery, with a price-integration measure about 61% higher than that of DEXs [31]. The specific methodology behind this figure (information share, the Gonzalo-Granger decomposition, or another price-discovery measure), the sample period, and the range of assets analyzed all significantly affect the estimate. Such estimates are highly sensitive to the sampling frequency and the lead-lag specification, and given the rapid change in DEX market share, even recent research may already be partly outdated. New information is reflected in prices at the CEX first and only then transmitted to DEXs through the funding rate and arbitrage mechanisms. This finding shows that although the direction of the migration trend is clear, and it has already completed the stage-transition from *impossible* to *observable*, it remains a considerable distance from the *irreversible* critical mass.

Figure 6-11 assesses the current state of the price-discovery migration across five key dimensions: liquidity depth (for some assets, such as BTC perpetual futures, this has approached or even surpassed the CEX) is the dimension closest to critical mass, whereas on the price-discovery-contribution (information-share) dimension, the CEX is still about 61% higher, the strongest evidence that the migration is incomplete; trading-volume share, institutional participation, and the persistence of liquidity after incentives are withdrawn are the key variables that will determine whether the migration can move from *observable* to *irreversible*.

6.4.3 A cautious reading of the current evidence

Regarding the price-discovery migration hypothesis, we must maintain a cautious scholarly stance and recognize several key limitations in the current market evidence.

Potential bias in data sources. Some of the prominent data currently supporting the migration hypothesis (such as order-book-depth comparisons at specific price levels)

come from public statements by application-chain team members or from reports by interested parties. Although these data have reference value, they carry an inherent conflict of interest. For example, the depth data made public by the Hyperliquid team selected the price levels and time windows most favorable to itself and may not reflect true conditions across all hours and all market states. These data need to be cross-validated through independent academic research and third-party microstructure data (from professional data providers such as Kaiko and Amberdata).

Snapshot versus norm. A depth snapshot at a single point in time or under a specific market environment cannot represent the market's persistent norm. Liquidity may temporarily surpass the CEX in certain periods (such as the peak of protocol token incentives, the anticipation phase of an airdrop, or when market sentiment is extremely optimistic) but may fall back significantly after incentives end or sentiment shifts. Assessing the true progress of the migration requires examining the retention rate of liquidity after a complete market cycle (including bull markets, bear markets, and extreme volatility events). "Incentives are highly effective at attracting volume but ineffective at retaining risk. Fee rebates and points programs can rapidly mobilize activity, but they do not by themselves generate durable open interest. Risk does not stay where it is subsidized; it stays where it is safely held." [32]

Depth does not equal price-discovery quality. Order-book depth is only one dimension of market quality. The quality of price discovery depends more on informational efficiency (the speed and accuracy with which new information is incorporated into prices) and on market resilience (the speed at which prices recover after being hit by a large order). A market with an extremely deep order book but slow information transmission may have lower price-discovery quality than a shallower market that reacts to information far more quickly. These multidimensional assessments require more precise microstructure tools—such as the Hasbrouck information share [33] and the Gonzalo-Granger permanent-transitory decomposition—for empirical testing, which is taken up in Chapters 13 through 15.

The role of seigniorage in liquidity formation. The perpetual futures DEX market is at a critical juncture in the transition from *subsidy-driven* to *market-structure-driven* [32]. The year 2025 is the first in which multiple perpetual futures DEXs began to pass this test at scale. Yet the liquidity of many emerging platforms still relies heavily on token-inflation incentives (seigniorage). Only when the main source of on-chain liquidity substantively shifts from token incentives to genuine trading demand and sustainable market-making profits can this migration of price discovery be considered truly sustainable.

> Conclusion: The directional judgment of the migration hypothesis is well-founded, but it is currently still at the stage of *a trend that is identifiable but not yet irreversible*. Chapters 13 through 15 will use more precise microstructure tools to assess this judgment.

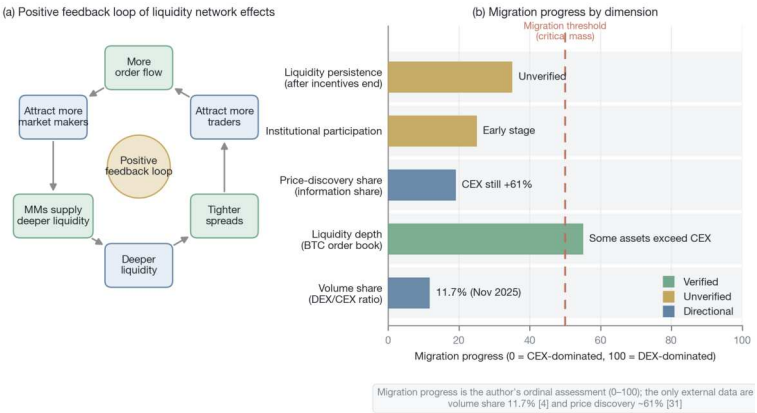


Figure 6-11. The positive feedback loop of liquidity network effects and the current-state assessment of the price-discovery migration (a positive-feedback-loop schematic overlaid with the author's migration scores, where 0 to 100 is the author's composite assessment, not a single quantitative metric; the only external empirical data are the 11.7% volume share, November 2025 [4], and the CEX price-information share being about 61% higher [31]; Data source: CoinGecko [4], Zhivkov [31], and Hyperliquid public data [29])

6.4.4 The shaping of trader behavior

The migration of price discovery is not merely a reallocation of macro market share; at the micro level it profoundly shapes traders' behavioral patterns. Market structure is by no means a passive *backdrop* to trading behavior but an active *shaper* of it. Understanding this has a decisive influence on assessing the quality of price discovery under different architectures.

Empirical research on the behavioral differences between CEX and DEX traders reveals the concrete impact of this shaping effect [10]. That study systematically compared the behavioral differences of traders on CEXs and DEXs and found a striking phenomenon: on CEXs, traders' behavioral patterns are closer to those of traditional financial markets, showing a relatively balanced response to price signals in both directions—roughly symmetric sensitivity to good news and bad news. On oracle-priced DEXs, however, uninformed traders respond significantly more strongly to positive news than to negative news, producing a systematic long bias.

The root of this behavioral difference lies in the design of the underlying mechanism. Oracle-pricing models (such as the virtual automated market maker, or vAMM, and oracle-based synthetic assets) structurally reduce the slippage and adverse-selection risk a trader faces when building a long position. When a trader goes long on an oracle-priced DEX, the execution price is determined directly by the external price the oracle provides and is not constrained by local order-book depth, so a large long order does not push up the execution price the way it would in a traditional central limit order book. This mechanical asymmetry regularly incentivizes going long and suppresses going short.

To present these behavioral differences systematically, Table 6-5 compares the behavior of CEX traders and oracle-priced DEX traders across four key dimensions and traces the mechanism root behind each difference.

Behavioral dimension	CEX traders	Oracle-priced DEX traders	Mechanism root of the behavioral difference
Response to positive news	Moderately increase long positions	Significantly increase long positions	Oracle pricing lowers long-side slippage
Response to negative news	Moderately increase short positions	Relatively muted response	Shorting still faces liquidity constraints
Leverage propensity	Fairly evenly distributed	Skewed toward higher leverage	Low slippage encourages more aggressive positions
Holding period	Relatively short (mostly intraday)	Relatively long	Lagging oracle-price updates may systematically depress the deviation between the mark price and the index price, thereby lowering the absolute value of the funding rate and the cost of holding a position (a hypothetical explanation, awaiting further empirical testing)

Table 6-5. Behavioral differences between CEX and oracle-priced DEX traders and their mechanism roots (Data source: compiled by the author based on Chen, Ma, and Nie [10])

Table 6-5 shows that these behavioral differences are not random but correspond systematically to the underlying pricing mechanism—the structural weakening of long-side slippage under oracle pricing simultaneously incentivizes a long bias, higher leverage, and longer holding periods across all four dimensions. Beyond the mechanistic explanation, a behavioral perspective offers a complementary understanding: the amplifying effect of social media (narrative contagion in Crypto Twitter and Telegram groups) creates a powerful long-direction herding effect that superimposes on and reinforces the structural incentives of oracle pricing. This systematic long bias has reflexive implications: deviation from the CEX funding rate creates a predictable arbitrage flow, and the arbitrage flow itself becomes a source of distortion in the cross-CEX-DEX price signal. This finding shows that different pricing mechanisms, fee structures, and risk-management rules regularly incentivize or suppress specific types of trading behavior, thereby systematically affecting the direction and efficiency of price discovery; it also provides the macro background for the deeper discussion of order-book microstructure and mechanism design in later chapters (Chapters 7 and 8): every adjustment of a protocol parameter is not merely optimizing system performance but redefining the game matrix of market participants and thereby shaping the price-discovery characteristics of the entire market.

6.5 The evolution of risk distribution

The evolution of market structure has changed not only the mode of competition and the pricing mechanism but also, fundamentally, the distribution of systemic risk. In assessing the migration from centralized to decentralized architectures, a common error is to seek a simple binary conclusion: *Is on-chain safer than a centralized exchange?* This way of framing the question is itself flawed, because it presupposes that risk can be compared along a single dimension. This section advances the *risk visibility-manageability proposition*, arguing that the evolution from CEXs to on-chain architectures is essentially not a simple increase or decrease in the total quantity of risk but a structural transformation in the character of risk—from *invisible but occasionally explosive* to *continuously visible but manageable*. This framework is a dynamic extension of the *risk substitution theorem* of Chapter 5 and is of a piece with the concept of *risk-surface transformation* advanced in Chapter 1.

6.5.1 The risk visibility-manageability proposition

A core axiom in the theory of financial risk management is this: risk that cannot be observed and measured cannot be effectively managed. This axiom runs through the entire body of risk-management knowledge, from the Basel Accords to modern portfolio theory. On this basis, we advance the risk visibility-manageability proposition.

> Proposition: The visibility of risk is a necessary but not sufficient condition for its manageability. Invisible risk cannot be accurately quantified and therefore cannot be effectively hedged or managed ex ante; it can only be dealt with passively, after the fact, when a crisis erupts. Yet visibility alone does not guarantee manageability; the effective management of risk also depends on adequate loss-absorption capacity, appropriate circuit-breaker mechanisms, and a tiered liquidation system.
> Corollary: The evolution from CEXs to on-chain architectures marks a substantial improvement in the visibility of systemic risk. This does not mean the absolute level of risk falls; rather, it means the nature of risk changes fundamentally, from tail-risk events that are *unmanageable but occasionally require enduring catastrophic consequences* [34] to a normalized challenge that is *manageable but requires the continuous investment of resources for monitoring and protection*. The JELLY incident provides an important boundary-condition test of this proposition: the attack was fully visible on-chain in real time, and the community completed a diagnosis of the attack vector within hours—visibility did indeed accelerate identification of the problem. Yet the protocol could not manage this risk through automated mechanisms and ultimately still required centralized intervention (unilateral delisting and forced liquidation), which at the operational level is equivalent to the CEX behavior this chapter criticizes. This shows that, above visibility, the *manageability* of risk may still depend on residual centralized power, a tension especially pronounced in application chains with limited decentralization.

Under the CEX paradigm, risk exhibits the characteristics of *low-frequency, invisible, and occasionally catastrophic*. In day-to-day operations, the user experience is usually extremely smooth: trades execute fast, the interface is friendly, and customer service responds promptly. Risk is hidden inside the black box of centralized servers, and the safety of a user's assets rests entirely on one-sided trust in the exchange's solvency

and internal governance. Yet once this invisible risk accumulates to a critical point, it often erupts in an extremely destructive way. The customer-fund losses caused by the FTX collapse and its bankruptcy proceedings, which lasted more than two years (see Section 6.1.1) [28]; the $1.5 billion security incident suffered by Bybit in February 2025 [35]; and the loss of roughly 850,000 BTC by Mt. Gox in 2014 (on the basis of the bankruptcy-trustee / civil-rehabilitation proceedings) [54], with its legal-recovery process stretching over nearly a decade, are all textbook cases of this pattern. Their common feature is that, before the disaster, external market participants had almost no warning, and the social cost was extremely high.

By contrast, risk under an on-chain architecture exhibits the characteristics of *high-frequency, visible, and manageable*. In a general-purpose public-chain or application-chain DEX, the MEV game plays out in real time in the public mempool, the logic and potential vulnerabilities of smart contracts are open to all auditors, and every parameter and trigger condition of the liquidation process is clearly queryable on-chain. Users and institutions can intuitively *see* these risks every day. Precisely because of this continuous visibility, market participants can build real-time monitoring systems, early-warning mechanisms, and hedging strategies. Take the JELLY attack Hyperliquid suffered in March 2025: the attacker exploited a design flaw in the liquidation mechanism, and the protocol vault consequently faced a peak unrealized loss of roughly $13.5 million (not a realized loss; the HLP vault recorded a small profit after final settlement—see the detailed analysis of this incident in Section 6.3.2); but the protocol-level response completed within hours: the token was delisted, positions were settled at the mark price, and parameters were tightened [36]. This pattern of *rapid exposure, rapid repair* stands in sharp contrast to the CEX pattern of *long concealment, sudden collapse*.

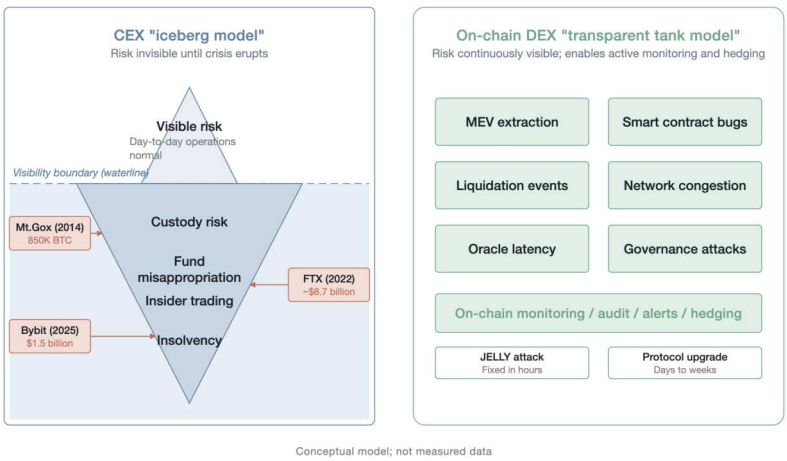


Figure 6-12. A visualization of the risk visibility-manageability proposition (a CEX's tail risks—custodial risk, misappropriation of funds, insolvency—are invisible in daily operation until a crisis erupts; on-chain DEX risk types—MEV extraction, smart-contract vulnerabilities, liquidation events—are continuously visible, supporting active monitoring and hedging; a conceptual-model schematic, not empirical data; the roughly $8.7 billion FTX figure is on the debtors' customer-shortfall / August 2024 court restitution-order basis, Bybit's $1.5 billion [7], and JELLY's within-hours repair [36]; drawn by the author)

In assessing the merits of these two patterns, the lessons of financial history repeatedly prove a profound truth: invisible risk is not absent risk; it means only that when it erupts, the system will be utterly unprepared. One root of the 2008 global financial crisis was precisely that complex structured financial products hid the risk of subprime loans within multiple layers of securitization, leaving regulators and market participants unable to accurately assess the true scale of systemic risk. The collapse of FTX is the crypto market's version of the same lesson. Therefore, the systemic destructive power of a tail-risk event far exceeds the economic cost of sustaining continuous security investment. The evolution of risk toward a visible state is a necessary condition for a market's maturation.

6.5.2 Statistical properties of risk

To characterize this evolution more precisely, we need to systematically compare the statistical properties of risk across the three service-provider architectures. This is not only a dynamic extension of the *risk substitution theorem* of Chapter 5 but also a concretization of the concept of *risk-surface transformation* of Chapter 1, providing an analytical framework for quantitatively assessing market quality.

Dimension	Centralized exchange	General-purpose public-chain DEX	Application-chain DEX
Frequency of risk events	Low (roughly five catastrophic events with single losses exceeding $1 billion between 2022 and 2025, including FTX and Bybit; the total is significantly higher if bankruptcies such as Celsius, BlockFi, and Voyager and mid-sized security incidents are included)	High (per EigenPhi data, daily MEV extraction on Ethereum exceeds $500,000 [18])	Medium (in 2025, Hyperliquid experienced three strategic attacks on the order of tens of millions of dollars)
Single-shock magnitude	Extremely high (FTX roughly $8.7 billion, on the debtors' customer-shortfall basis [1]; Bybit $1.5 billion [35])	Low to medium (a single MEV extraction is usually thousands to tens of thousands of dollars; a contract-vulnerability attack can reach millions)	Medium (the three 2025 incidents were roughly $4 million, $13.5 million, and $4.9 million, respectively [20][37][25])
Ex ante predictability	Extremely low (operates inside a black box; not externally observable)	High (on-chain data and the mempool are globally traceable)	Medium to high (traceable on-chain, but strategic attacks have a degree of covertness)
Crisis-recovery path	Lengthy legal and bankruptcy proceedings (FTX about 2 years, Mt. Gox about 10 years)	Smart-contract repair and deployment (on the order of days to weeks)	Governance votes and protocol-level emergency upgrades (Hyperliquid's JELLY incident about 6 hours [36])

Table 6-6. A comparison of the statistical properties of risk across the three paradigms (Data source: compiled by the author based on [1][18][20][25][35][36][37])

The comparison in Table 6-6 presents a clear frequency-versus-shock-magnitude trade-off: the CEX has the lowest frequency of risk events but the highest single-shock magnitude (on the order of billions of dollars), whereas on-chain architectures have more frequent risk events but relatively controllable single losses (on the order of millions to tens of millions of dollars). The difference in the crisis-recovery path is especially striking: CEX bankruptcy recovery is measured in years, while on-chain protocol repair is measured in hours to weeks—an order-of-magnitude gap that directly corroborates the core assertion of the risk visibility-manageability proposition. Looking at the distribution of specific risk types, the three paradigms present markedly different profiles, echoing the risk substitution theorem of Chapter 5: rather than eliminating risk, different architectures substitute one form of risk for another.

Beyond the macro trade-off between frequency and shock magnitude, the three paradigms also present markedly different profiles in the distribution of specific risk types. From seven core risk categories—custodial/counterparty risk, smart-contract

risk, MEV/information leakage, liquidation/shortfall, economic security, cross-chain bridges, and regulation/compliance—Table 6-7 compares, one by one, the degree of exposure and the manifestation of each risk in the three architectures.

Risk type	CEX paradigm	General-purpose public-chain DEX paradigm	Application-chain DEX paradigm
Custodial/counterparty	Extremely high: assets are custodied by the exchange, facing theft, misappropriation, and bankruptcy risk	Extremely low: self-custodial wallet, managed by smart contracts	Low: on-chain account management, but the cross-chain bridge is the weak link
Smart contract	None: no on-chain smart contracts involved	High: contract vulnerabilities can cause fund losses	Medium: core logic embedded at the protocol layer, with a smaller attack surface
MEV/information leakage	Implicit: information asymmetry exists inside the exchange	Explicit and high: the public mempool exposes trading intent	Controllable: the private mempool mitigates traditional MEV, but validator MEV is a new risk
Liquidation/shortfall	Socialized sharing by the insurance fund; in extreme cases, transferred via ADL	Managed by the protocol insurance fund; the liquidation process is public and transparent	Managed by the protocol insurance fund; governance can adjust liquidation parameters
Economic security	None: does not depend on a token-economic model	Low: inherits the economic security of the underlying blockchain	Medium to high: relies on its own token's market capitalization to maintain network security
Cross-chain bridge	None: no cross-chain operations involved	Low: the official L2-to-L1 bridge has relatively strong security guarantees	Medium to high: assets must pass through a cross-chain bridge to enter the application chain
Regulation/compliance	High: a regulated entity, facing license revocation and asset freezes	Medium: decentralization raises the difficulty of enforcement, but the front end can be regulated	Medium: between a CEX and a pure DEX, with an as-yet-unclear regulatory status

Table 6-7. A comparison of the distribution of seven core risks across the three architectures (Data source: compiled by the author)

Table 6-7 clearly displays the concrete paths of risk substitution: the CEX's custodial/-counterparty risk is largely eliminated in on-chain architectures, but it is replaced by smart-contract risk and cross-chain-bridge risk; the implicit information asymmetry of the CEX turns into explicit MEV extraction in the general-purpose public-chain DEX and then evolves into the more covert validator MEV in the application-chain DEX. No architecture achieves the absolute elimination of risk; what differs is only the form of risk exposure and the degree to which it is manageable.

Figure 6-13 integrates Tables 6-6 and 6-7 into four complementary visual perspectives: (a) a frequency-versus-shock-magnitude scatter plot, (b) a radar chart of the seven risk types, (c) a comparison of recovery-path times, and (d) a visibility-manageability quadrant chart, intuitively presenting the overall direction of risk evolving from *low-frequency, high-shock, unmanageable* to *high-frequency, low-shock, manageable*. Here, the order-of-magnitude gap between recovery paths measured in years (CEX: FTX about 2 years, Mt. Gox about 10 years) and those measured in hours to weeks (on-chain: JELLY about 6 hours) is the most intuitive corroboration of the risk visibility-manageability proposition.

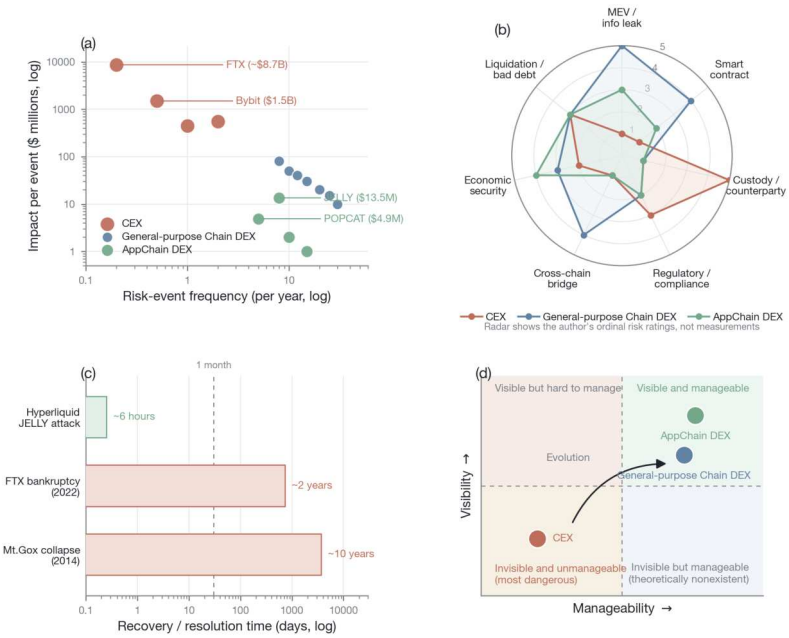


Figure 6-13. The evolution of risk distribution across the three paradigms: a frequency-versus-shock-magnitude scatter plot, a radar chart of the seven risk types, a comparison of recovery-path times, and a visibility-manageability quadrant chart (panels a and c use public event data, sourced event by event: FTX roughly $8.7 billion, on the debtors' customer-shortfall / restitution-order basis; Bybit $1.5 billion [7]; POPCAT roughly $4.9 million [25]; JELLY roughly $13.5 million, peak unrealized loss [37], recovered within hours [36]; panels b and d use the author's ordinal risk ratings, not quantitative measurements; Data source: public event records, 2014 to 2025; the DEX/CEX volume ratio additionally follows CoinGecko [4])

6.5.3 Cross-paradigm risk transmission

Although the three paradigms differ significantly in architectural design and risk distribution, they are not systems that operate in isolation. In the current crypto ecosystem, they are tightly connected by the price of the same underlying asset, constituting a complex financial system of *communicating vessels*. While this interconnection improves overall market efficiency (eliminating price gaps through cross-market arbitrage), it also creates a lethal path for cross-paradigm risk transmission. Beyond the oracle-mediated price-transmission channel detailed below, the stablecoin liquidity channel constitutes

another equally critical but more covert cross-paradigm transmission path. DEX perpetual futures use USDT and USDC as their principal collateral, and the credit basis of these stablecoins derives from off-chain fiat reserves and short-term Treasury portfolios. In a macro-stress scenario (as demonstrated by the USDC depeg event triggered by the collapse of Silicon Valley Bank in March 2023), transmission occurs not through the oracle price but through the collateral layer itself: a stablecoin confidence crisis would simultaneously impair the collateral quality of all three paradigms, triggering correlated liquidations. This channel bypasses all oracle-based circuit-breaker designs and constitutes one of the hardest-to-defend systemic risks in the current market structure.

In extreme market volatility, this transmission usually completes within minutes, and its typical cyclical path is shown in Figure 6-14. Because CEXs aggregate large numbers of highly leveraged retail traders, when the market suffers a sudden shock, the CEX is the first to trigger large-scale forced liquidations, causing prices to fall sharply. This negative feedback loop of *forced liquidation → price decline → more liquidations* is precisely the crypto-market realization of the *loss spiral* that Brunnermeier and Pedersen (2009) formalized in *Market Liquidity and Funding Liquidity*, in which the mutually reinforcing decay of market liquidity and funding liquidity is sharply amplified in a high-leverage environment [38]. Subsequently, the oracle network rapidly transmits the CEX crash price on-chain, triggering a chain of liquidations of highly leveraged DEX positions; on-chain liquidation flow strikes market makers' liquidity pools or order books, causing the DEX price to deviate significantly from the CEX; cross-market arbitrageurs then buy on the lower-priced DEX and sell to hedge on the CEX, transmitting new selling pressure back to the CEX and thereby potentially opening a new round of the negative feedback loop.

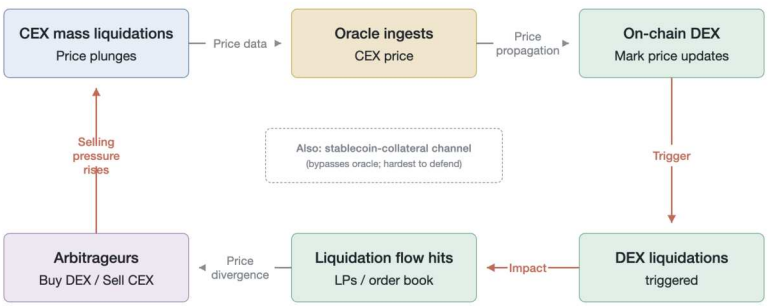


Figure 6-14. The cyclical path of cross-paradigm risk transmission (the price shock triggered by CEX liquidations is transmitted on-chain to DEXs via the oracle, triggering a chain of liquidations, and arbitrageurs' cross-market operations feed selling pressure back to the CEX; a conceptual-cycle schematic, the oracle-mediated channel, not empirical data; drawn by the author, based on FTI Consulting [39])

There are two key amplifiers in this transmission mechanism. The immediacy of the oracle is one. In traditional finance, the settlement delay between different markets often serves as a buffer—for example, the daily settlement mechanism of futures markets gives participants a time window to adjust positions. In the crypto market, however, oracle networks differ fundamentally in architecture: Chainlink uses a push-based heartbeat-

plus-deviation-threshold model (updating automatically when the price moves more than, say, 0.5% or after N minutes), while Pyth uses a pull-based model (consumers request the latest price on demand). This difference has different effects on the speed of cross-paradigm transmission; a pull-based oracle may in fact amplify the transmission effect, because liquidators actively pull the latest, most unfavorable price to trigger liquidations and capture rewards. Regardless of the model, oracles typically reflect price changes at second-level frequency, eliminating the settlement buffer of traditional finance, so that a price crash on the CEX side can be transmitted within seconds to all on-chain protocols relying on that oracle, triggering synchronized chain liquidations. Superimposed on the immediacy of the oracle is the blind spot of each platform's local perspective: whether it is a CEX's liquidation engine or a DEX's protocol parameters, their risk-control mechanisms can only make decisions based on local data from their own platform, and no single platform can assess the systemic leverage level of the entire market. A CEX's risk-control system cannot know how large a leveraged position the same batch of traders holds on an on-chain DEX, and vice versa. This *locally reasonable but globally inadequate* parameter setting (such as the liquidation-penalty rate, the maintenance-margin ratio, and the auto-deleveraging trigger conditions) is extremely fragile in the face of a cross-market resonant shock.

The crypto-market flash crash of October 2025 provides a vivid empirical illustration of this transmission mechanism. According to FTI Consulting's analysis, Bitcoin fell roughly 14% within hours, and the visible notional value of forced liquidations across CEXs and DEXs exceeded $19 billion within about 24 hours. The wave of CEX-side liquidations erupted first, and the on-chain DEX liquidation wave followed close behind, with the lag between them only a few minutes, during which the price deviation of major oracles at one point reached over 1% [39]. High leverage, shrinking order-book depth, the brief failure of oracles, and cross-platform auto-deleveraging mechanisms together constituted a severe liquidity crisis. In that event, the liquidation engines of multiple platforms triggered large-scale forced liquidations within the same time window, while each platform's risk-control system saw only its own internal risk exposure and could not assess the systemic leverage level of the entire market. This cross-paradigm risk transmission is the most severe systemic challenge facing today's perpetual futures market and a core problem that future regulatory frameworks and protocol engineering must confront head-on. Chapter 20 provides a formal analysis of this cross-paradigm transmission from the liquidity perspective, and Chapter 22 from the volatility perspective.

6.6 The paradigm endgame

After a systematic dissection of the competition modes, MEV forms, and risk distribution of the three paradigms—centralized exchanges, general-purpose public-chain decentralized exchanges, and application-chain DEXs—a core macro question must be addressed: Where are these three paradigms headed? Does the evolution of market structure necessarily follow the one-way, linear logic of *decentralization replacing centralization*, or will it display a more complex dynamic pattern?

The analysis in this section shows that the three paradigms will not converge on an

endgame of *one model rules them all*; instead, they will coexist over the long term amid competition and converge on one another through mutual borrowing. The key uncertainty is where the terminal state of this convergence will land, and whether advances at the technological frontier can effectively expand the area of the *impossible triangle* and thereby redefine the boundary of what is possible for market structure. At the same time, two fundamental unresolved problems—the effectiveness boundary of decentralization and the dual effect of transparency in a high-leverage environment—will continue to shape the evolutionary direction of the next generation of market structure.

6.6.1 Coexistence and the competitive landscape

Although decentralized exchanges have grown rapidly over the past few years, the market data clearly reveal a multipolar coexistence rather than a simple relationship of replacement. Over the full year of 2025, the top ten perpetual futures exchanges (spanning both CEXs and DEXs) handled as much as $92.9 trillion in volume, up 64.6% year over year [40]. Within this volume growth, DEX perpetual futures volume surged 346% to a record high of $6.7 trillion, and the ratio of DEX to CEX perpetual futures volume climbed further to 11.7% (for the starting point and trajectory of this ratio, see this chapter's introduction and Section 6.4) [40]. This does not mean, however, that CEXs are being comprehensively replaced; in absolute terms, CEXs still hold roughly 85% or more of market share.

This coexistence pattern arises because the three paradigms each occupy an irreplaceable niche. The centralized exchange remains the venue of choice for extremely latency-sensitive high-frequency market-making strategies, for retail investors with limited technical ability, and for institutional capital that depends on deep fiat on- and off-ramps. Through millisecond-level matching engines and internal-ledger settlement, CEXs offer the highest capital efficiency and liquidity depth achievable under current technical conditions. General-purpose public-chain DEXs (such as protocols deployed on Ethereum or its general-purpose Layer-2s), although unable to rival CEXs in performance, satisfy DeFi-native strategies with a strong demand for *composability*; on these platforms, perpetual futures can interact seamlessly with lending protocols and yield aggregators to form complex structured products. Application-chain DEXs (such as Hyperliquid) and the neighboring application-specific ZK-Rollups that anchor finality to Ethereum's settlement layer (such as Lighter, which strictly speaking does not have independent consensus—see Section 6.6.3) are opening a new battlefield that balances performance and verifiability. By building a dedicated Layer-1 blockchain, Hyperliquid achieved sub-second trade confirmation and throughput of roughly 200,000 orders per second, and in 2025, with $2.9 trillion in annual volume, leapt to become the world's seventh-largest perpetual futures exchange, even surpassing established CEXs such as Coinbase International [40].

From four dimensions—core user base, core advantage, market share, and representative platforms—Table 6-8 summarizes the niche each of the three paradigms occupies. This comparison helps explain why a simple *replacement thesis* cannot describe the market's true state, because each paradigm serves different user needs and trading scenarios, and

its core advantage cannot yet be fully replicated by the others under current technical conditions.

Dimension	CEX	General-purpose public-chain DEX	Application-chain DEX
Core user base	High-frequency market makers, retail, institutions	DeFi-native strategists	Mainstream traders, institutions
Core advantage	Low latency, fiat channels, liquidity depth	Composability, permissionless innovation	The optimal intersection of performance and verifiability
2025 market share	Roughly 85%+	Roughly 3%–4% (perpetual futures)	Roughly 8%–9% (perpetual futures)
Representative platforms	Binance, OKX, Bybit	GMX, Synthetix	Hyperliquid, Lighter

Table 6-8. The core niche differences among the three paradigms (Data source: compiled by the author)

Table 6-8 shows that the market-share distribution of the three paradigms corresponds closely to the size and capital volume of their core user bases. The criterion distinguishing a *general-purpose public-chain DEX* from an *application-chain DEX* in the classification above is whether the platform has an independent blockchain and consensus mechanism; for example, after dYdX v4 migrated from a general-purpose public chain (v3 on StarkEx) to an independent Cosmos-based application chain, it was reclassified from the former to the latter. In addition, the volume metric may be inflated by wash trading or incentive-driven exchanges, and open interest (OI) is a more robust indicator of genuine risk-bearing and price-discovery contribution; but because the availability of DEX OI data is limited, volume remains the primary comparison dimension here. The market may ultimately consolidate to one or two dominant platforms in each segment [41]. This means the future competitive landscape may evolve into a bipolar competition between *the top CEX and the top application-chain DEX*, while general-purpose public-chain DEXs continue to play their distinctive role in the composability niche. The coexistence of the three paradigms may constitute a long-term equilibrium of market structure, but this judgment must be treated with caution: the current coexistence may be sustained in part by user habits, platform lock-in effects, and switching costs, which makes it closer to a focal equilibrium sustained by behavioral inertia than to a Nash equilibrium in the strict sense. Should a major CEX regulatory crackdown occur, a technological breakthrough eliminate the application chain's performance advantage, or the liquidity network effect of one paradigm break through a critical point, the current tripolar coexistence could be disrupted.

6.6.2 Convergence dynamics

The evolution of market structure is not a static standoff but a dynamic process of mutual borrowing and fusion. The fundamental driver of this convergence is precisely the competitive compression of the trust tax that we examined in Chapter 1 (Section 1.4.3, *the layered hybrid of the three trust paradigms*) and Chapter 3 (the *institutional*

convergence thesis). To attract and retain liquidity, both CEXs and DEXs strive to make up for their own structural shortcomings, moving toward a middle ground of *lower trust tax and higher verifiability*.

The trend of CEXs moving closer to on-chain models accelerated markedly after the FTX collapse. To rebuild market trust, mainstream CEXs widely introduced cryptography-based proof of reserves, seeking to add a layer of verifiability outside their opaque internal ledgers. After the FTX event, proof of reserves (PoR) became standard practice for leading CEXs, but current PoR implementations have several known limitations: most are point-in-time snapshots rather than continuous audits, usually cover only the asset side without liability information, and cannot prevent *window dressing*—that is, an exchange temporarily borrowing assets before the snapshot to meet the proof requirement. The analysis of Vidal-Tomas (2025) further points out that in the absence of independent audit verification, the information content of PoR may be far lower than the market expects [42]. A deeper transformation is reflected in the evolution of asset-custody models: between 2024 and 2025, Binance actively promoted its keyless, seedless, multi-chain-supporting Web3 self-custodial wallet, allowing users to seamlessly access CEX liquidity while retaining control over their assets [43]. Institutional-grade custody services (such as Coinbase Prime) are exploring hybrid-custody and on-chain-settlement mechanisms, seeking to move toward the *transparency* axis within the impossible triangle. The essence of these moves is that CEXs, without giving up their performance advantage, are actively absorbing the verifiability elements of the on-chain ecosystem.

At the same time, DEXs are moving closer to institutional structure. As the scale of on-chain trading has grown and regulatory scrutiny has intensified, the fully *permissionless* ideal is giving way to an *orderly open market*. Protocols such as dYdX have begun implementing geo-fencing at the front end to restrict access by users from sanctioned regions [44]. In February 2026, Hyperliquid formally established a policy center, marking the beginning of leading on-chain protocols' active participation in exploring and shaping compliance frameworks [45]. Whether front-end geo-fencing is legally sufficient to avoid jurisdiction, however, remains fundamentally in doubt: the position of the SEC and the CFTC is that a protocol remains within their jurisdiction as long as users can access it through a VPN or by interacting directly with the smart contract. The CFTC's 2023 enforcement action against Ooki DAO established an important precedent: a DAO can be held legally liable as an *unincorporated association*. More subtly, the ability Hyperliquid demonstrated in the JELLY incident to unilaterally delist an asset and force liquidation can itself be cited by regulators as evidence that the protocol has centralized control and therefore needs to register. These legal issues are analyzed in depth in Chapter 26. At the on-chain-monitoring level, tools such as AMLBot have already extended to the Hyperliquid blockchain, providing the protocol with anti-money-laundering capabilities. At the protocol layer, DEXs have also begun introducing stricter position-size limits and governance of risk-control parameters, evolving from *fully permissionless* toward an *orderly open market* to prevent systemic collapse under extreme market conditions.

This two-way convergence is not merely a fusion of technology and products but a deep blending of two ecosystem cultures. The CEX ecosystem, originating in traditional

finance, emphasizes efficiency, liquidity, and top-down innovation, presenting a pyramidal participant structure; the DEX ecosystem, originating in the crypto community, prizes transparency, asset sovereignty, and bottom-up evolution, presenting a flatter participant structure. As traditional financial institutions participate deeply in on-chain market making (Jump Trading invested in Firedancer, a Solana validator client, and Cumberland became a data provider for the Pyth oracle) and CEXs continuously absorb on-chain elements, the boundary between the two is blurring.

> Both are moving toward a middle ground of *lower trust tax plus higher verifiability*. The driver of convergence is the competitive compression of the trust tax (Chapters 1 and 3). The ultimate result may be a *hybrid model*, in which different trust paradigms are chosen at different functional layers—precisely the *layered hybrid* predicted in Section 1.4.3, for which Chapter 28 will design a target architecture.

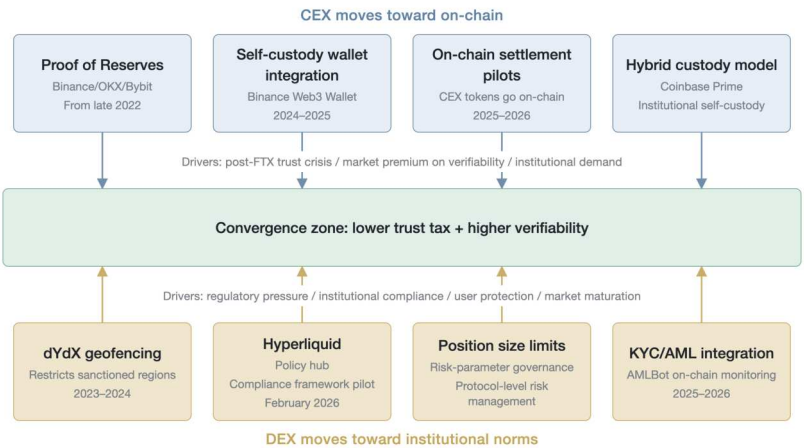


Figure 6-15. The two-way convergence trend of the CEX and DEX ecosystems (a conceptual convergence timeline, with event points through 2024 to 2026, not empirical data; Data source: Binance [43], AInvest [45], and dYdX Terms of Use [44])

Figure 6-15 presents the concrete paths of this two-way movement: the upper half shows the four paths by which CEXs move toward on-chain verifiability—through proof of reserves, self-custodial wallets, and the like—while the lower half shows the four paths by which DEXs move toward institutional compliance—through geo-fencing, policy centers, and the like—with the two sets of paths meeting in the middle ground of *lower trust tax and higher verifiability*. This convergence trend indicates that the next generation of market structure may no longer be a simple either-or choice between CEX and DEX but a layered hybrid architecture that chooses different trust paradigms at different functional layers.

6.6.3 Expanding the impossible triangle

In Section 5.1.3 of Chapter 5, we examined the *impossible triangle* that constrains the architectural design of an exchange: performance, transparency/verifiability, and

decentralization. The traditional view holds that a system can choose only two of these three dimensions and must sacrifice the third. Technological progress, however, is attempting to *expand the area of the triangle*, so that dimensions that once had to be sacrificed can be achieved simultaneously at a higher level. Four frontier technologies are currently reshaping this boundary.

Zero-knowledge proofs are the most transformative of these technologies. ZK allows one party to prove to another the correctness of a computation without revealing its details. In the exchange context, this means fully verifiable matching and liquidation can be achieved while keeping trading strategies private and on-chain data costs extremely low. ZK essentially creates a new dimension on the *transparency* axis—*verifiable but not transparent*—breaking a combination that was impossible in the traditional triangle. Its scope of application is expanding from proof of reserves to verifiable off-chain computation and cross-chain communication, providing a mathematical-grade guarantee for enhancing privacy and interoperability. The zero-knowledge-proof (ZKP) market is projected to expand from $1.535 billion in 2025 to $7.586 billion in 2033, at a compound annual growth rate of 22.1% [46].

Intent-based trading represents a fundamental shift in the execution paradigm. Under this architecture, users no longer submit specific trading instructions but express high-level execution goals, and a network of professional solvers competitively finds the optimal execution path [47]. This paradigm shifts the complexity of execution from the user side to professional nodes and has already been validated in the automated market maker (AMM) spot-swap scenario through protocols such as UniswapX and CoW Protocol, effectively reducing slippage and the negative impact of MEV. The migration of the intent-based architecture to the perpetual futures scenario, however, faces additional complexity: perpetual futures involve state-dependent operations such as margin management, the position lifecycle, and funding-rate settlement, which are hard to reduce to stateless *intent* expressions. When handling the opening, adjustment, and closing of a leveraged position, a solver must assess margin adequacy and liquidation risk in real time, a complexity far exceeding the single-path optimization of a spot swap. Intent-based trading in the perpetual futures domain is therefore more likely to be applied first to peripheral links such as order routing and cross-platform optimal execution, rather than replacing the core order-book matching logic. In essence, this technology raises system efficiency on the *performance* axis through fine-grained division of labor, but its boundary of applicability in the derivatives scenario still awaits testing in practice.

Modular blockchains, by separating data availability, the consensus mechanism, and the execution layer into different specialized networks (such as Celestia or EigenDA), allow application-chain DEXs to outsource some core functions to specialized underlying infrastructure. This architecture enables a protocol to achieve performance far beyond that of a monolithic blockchain without sacrificing underlying security, raising the system's overall security on the *decentralization* axis through specialized division of labor.

The proliferation of account abstraction narrows the gap between on-chain and CEX at the user-experience level. With the integration of EIP-7702 in Ethereum's Pec-

tra upgrade in May 2025, the adoption of smart-contract wallets accelerated markedly [48]. Account abstraction greatly lowers the threshold for on-chain interaction, allowing users to use features such as gasless transactions, social recovery, and batch operations. Without sacrificing the core principle of self-custody, it brings the on-chain experience remarkably close to that of a CEX, thereby narrowing the gap on the *performance/-convenience* dimension.

Table 6-9 systematically summarizes the four technologies above along the impossible-triangle dimension they expand, their core mechanism, and their representative applications. It shows that the four technologies do not each exert force in isolation on a single dimension; rather, each pushes outward along a different axis of the impossible triangle, and their fusion in real projects is redefining the boundary of what is possible for market structure.

Frontier technology	Dimension expanded	Core mechanism	Representative application
Zero-knowledge proofs	Transparency/verifiability	A new dimension of *verifiable but not transparent*	Lighter ZK-Rollup, proof of reserves
Intent-based trading	Performance	User-solver division of labor, reducing MEV	UniswapX, CoW Protocol
Modular blockchains	Decentralization	Separation and specialization of DA/-consensus/execution	Celestia, EigenDA
Account abstraction	Performance/convenience	Smart-contract wallets, approaching the CEX experience	EIP-7702, Safe

Table 6-9. The mechanisms by which four frontier technologies expand the dimensions of the impossible triangle (Data source: compiled by the author)

As Table 6-9 shows, the four technologies each push outward along a different axis of the impossible triangle: zero-knowledge proofs create the new possibility of *verifiable but without exposing details* on the transparency axis; intent-based trading and account abstraction raise execution efficiency and user experience along the performance axis; and modular blockchains strengthen underlying security along the decentralization axis through specialized division of labor. These technologies do not work in isolation but present a fusion trend in real projects, and Lighter's case is a representative practice of this fusion. Lighter's rise in 2025 is a representative case of these frontier technologies reshaping market structure. As an application-specific ZK-Rollup built on Ethereum, Lighter, with $1.3 trillion in annual volume, joined the ranks of the world's tenth-largest perpetual futures exchange in 2025 and completed a $68 million financing round at a $1.5 billion valuation in November 2025 [40][49]. This figure, however, warrants the same scrutiny applied to Hyperliquid's depth data in Section 6.4.3: as a zero-fee protocol, Lighter faces a major concern about volume inflated by wash trading and incentive-driven trading, and zero fees may attract large amounts of high-frequency noise trading rather than genuine price-discovery flow. The *subsidy-driven versus market-structure-driven* analytical framework advanced in Section 6.4.3 applies equally here: under a zero-fee model, whether volume can be sustained after a business-model transition

(the introduction of fees) is a key test of its market-structure significance. Its core innovation lies in a custom proving engine that can generate execution proofs for hundreds of thousands of transactions in parallel and compress them, through multi-layer aggregation, into a single batch proof submitted to the Ethereum mainnet for verification, achieving an effective combination of high-speed off-chain order-book matching and mathematical-grade on-chain security. User assets are always custodied by an Ethereum smart contract, and an *escape hatch* mechanism guarantees that in extreme situations users can withdraw assets directly on Layer-1. The verification a ZK proof provides is retroactive: it confirms already-executed transactions rather than preventing sequencer misbehavior in real time; during proof generation and L1 confirmation, the system in fact operates under the same trust assumptions as a centralized sequencer. Although the *escape hatch*, through forced L1 withdrawal, provides an ultimate safety guarantee, in practice it may take hours to days to complete. In contrast to application chains that rely on their own small validator networks (such as Hyperliquid), Lighter anchors ultimate security to Ethereum's settlement layer, and its combination of zero fees, non-custody, and verifiable matching demonstrates how ZK technology can redefine the boundaries of a decentralized exchange.

> These four technologies and the Lighter case together show that the impossible triangle is not fixed; its area can be expanded by technological progress, so that dimensions that once had to be sacrificed can be achieved simultaneously at a higher level. The triangle does not disappear, but its area grows—and that is the meaning of technological progress.

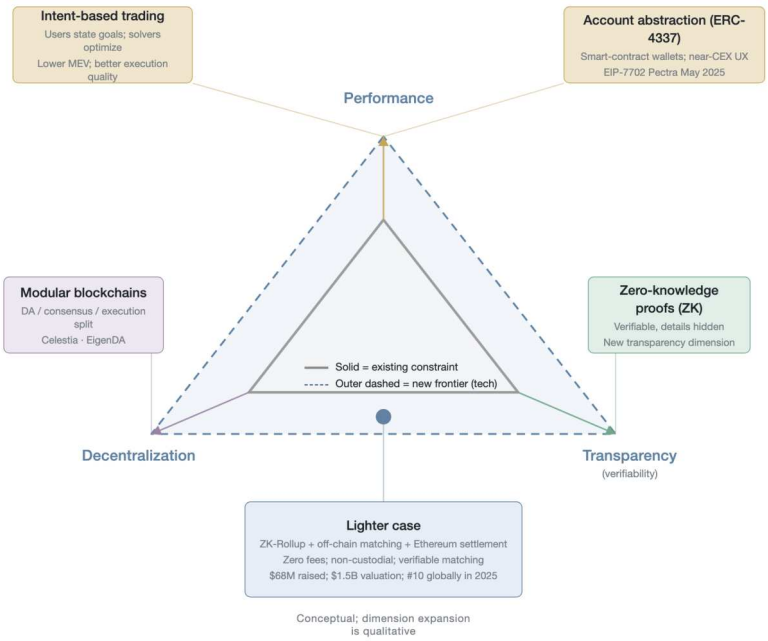


Figure 6-16. The mechanisms by which four frontier technologies expand the impossible triangle, with the Lighter case (a conceptual schematic, with dimension expansion represented qualitatively, not empirical data; Lighter's $68 million financing and $1.5 billion valuation [49] and its position as the world's tenth-largest perpetual futures exchange in 2025 [40][49] are both marked with sources; drawn by the author, based on Grand View Research [46], Paradigm [47], and CoinDesk [49])

Figure 6-16 integrates the analysis above into a unified visual framework: the inner triangle represents the existing constraints under current technical conditions, and the outer dashed triangle represents the new boundary that technological progress can push out; Lighter, as a fusion of ZK-Rollup, high-speed off-chain matching, and Ethereum settlement, is marked as the practical case closest to the outer boundary at present, validating the practical feasibility of technological fusion expanding the impossible triangle.

6.6.4 Open problems

Although technological progress keeps expanding the boundary of the possible, the evolution of the next generation of market structure still faces two fundamental unresolved problems, whose solutions will define the ultimate form of the perpetual futures market.

On the problem of the effectiveness boundary of decentralization, an application chain pursuing high performance inevitably must compromise on the scale of its validator network. Take Hyperliquid: although it achieved rapid growth in 2025 (users grew from 300,000 to 1.4 million, and annual volume reached $2.9 trillion), its number of validators remained at just 16 to 24 for a long time [50][51]. By early 2026, the number of validators had increased to 24; compared with Ethereum's hundreds of thousands of validating nodes, this is still a highly concentrated consensus network [9]. This concentration has raised persistent community concern about censorship resistance and

the potential risk of collusion.

Decentralization is not a black-and-white binary state but a continuous variable. This raises a profound question of governance philosophy: *How much decentralization is enough?* The answer often depends on the specific application scenario and the user's risk preference. There is a fundamental physical tension between the scale of the validator network and system performance: more validators mean higher security and censorship resistance but inevitably lead to increased consensus latency, while fewer validators can provide a CEX-comparable sub-second experience but introduce a stronger trust assumption. How to find that dynamic *effectiveness boundary* between performance and decentralization is a challenge every application chain must confront over the long term. The regulatory dimension of this problem is explored further in Chapter 26.

Parallel to the decentralization-boundary problem, the second problem is the trilemma of tension among transparency, leverage, and security, which is the ultimate extension, at the market-structure level, of the *transparency-fragility paradox* examined in Section 4.4.3 of Chapter 4. In on-chain perpetual futures markets, a fully transparent order book, leverage ratios as high as 50x to 100x, and algorithmic high-frequency execution together constitute an extremely fragile complex system. Application chains have mitigated traditional atomic MEV through protocol design, but the fully transparent liquidation price instead gives rise to strategic MEV, a deeper form of attack, as analyzed in detail in Section 6.3.

The three strategic attacks that occurred consecutively on Hyperliquid from March to November 2025—the whale's self-liquidation after building a roughly $300 million ETH long position at 50x leverage, the JELLY token attack, and the November cross-market price manipulation—provide the most vivid footnote to this tension (the loss magnitudes and mechanisms of the three incidents are detailed in Section 6.3.2).

These events lay bare the trilemma of tension among transparency, high leverage, and system security. Under the combination of *transparency plus high leverage*, the liquidation levels and liquidity depth of all participants are fully visible to attackers, making precise sniping possible and giving rise to strategic liquidation attacks. Under the combination of *high leverage plus security*, the system must possess extremely strong anti-manipulation capability and a deep insurance fund; otherwise, shortfall risk will inevitably be socialized, transferred through auto-deleveraging or vault losses onto innocent liquidity providers. And the more fundamental question is whether 50x to 100x leverage matches the actual loss-absorption capacity of an on-chain protocol. The maximum leverage of traditional futures markets (such as CME and Eurex) is constrained by clearinghouse margin requirements based on decades of volatility data, backed by a mutualized default fund and an explicit central-counterparty guarantee. That on-chain perpetual futures protocols offer equal or even higher leverage in the absence of comparable loss-absorption capacity (both the size of the insurance fund and the depth of the HLP vault being far smaller than a traditional CCP's default fund) makes the reasonableness of their risk pricing questionable. Under the combination of *transparency plus security*, the most counterintuitive insight emerges: in a high-leverage financial game, full transparency may not be net positive. To protect system security, some form of

selective opacity (for example, using ZK technology to achieve *verifiable but without exposing the liquidation threshold*) may be safer than unprotected full transparency. Traditional finance already has mature precedents for such calibrated transparency: when the TRACE (Trade Reporting and Compliance Engine) system in the U.S. corporate bond market introduced post-trade transparency, it imposed delayed reporting on large trades to protect institutional investors (Bessembinder, Maxwell, and Venkataraman, 2006 [52]); the EU's MiFID II framework likewise provides transparency waivers for large trades. These experiences show that *selective opacity* is not a new invention of the crypto market but a well-tested institutional arrangement in financial-market design.

This insight connects the strategic-MEV analysis of Section 6.3, the ZK-technology discussion of Section 6.6.3, and the embedded-compliance design of Chapter 30 into a complete logical chain. Defending against strategic MEV cannot rely solely on increasing matching speed (this is the mental inertia of latency competition) but must involve a fundamental restructuring at the mechanism-design level, which also lays the theoretical foundation for the book's later discussions of the liquidation mechanism (Chapter 11), insurance-fund governance (Chapter 12), and *strategic-MEV-immune* clearing and settlement architecture design (Chapter 29).

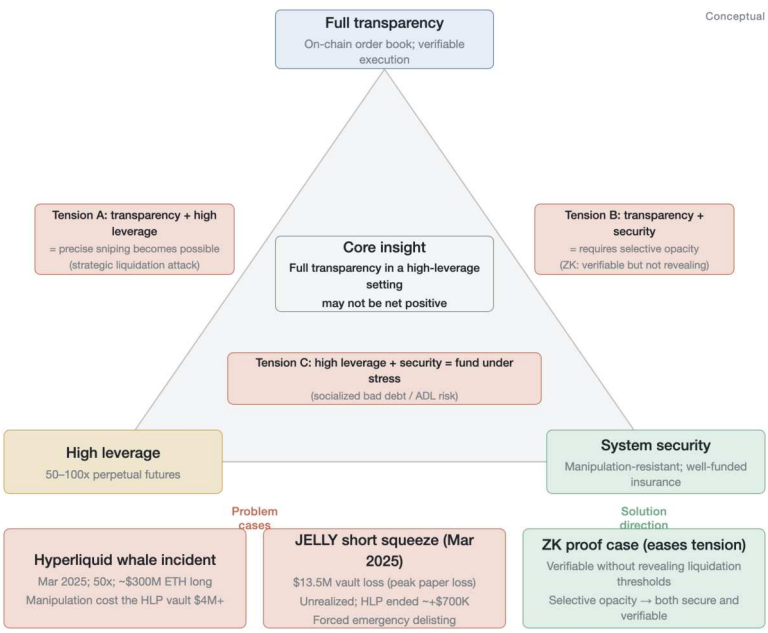


Figure 6-17. The inherent tension among transparency, high leverage, and system security (a conceptual-triangle schematic, with vertex positions representing qualitative tension relationships, not empirical coordinates; the roughly $13.5 million for JELLY is a peak unrealized loss, not a realized loss, with the HLP vault ultimately recording a small profit of about $700,000 [37]; the whale roughly $300 million, HLP roughly $4 million [20], and 50x to 100x leverage are all non-empirical data; Data source: Halborn [25], CoinDesk [37], and Arkham Intelligence [20])

Figure 6-17 visualizes this trilemma of tension as a triangular structure, whose three

edges correspond to three sets of constraint relationships: the combination of transparency and high leverage gives rise to precise liquidation sniping (edge A), the combination of high leverage and security requires a deep insurance-fund reserve (edge C), and the combination of transparency and security points to the counterintuitive conclusion that in a high-leverage environment *selective opacity* may be preferable to full transparency (edge B). The bottom of the figure contrasts the three real 2025 attack cases with the technical logic of ZK proofs as a potential solution path, connecting the problem diagnosis and the engineering solution into a complete logical chain.

6.7 Chapter summary

This chapter concludes Part II, *The Ecosystem: The Trading Ecosystem of Perpetual Futures*. Chapter 4 mapped six classes of participants and advanced the *transparency-fragility paradox*; Chapter 5, through a static dissection of the three architectures, advanced the *risk substitution theorem*; and this chapter pushes the perspective toward dynamic evolution, answering the question of *how different structures compete and converge*. The chapter established four core theoretical contributions. The competition-mode migration thesis argues that the evolution of market structure is not a simple case of *decentralization replacing centralization* but a paradigm shift in the focus of competition—from the latency competition of CEXs, to the information competition of general-purpose public-chain DEXs, to the mechanism competition of application-chain DEXs—with each migration driven by the diminishing economic returns of the previous mode. The definition of strategic MEV elevates MEV from within-block technical arbitrage to a cross-cycle, cross-market economic game, pointing out that its defense can rely only on mechanism redesign, not a speed race. The price-discovery migration hypothesis holds that three structural drivers—the trust premium of verifiable liquidity, the efficiency gains of protocol-level market making, and the on-chain migration of institutions—are shifting pricing power toward application-chain DEXs, and that once the liquidity network effect crosses a critical mass, this migration may become irreversible. The risk visibility-manageability proposition holds that, from CEXs to on-chain venues, risk shifts from *invisible but occasionally catastrophically explosive* to *continuously visible but manageable*, with visibility a necessary precondition for risk management.

After completing this macro analysis of the competition among the three architectures, one key question remains unresolved: How does a macro structural advantage translate into micro trade execution? How do orders express traders' intent, how does the order book aggregate information and form prices, and how do differences in matching rules affect market liquidity? Part III—beginning with the anatomy of the order (Chapter 7), the microdynamics of the order book (Chapter 8), and Hyperliquid's architectural innovations (Chapter 9)—establishes the micro foundations for the subsequent quantitative analysis of price discovery, arbitrage, and liquidity.

References

[1] Reuters. (2022, November 11). FTX collapse: How the crypto exchange lost billions.

Reuters. https://www.reuters.com/technology/ftx-crypto/

[2] European Securities and Markets Authority. (2025). "Maximal Extractable Value Implications for crypto markets." https://www.esma.europa.eu/sites/default/files/2025-07/ESMA50-481369926-29744_Maximal_Extractable_Value_Implications_for_crypto_markets.pdf

[3] CoinGecko. (2026). "2025 Annual Crypto Industry Report." https://www.coingecko.com/research/publications/2025-annual-crypto-report

[4] CoinGecko Research. (2025). "DEX to CEX Volume Ratios Reach New Highs in 2025." https://www.coingecko.com/research/publications/dex-to-cex-ratio

[5] Coinage Media. (2026). "How Hyperliquid Is Looking to Maintain Its Lead Among Perp DEXes." https://www.coinage.media/2026/how-hyperliquid-is-looking-to-maintain-its-lead-among-perp-dexes

[6] Harris, L. (2003). *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press.

[7] Crawley, J. (2025, February 21). Bybit suffers $1.5 billion security breach in largest crypto hack. *CoinDesk*. https://www.coindesk.com/business/2025/02/21/bybit-suffers-1-5-billion-security-breach/

[8] DeFiLlama & Rekt. (2025). Cross-chain bridge exploits leaderboard [Database]. DeFiLlama (https://defillama.com/hacks) and Rekt News (https://rekt.news/leaderboard/). Data as of 2025; the statistical basis varies with how cross-chain bridges are defined and how incidents are classified.

[9] Eco.com. (2026). What is Hyperliquid? The High-Performance DEX Explained. https://eco.com/support/en/articles/11972709-what-is-hyperliquid-the-high-performance-dex-explained

[10] Chen, E., Ma, Y., & Nie, Z. (2024). Perpetual future contracts in centralized and decentralized exchanges: Mechanism and traders' behavior. *Electronic Markets*, *34*(1). https://doi.org/10.1007/s12525-024-00715-1

[11] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100. https://doi.org/10.1016/0304-405X(85)90044-3

[12] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

[13] Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, *130*(4), 1547–1621. https://doi.org/10.1093/qje/qjv027

[14] Aquilina, M., Budish, E., & O'Neill, P. (2022). Quantifying the high-frequency trading arms race. *The Quarterly Journal of Economics*, *137*(1), 493–564. https://doi.org/10.3386/w29011

[15] Amazon Web Services. (2025). Optimize tick-to-trade latency for digital assets exchanges and trading platforms on AWS.

[16] O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial*

Economics, *116*(2), 257–270. https://doi.org/10.1016/j.jfineco.2015.01.003

[17] Foucault, T., Kadan, O., & Kandel, E. (2013). Liquidity cycles and make/take fees in electronic markets. *The Journal of Finance*, *68*(1), 299–341. https://doi.org/10.1111/j.1540-6261.2012.01801.x

[18] CryptoSlate. (2026). Crypto privacy has turned into an economic crisis as MEV bots burn over 50% of gas fees. (Citing EigenPhi data for Dec 2025 - Jan 2026).

[19] Mancino, D., & Rezzoli, D. (2025). Sandwiched and silent: Behavioral adaptation and private channel exploitation in Ethereum MEV. arXiv preprint arXiv:2512.17602.

[20] Arkham Intelligence. (2025, March 12). Hyperliquid Whale Passes $4M Loss to HLP Vault.

[21] Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, *30*(5), 15–29. https://doi.org/10.3905/jpm.2004.442611

[22] Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, Y., Bentov, I., Breidenbach, L., & Juels, A. (2020). Flash boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. *2020 IEEE Symposium on Security and Privacy (SP)*, 910–927. https://doi.org/10.1109/sp40000.2020.00040

[23] Gramlich, V., Guggenberger, T., Principato, M., Schellinger, B., & Urbach, N. (2024). A systematic literature review of maximal extractable value. *Electronic Markets*, *34*, 58.

[24] Talos. (2025). $JELLY's Last Jam: Adventures in Hyperliquid Margin Risk Management. Talos Insights.

[25] Halborn. (2025). Explained: The Hyperliquid Hack (November 2025). https://www.halborn.com/blog/post/explained-the-hyperliquid-hack-november-2025

[26] Obadia, A., Salles, A., Sanber, L., Tribble, T., Rao, V., & Juels, A. (2021). Unity is strength: A formalization of cross-domain maximal extractable value. arXiv preprint arXiv:2112.01472. https://doi.org/10.48550/arXiv.2112.01472

[27] Capponi, A., Jia, R., & Wang, Y. (2024). The ordering of blockchain transactions and its impact on market quality. *Management Science*.

[28] Michaels, D., & Osipovich, A. (2023, October 5). Here's how FTX executives secretly spent $8 billion in customer money. *The Wall Street Journal*. https://www.wsj.com/finance/currencies/ftx-customer-money-spending-39439e3c

[29] Hyperliquid Team. (2026). "Order Book Depth Comparison: Hyperliquid vs. Centralized Exchanges." Public data release, January 2026.

[30] Hayward, A. (2022, August 16). Jump Crypto plans new Solana validator client to boost performance, decentralization. *Decrypt*. https://decrypt.co/107521/jump-crypto-plans-solana-validator-client-boost-performance-decentralization

[31] Zhivkov, P. (2026). The two-tiered structure of cryptocurrency funding rate markets. *Mathematics*, *14*(2), 346. https://doi.org/10.3390/math14020346

[32] Monarq Asset Management. (2026). "Perp DEXs in 2025: The Shift From Subsidies to Market Structure." Medium. https://medium.com/@Monarq_Mgmt/perp-

dexs-in-2025-the-shift-from-subsidies-to-market-structure-68a1138f4c10

[33] Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *The Journal of Finance*, *50*(4), 1175–1199. https://doi.org/10.1111/j.1540-6261.1995.tb04054.x

[34] Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House.

[35] CNBC. (2025). "Hackers Steal $1.5 Billion from Exchange Bybit in Biggest-Ever Crypto Heist." CNBC, February 2025.

[36] BlockEden. (2025). "The JELLY Exploit Exposed Hyperliquid's Centralization." https://blockeden.xyz/forum/t/the-jelly-exploit-exposed-hyperliquids-centralization/406

[37] CoinDesk. (2025). HyperLiquid Delists JELLY After Vault Squeezed in $13M Tussle. https://www.coindesk.com/markets/2025/03/26/hyperliquid-delists-jellyjelly-after-vault-squeezed-in-usd13m-tussle

[38] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn014

[39] FTI Consulting. (2025). "Crypto Crash Oct 2025: Leverage Meets Liquidity." https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity

[40] CoinGecko. (2026). Is the Future of Crypto Perpetual? The Meteoric Rise of Perp DEXs. CoinGecko Learn. https://www.coingecko.com/learn/rise-of-perpetuals-and-perp-dexs

[41] Grayscale Research. (2025). DEX Appeal: The Rise of Decentralized Exchanges. https://research.grayscale.com/reports/dex-appeal-the-rise-of-decentralized-exchanges

[42] Vidal-Tomas, D. (2025). Centralized Crypto Exchanges and Proof-of-Solvency. Working Paper.

[43] Binance. (2025). Everything You Need to Know About Our Self-Custody Web3 Wallet. Binance Blog.

[44] dYdX. (2025). Terms of Use & Geo Restrictions. dYdX Help Center.

[45] AInvest. (2026). Hyperliquid's Strategic Expansion in Perpetual Futures Trading. https://www.ainvest.com/news/hyperliquid-strategic-expansion-perpetual-futures-trading-assessing-leverage-token-utility-market-capture-2026-2601/

[46] Grand View Research. (2025). Zero Knowledge Proof Market Size, Share & Trends Analysis Report, 2025-2033.

[47] Paradigm. (2023). Intent-Based Architectures and Their Risks. Paradigm Research.

[48] Alchemy. (2025). What is ERC-4337? The Complete Guide to Account Abstraction.

[49] CoinDesk. (2025). Lighter Raises $68M at $1.5B Valuation. https://www.coindesk.com/business/2025/11/04/lighter-raises-68m-at-1-5b-valuation/

[50] The Block. (2025). Hyperliquid responds to community concerns over validator issues. https://www.theblock.co/post/333559/hyperliquid-responds-to-community-concerns-over-validator-issues

[51] Ian Unsworth. (2025). Hyperliquid validator set expansion from 21 to 24. X/Twitter.

[52] Bessembinder, H., Maxwell, W., & Venkataraman, K. (2006). Market transparency, liquidity externalities, and institutional trading costs in corporate bonds. *Journal of Financial Economics*, *82*(2), 251–288. https://doi.org/10.1016/j.jfineco.2005.10.002

[53] CoinDesk. (2023, June 26). FTX bankruptcy team says the exchange owed customers $8.7B. *CoinDesk*. https://www.coindesk.com/policy/2023/06/26/ftx-bankruptcy-team-says-the-exchange-owed-customers-87b

[54] Kobayashi, N. (Mt. Gox Civil Rehabilitation Trustee). Announcements regarding the rehabilitation proceedings of MtGox Co., Ltd. *Mt. Gox*. https://www.mtgox.com/

Part III
Microstructure

Core question: In a perpetual futures market, how does a trade go from intent to execution, and how is price discovered and formed at the microstructural level?

Orders and the Language of Perpetual Futures Trading

In today's crypto derivatives markets, buying a large position—say, a $10 million Ethereum long—is really a multilayered act of transmitting financial information. How this order is ultimately handled depends on the execution architecture chosen. If the limit order is sent to a centralized exchange such as Binance or OKX, it reaches the exchange's private servers instantaneously over an encrypted API. Within the matching engine's memory, the order is processed in sub-millisecond time. If no matching sell order exists in the current order book, it rests there. Apart from the exchange's internal systems and the handful of high-frequency market makers who can observe partial information through private data feeds, the market as a whole cannot access this $10 million buying intent. The trader can cancel the order at any time at zero cost, leaving no public record.

If the trader instead submits the identical order on an on-chain central limit order book protocol such as Hyperliquid or dYdX, the processing flow differs fundamentally. Once signed by the private key, the order is broadcast to the application chain's validator network. Roughly 200 milliseconds later, it is packed into a block and becomes an immutable on-chain commitment, displayed publicly to the entire world. Every full-node operator, on-chain analytics tool, and arbitrage bot can immediately "see" this $10 million buying intent. If the order does not fill immediately, it hangs publicly on the on-chain order book, becoming a focal point for strategic maneuvering among all parties. To cancel it, the trader must submit another on-chain transaction, which not only incurs a network fee but also endures the latency of block confirmation, and every attempt and withdrawal is permanently recorded as historical data.

These two starkly different processing paths reflect the ongoing evolution of the crypto derivatives market microstructure. In 2025, the crypto derivatives market recorded average daily trading volume of approximately $264.5 billion [1]. Within this vast market, perpetual futures have become the instrument with the largest share of trading volume: they have not only surpassed the trading scale of the spot market but have also become the core mechanism for pricing and liquidity discovery across the entire crypto-asset space. Decentralized derivatives exchanges also grew rapidly over the same period. Driven by high-performance application chains such as Hyperliquid, decentralized exchange (DEX) perpetual futures volume surged 346% in 2025 to reach $6.7 trillion [2]. Hyperliquid in particular reached $2.9 trillion in annual volume, surpassing compliant

giants such as Coinbase International, and in October set a historical record by processing more than $300 billion in volume in a single month [2]. The DEX share of the perpetual futures market jumped from less than 4% at the start of the year to more than 13% by year-end—a structural shift that marks the arrival of on-chain derivatives trading in the mainstream market.

Migrating complex derivatives trading and high-frequency order books from private servers onto public blockchains, however, is not merely a straightforward replacement of infrastructure. It constitutes a reconstruction at the level of market microstructure. In an on-chain environment, an order is no longer merely an expression of trading intent; it becomes a public signal with high information content in its own right. Starting from the theoretical framework of "the order as a signal," this chapter analyzes information leakage and adverse selection in the transparent on-chain environment, the defensive mechanisms market makers deploy within the order book, the "intent-execution gap" faced by ordinary traders, and—progressing from basic to advanced—the toolbox of order types, the three-dimensional strategic extension of perpetual futures, and how the physical constraints of blockchain reshape the economic meaning of orders.

7.1 The signaling function of orders

In the microstructure theory of financial markets, the order is the most fundamental building block [3]. Traditional textbooks tend to treat the order as a purely operational tool ("a market order means immediate execution at the current price; a limit order means resting at a specified price and waiting"). This understanding is necessary but far from sufficient. In the environment of modern high-frequency trading and transparent on-chain ledgers, every order submitted to the market—whether or not it ultimately fills—transmits rich information to the market. This section establishes a conceptual framework: every order simultaneously performs a "trading function" and a "signaling function," and the trader's core strategic challenge is to find the optimal balance between the two. This "order as a signal" framework runs through the entire chapter and provides a unified explanation of the strategic implications of all order types.

7.1.1 The trading function and the signaling function

In crypto derivatives markets, and especially in the on-chain central limit order book (CLOB), every order has an inseparable dual function: an explicit trading function and an implicit signaling function.

The trading function is the surface-level purpose of an order. When a trader submits a buy order, the direct purpose is to acquire a long position in the target asset at a specific price. This function is deterministic and mechanical, executed strictly by the matching engine according to rules such as price priority and time priority. Whether the trading function succeeds is reflected directly in the trader's account balance and position changes. In the context of perpetual futures, the trading function carries additional dimensions: a buy order not only changes the direction and size of a position but also simultaneously changes the effective leverage, the liquidation price, and the proportion

of margin utilized. These derivative effects make order decisions in perpetual futures an order of magnitude more complex than in the spot market.

The signaling function is the deeper attribute of an order. In a market with information asymmetry, a trader's behavior is itself the largest source of information. The type, price, size, and submission time of an order—and even the frequency with which it is canceled—all leak the trader's private information to other market participants, especially market makers and arbitrageurs. Specifically, the signaling function comprises the following dimensions:

Signal dimension	Information leaked	Inference logic
Order direction	The trader's view on price movement	A buy order implies a bullish expectation; a sell order implies a bearish expectation
Order type	The trader's degree of urgency	A market order signals extreme urgency; a limit order signals a willingness to wait
Order size	The trader's capital base and level of conviction	A large order suggests an institutional-grade participant or a strong directional belief
Submission time	The timeliness of the information	Orders submitted around major events are more likely to contain private information
Cancellation frequency	The order's true intent	Frequent cancellation may suggest "probing" orders or market-making adjustments

Table 7-1. The five dimensions of the order's signaling function (Data source: compiled by the author)

Table 7-1 systematically lays out the five dimensions of the order's signaling function: a market maker's algorithms analyze an order's direction, type, size, and timing in combination to infer the identity category and information content of the submitter. Figure 7-1 uses a flowchart to show the transmission path of the trading function and the signaling function from order submission to the trader's optimization objective.

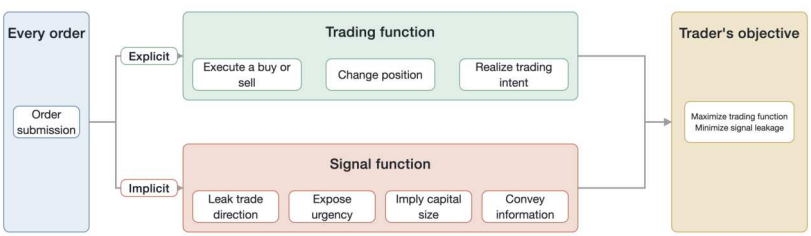


Figure 7-1. The dual function of an order (conceptual/mechanism flow illustration, not empirical data)

This recognition of a dual function can be formalized as a concise optimization framework. For any order, the trader's objective function can be stated as:

> Order optimization objective = maximize (trading-function efficiency) × minimize (degree of signal leakage)

The framework above is a normative benchmark: it describes how a rational trader ought to choose an order type, not how actual traders in fact behave. A large body of empirical research shows that retail traders' order choices are often driven by behavioral factors such as platform UI defaults, urgency bias, and the availability heuristic, and thus systematically deviate from this theoretical optimum. Different order types offer different Pareto-optimal solutions within this objective function. A market order pushes trading-function efficiency to its limit (immediate execution) but also maximizes signal leakage; an iceberg order minimizes signal leakage but sacrifices execution speed and certainty. Understanding this trade-off is the theoretical foundation for understanding all of the order-type analysis that follows.

On a centralized exchange, the signaling function is partly masked by dark pools, iceberg orders, and the exchange's internal data-segregation mechanisms. A trader's limit order is visible only to the exchange and to authorized counterparties, cancellation leaves no public trace, and the signal's range of propagation is strictly confined within the walls of the exchange. In an on-chain CLOB protocol, however, the global visibility of the blockchain ledger dramatically amplifies the order's signaling function. Any participant able to run a full node or access an RPC endpoint can monitor every minute change in the order book in real time. In this fully transparent market, every order transmits information, and it does so to all participants.

7.1.2 The signal-strength spectrum

Not all orders transmit signals of the same strength. Depending on the order type and its parameters, we can construct a "signal-strength spectrum." Along this spectrum, different types of orders exhibit a clear trade-off between execution urgency and the degree of signal leakage.

At one end of the spectrum is the large market order. Orders of this kind transmit strong execution urgency, indicating that the trader is willing to accept any price on the current order book in order to complete the trade immediately. In microstructure theory, market orders are generally regarded as containing the highest proportion of "private information" [4]. Their signal content can be read as: "I need to fill immediately, whatever the slippage, because I may hold time-sensitive material information." When market makers observe a run of one-directional large market orders, they immediately infer that an informed trader is building a position, and they promptly withdraw liquidity or adjust their quotes. Yet the signal from a large market order is not noise-free: panic selling and forced liquidations also generate large market orders, and market makers cannot tell the two apart perfectly in advance—which is precisely the root of the adverse-selection problem.

The middle region of the spectrum holds the various limit orders. Aggressive limit orders (priced very close to, or even better than, the current best bid or offer) transmit a signal strength second only to that of market orders; their implicit message is: "I have a directional view, but I am not especially urgent, and I am willing to wait briefly for a

better price." Such orders seek a balance between urgency and execution price. Passive limit orders placed far from the market price, by contrast, transmit a relatively weak signal; they are seen more as the passive positioning of a liquidity provider—the typical choice of a market maker—or as the "buy-the-dip" strategy of a long-term investor.

At the other end of the spectrum are algorithmic orders (such as TWAP/VWAP) and iceberg orders. These orders are designed from the outset to conceal the trader's true intent and total size. An iceberg order slices a large order into small blocks displayed in sequence, deliberately hiding the true size; time-weighted average price (TWAP) and volume-weighted average price (VWAP) algorithms spread a large order over a longer time window, so that the signal of each small order is so faint as to be nearly indistinguishable from background market noise. Their goal is to split a large order into subordinate orders small enough that each child order's signal strength approaches the level of market noise.

Order type	Signal strength	Signal content	Execution certainty	Typical users
Large market order	Very strong	Urgent need; may hold material information	Highest	Informed traders, panic traders
Ordinary market order	Strong	Has a directional view; needs immediate execution	High	Speculators, trend followers
Aggressive limit order	Moderately strong	Has a directional view; willing to wait briefly	Medium-high	Active traders
Ordinary limit order	Medium	Conviction about value; not in a hurry to fill	Medium	Market makers, value investors
Far-from-market limit order	Weak	Providing liquidity or positioning to buy the dip	Low	Market makers, long-term investors
TWAP/VWAP algorithmic order	Very weak	Spreads size across time	Medium (in tranches)	Institutional traders
Iceberg/hidden order	Very weak	Deliberately conceals true size	Medium-low	Large traders

Table 7-2. The correspondence between order types and signal strength (Data source: compiled by the author)

Table 7-2 reveals a core regularity: the true trade-off axis in order selection is not "signal strength ↔ execution certainty" but rather "execution urgency ↔ price improvement and signal concealment." The large market order sits at the high-urgency end, achieving both the highest execution certainty and the strongest signal leakage (the two move in the same direction at this end); iceberg orders and TWAP sit at the low-urgency end, trading away execution speed in exchange for the weakest signal leakage. Signal strength and execution certainty therefore move together at the passive end and diverge at the active end, rather than forming a simple one-dimensional inverse relationship. (Aggressive limit orders and immediate-or-cancel (IOC) orders, for example, may under

specific conditions combine relatively high signal strength with high execution certainty; moreover, in an on-chain environment, the execution pattern of TWAP/VWAP child orders may be reassembled through pattern recognition by on-chain analytics tools, so that the actual efficiency of signal concealment is lower than the table suggests.) Figure 7-2 arranges the various order types along the two poles of "execution urgency" and "degree of signal leakage."

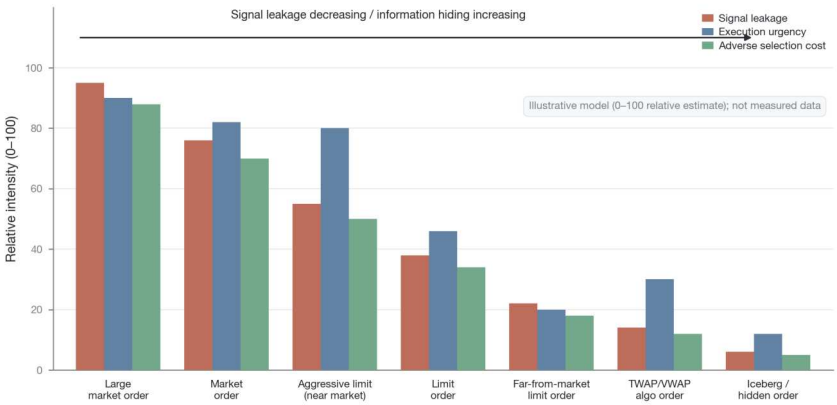


Figure 7-2. The signal-strength spectrum of orders (conceptual model illustration: the 0–100 scale represents relative strength assigned by the author, not empirical data; theoretical basis: Copeland & Galai 1983 [4], Harris 2003 [3], Easley et al. 2012 [29])

> Core strategic implication: traders face a fundamental trade-off between execution urgency (speed) and information concealment (and price improvement), rather than a simple opposition of "execution certainty versus information leakage." A market order has the highest execution certainty but the greatest signal leakage; an iceberg order has the smallest signal leakage but lower execution certainty and greater operational complexity. To choose an order type is to choose one's position within this trade-off.

In an on-chain environment, this trade-off becomes especially acute. Because there is no native privacy-preserving mechanism, a genuine iceberg order is difficult to implement (its transparency tension and formal definition appear in Section 7.3.2, and the mitigation via protocol-level selective opacity appears in Section 7.5.4). On today's mainstream on-chain CLOBs such as Hyperliquid and dYdX, traders rely mainly on off-chain computation and high-frequency order splitting to simulate algorithmic execution; but every child order remains visible on-chain, and a sufficiently sophisticated on-chain analytics tool may still re-correlate these fragments through pattern recognition to infer the trader's true intent. The structural limits the on-chain environment places on signal concealment are the core analytical starting point for understanding the microstructural differences between the on-chain CLOB and the CEX. This section only establishes that conceptual framework; Section 7.5 develops the public-commitment property of on-chain orders in depth from the standpoint of MEV exposure and commitment cost.

7.1.3 The microfoundations of adverse selection

When orders "speak," the most important listener in the market is the market maker. Market makers earn the bid-ask spread by simultaneously providing quotes on both the bid and ask sides; they are the core suppliers of market liquidity. This business model, however, faces one central risk: adverse selection.

The concept of adverse selection was first formalized by Copeland and Galai in their seminal 1983 paper [4]. Glosten and Milgrom (1985) subsequently proved, in a sequential-trade model, that the equilibrium level of the bid-ask spread is determined by the market maker's posterior estimate of the proportion of informed traders [5]; Kyle (1985), starting from the optimal strategy of the informed trader, proved that the price-impact coefficient λ is proportional to the precision of the informed trader's information [6]. Together, these three classic works build the microfoundations of adverse-selection theory. Copeland and Galai's key insight is this: the act of submitting a limit order is, in economic terms, equivalent to giving the market a free option. When a trader posts a buy limit order, they in effect grant anyone in the market the right to sell the asset to them at that price. If the market price remains stable or rises, this "option" is not exercised and the limit order remains unfilled; but if an informed trader learns of information that will cause the price of ETH to crash, they will immediately "exercise" the option, selling their soon-to-depreciate ETH to the limit-order submitter at \$3,000. For the limit-order submitter, being passively filled often means bearing an adverse-selection loss.

As the largest providers of limit orders in the market, market makers face adverse-selection risk continuously. To compensate for this potential loss, market makers must charge an "information premium" to all traders—including uninformed ones—by widening the bid-ask spread. This cost is shared across the entire trading population: the presence of a small number of informed traders systematically raises transaction costs for all participants.

In practice, the way market makers "listen" to orders relies mainly on real-time monitoring of order-flow imbalance. By analyzing the volume difference between aggressive buy orders and aggressive sell orders within a given time window, market makers can infer the direction in which market microstructural momentum is tilting. Research by Bieganowski and Ślepaczuk (2026) confirms that order-flow imbalance is one of the strongest predictive signals of short-term price direction [7]. When a market maker's algorithm detects a sharp rise in buy-side aggression (for example, a succession of large market buy orders), it reacts immediately: widening the ask-side quote and reducing ask-side depth. This is the microstructural mechanism of a "rising price": not some abstract "market force," but rather the market maker "hearing" in the order flow the signal that informed traders are entering, and adjusting risk exposure accordingly. Figure 7-3 formalizes this dynamic process as a closed-loop feedback model: the market-order impact of informed traders triggers the market maker's perception of adverse selection; the market maker defends by widening the spread and withdrawing quotes; and this defensive response in turn further intensifies the market's liquidity contraction, forming a self-reinforcing loop of adverse selection.

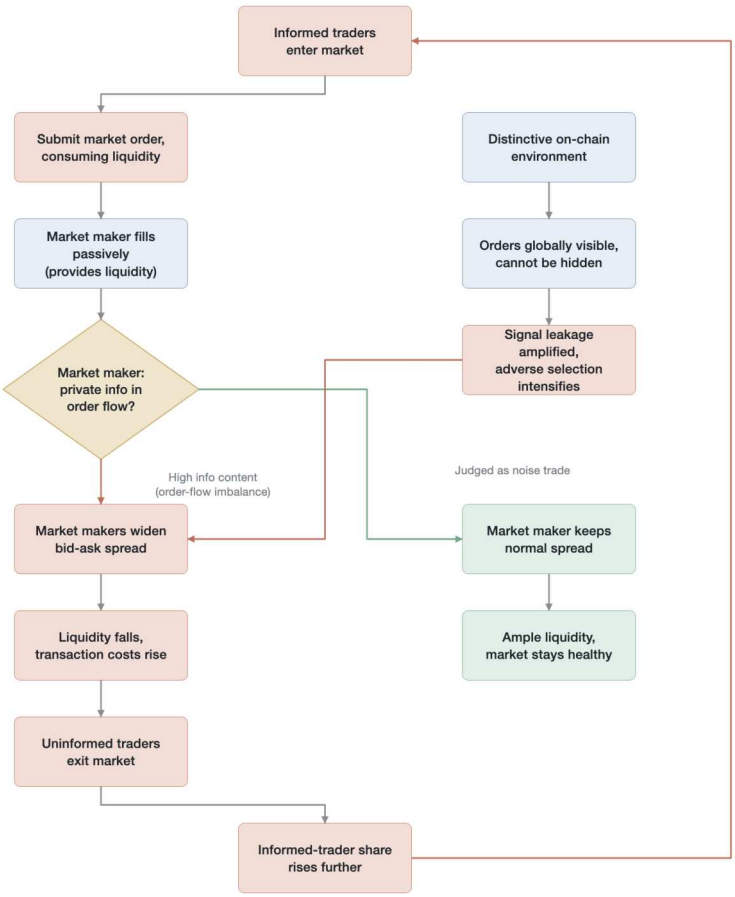


Figure 7-3. The market maker's adverse-selection feedback loop (mechanism feedback-loop illustration, not empirical data; order-flow imbalance is one of the strongest short-term price-prediction signals, per Bieganowski & Ślepaczuk 2026 [7])

Adverse selection in perpetual futures has a distinct amplifying effect. Empirical research by Ruan and Streltsov (2022) finds that a market's transaction costs may actually rise after perpetual futures are introduced [8]. The study's sample period (2020–2022), however, coincided with a period of large-scale retail entry into crypto markets and globally accommodative monetary policy; the rise in transaction costs may partly reflect the dramatic change in the composition of market participants rather than an inherent effect of perpetual futures, and the finding should therefore not be casually generalized as a universal feature of perpetual-futures markets. This seemingly counterintuitive phenomenon is a direct manifestation of an intensified adverse-selection effect. The high-leverage nature of perpetual futures means that an informed trader can command a larger position with less capital, greatly improving the efficiency with which their informational advantage is monetized; at the same time, round-the-clock uninterrupted trading means that market makers must maintain quotes at all times and cannot pause trading after a close as in traditional markets. More fundamentally,

the introduction of perpetual futures creates a new source of adverse selection for spot market makers: informed traders can choose to express their informational advantage in perpetual futures rather than in the spot market, so that the informed component of the order flow facing spot market makers rises and their market-making profit falls. These two factors combined force market makers to widen spreads to compensate for the higher adverse-selection risk.

The extreme market conditions of October 10, 2025, provide a representative observation sample. On that day, amid a macro geopolitical shock, the crypto market suffered a flash crash: open interest in perpetual futures market-wide evaporated by roughly $19 billion in about 36 hours, accompanied by forced liquidations of the same order of magnitude [9]. In this event, market makers on the on-chain CLOB monitored the sudden surge of one-directional market sell orders and enormous liquidation orders (a liquidation order is essentially a form of forced market order) and quickly "understood" the market's panic signal. Within milliseconds, market makers' algorithms responded defensively: withdrawing existing buy limit orders on a large scale and sharply lowering new quotes. This behavior caused order-book depth within the $\pm 2\%$ price band for major pairs (such as BTC-USDT and ETH-USDT) to evaporate by more than 90% in a short time, and bid-ask spreads widened sharply [10]. This is precisely the defensive manifestation of the adverse-selection mechanism under extreme stress. Although it protected market makers' capital, it also caused an instantaneous depletion of market liquidity, so that every trader attempting to execute an order at that moment faced slippage costs far above normal levels.

7.1.4 The intent-execution gap

For the ordinary trader, understanding the signaling function of orders and the defensive mechanisms of market makers ultimately serves the purpose of crossing one central obstacle—the "intent-execution gap." This is the second core theoretical contribution of this chapter.

This gap refers to the difference between the theoretical price a trader has in mind when forming a trading decision ("I want to buy 50 ETH long around $2,050") and the price at which the order actually fills in the market ("actual average fill price $2,053.7, slippage 0.18%, gas fee $1.2, an additional $12 lost to a sandwich attack"). The concept derives from the "implementation shortfall" framework proposed by Perold (1988) [11], but in the context of on-chain perpetual futures the composition and drivers of the gap exhibit a distinctive distribution.

In on-chain perpetual futures trading, the intent-execution gap can be decomposed into the following cost components (Table 7-3):

First, the latency cost arises from the time lag between the trader clicking "buy" and the order being packed into a block and processed by the matching engine. On a CEX this is typically a few milliseconds or even sub-millisecond; on-chain, even on a highly optimized application chain like Hyperliquid (block time of about 200 milliseconds), it takes several hundred milliseconds from order signing to final confirmation, and on a DEX built on a general-purpose public chain it may take several seconds or even tens of seconds. Within this seemingly brief time window, the market price may already have

moved unfavorably, especially during periods of high volatility. At the same time, the size of the trader's own order consumes liquidity on the order book, causing the average fill price to be worse than the initial best quote; this is the source of price impact. The larger the order size, the more pronounced the price impact; on an on-chain CLOB, where liquidity depth is generally lower than on top-tier CEXs, an order of the same size produces a more severe price impact. In addition, the trader must bear the spread cost—the bid-ask spread paid to market makers—which is essentially compensation for the market maker providing immediate liquidity and bearing adverse-selection risk. As noted earlier, the higher adverse-selection risk in an on-chain environment often results in wider spreads.

Beyond these market-microstructure costs, on-chain trading faces two additional, distinctive frictions. The first is network and protocol fees, including the gas fee paid to validators and the trading fee paid to the protocol. Although application chains such as Hyperliquid have internalized the gas fee (users need not pay gas directly), the protocol fee remains an explicit cost component. The second is slippage and MEV cost—the deviation of the actual fill price from the expected price due to network congestion, front-running, or maximal extractable value (MEV) attacks. The full on-chain transparency makes an order vulnerable, before it is packed (while in the mempool), to exploitation by MEV searchers, who can insert their own buy order ahead of the target buy order (raising the price) and then sell after the target buy order to profit—the classic sandwich attack.

Gap component	Magnitude on CEX	Magnitude on on-chain CLOB	Core driver
Latency cost	Very small (sub-millisecond)	Small to medium (200 ms to seconds)	Architectural performance and consensus latency
Price impact	Small (deep liquidity)	Medium (shallower liquidity)	Order size and order-book depth
Spread cost	Small (highly competitive)	Medium (higher adverse selection)	Market maker's adverse-selection compensation
Gas/protocol fee	None	Present (application chains typically internalize gas)	The fee model of the underlying chain
MEV/slippage	Implicit (internal information use)	Explicit (sandwich attacks/front-running)	Transaction visibility and ordering mechanism
Additional liquidation risk	Present	Present, with the liquidation price publicly visible	High leverage layered onto transparency

Table 7-3. A comparison of the cost components of the intent-execution gap (Data source: compiled by the author)

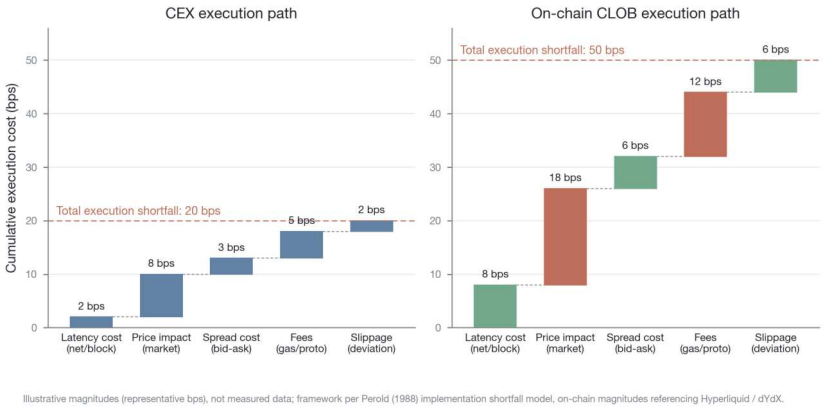


Figure 7-4. A decomposition of the costs of the intent-execution gap (conceptual model illustration: the basis-point components are representative values, not empirical; framework per Perold 1988 [11], with on-chain magnitudes referencing Hyperliquid/dYdX)

> The analytical value of the "intent-execution gap" is that it provides a unified metric dimension for assessing execution quality across different market structures. The point is not to judge simply "which is better," but to understand "how the composition and width of the gap differ." The gap on a CEX is generally narrower, but it contains an unquantifiable "implicit" portion—the risk that the exchange internally exploits order-flow information cannot be audited externally; the gap on an on-chain CLOB may be wider, but every component of it is on-chain observable, quantifiable, and auditable. This is precisely the concrete manifestation, at the level of order execution, of the "transparency paradox" discussed in Chapter 1.

The nonlinear widening of the intent-execution gap during a crisis is a particularly critical phenomenon. Kaiko Research's (2025) analysis of the October 2025 market crisis shows that, under extreme conditions, order-book depth plunged by more than 90% and liquidity contracted sharply [10]. The same $10 million market order incurs slippage that, under normal market conditions, varies by platform type: about 1–5 basis points on a top-tier CEX (such as Binance or OKX), about 10–30 basis points on an application-chain DEX (such as Hyperliquid), and possibly 50–100 basis points or more on a general-purpose public-chain DEX; but in a crisis, slippage may spike to 5% or even 10%, implying an additional cost of $500,000 to $1 million. In other words, the intent-execution gap widens nonlinearly at precisely the moment it most needs to be crossed—when the market is swinging violently and the trader urgently needs to adjust their position. This "liquidity illusion" (liquidity that appears ample in normal times collapsing sharply in a crisis) is an inherent fragility of all order-book-based market structures. The game-theoretic mechanism behind it can be explained by the liquidity-spiral model of Brunnermeier and Pedersen (2009) [12]: when an exogenous shock triggers risk-off sentiment, market makers are forced to withdraw their liquidity supply in unison—each market maker's cancellation is an individually rational defensive act, but the simultaneous withdrawal constitutes a fallacy of composition, causing a collective evaporation of order-book depth. The full transmission chain is: macro or geopolitical shock → the spread of risk-off sentiment → cross-market deleveraging

→ margin calls → forced selling → evaporation of order-book liquidity → amplified price impact → further margin calls. Possible institutional mitigations include imposing minimum-quoting obligations and minimum-quote-time requirements on market makers, but in today's permissionless on-chain DEX environment such constraints are difficult to enforce. Because the on-chain CLOB has a smaller liquidity base, this fragility is even more pronounced.

To bridge this gap, the modern crypto trader must think like an institutional quantitative fund. They no longer focus only on "what to buy" and "when to buy," but devote considerable effort to "how to buy." By allocating orders across multiple liquidity sources through smart order routing, by splitting large orders into smaller ones spread across time through a time-weighted average price algorithm, and by outsourcing order matching to specialized executors through intent-centric solver networks, traders seek to protect their order signals as much as possible in the fully transparent on-chain environment and to secure the optimal execution path. The specific mechanisms of these execution strategies are discussed separately in later sections.

7.2 Basic order types

In the crypto perpetual futures market, the only way for a trader to interact with an exchange's matching engine is to send an order. Orders constitute the basic instruments of trading, and the basic order types—market orders, limit orders, and stop-loss/take-profit orders—are their fundamental types. Understanding these basic order types is not only a prerequisite for mastering a platform's operating interface but also a cornerstone for understanding market microstructure and the price-discovery mechanism.

This section analyzes, from the perspective of microeconomics and market microstructure, the distinctive properties, risk characteristics, and interplay of these three basic order types in the perpetual futures market.

7.2.1 Market orders

A market order is an order that instructs the exchange to execute immediately at the best available price on the current order book. Its core characteristic is to prioritize execution certainty while relinquishing control over the execution price. In microstructure theory, the market order is a "liquidity consumer." When a trader submits a market order, they are in effect "crossing the bid-ask spread," consuming the existing limit-order depth on the order book. The market order is therefore a direct force driving price movement. As Bieganowski and Ślepaczuk (2026) note in their study of crypto microstructure, order-flow imbalance (the net difference between buy market orders and sell market orders) is one of the most critical features for predicting short-term price direction [7].

In the perpetual futures market, market orders have several special properties: high leverage amplifies the impact of slippage. Slippage is the difference between the expected execution price of an order and its actual execution price. In the spot market, 0.1% slippage may be negligible; but in perpetual futures, if a trader uses 50× leverage, 0.1% slippage directly consumes 5% of the initial margin. This means that the implicit

cost of a market order is significantly amplified under leverage. At the same time, in extreme market conditions—precisely when certainty is most needed—the cost is often highest. During a liquidation cascade, order-book depth dries up sharply. A market order submitted at that moment, though guaranteed to fill, will keep "eating" down (or up) the order book for lack of counterparties, producing enormous slippage.

Together these factors point to the fact that a market order transmits a strong information signal. In microstructure models, a large market order is generally seen as a mark of an informed trader entering. When a large market order appears, market makers quickly adjust their quotes to guard against adverse-selection risk, which further raises the execution cost of subsequent market orders.

7.2.2 Limit orders

A limit order is an order that specifies a particular price (or better) and executes only when the market price reaches or improves upon that specified price. In contrast to a market order, a limit order prioritizes price certainty while relinquishing execution certainty. Limit orders are "liquidity providers"; they constitute the depth of the order book. In submitting a limit order, a trader is essentially trading "patience" for a price advantage and lower fees. Limit orders, however, are by no means risk-free. As discussed in Section 7.1.3, submitting a limit order is economically equivalent to giving the market a free option [4]: a buy limit order is equivalent to selling a put, and a sell limit order is equivalent to selling a call.

Building on this theory, limit orders face two main risks. The first is execution risk: when the market moves in the direction favorable to the trader, the limit order may fail to fill, causing the trader to miss potential profit. The second is adverse-selection risk: when a limit order does fill, it is often because the market price has already broken through that limit. Specifically, when a limit order fills, the trader is more likely than not to be in a temporary loss position. In perpetual futures, the empirical research of Ruan and Streltsov (2022) shows that the perpetual-futures funding-rate settlement mechanism further amplifies the adverse-selection risk present in the spot market [8].

In perpetual futures, managing limit orders is more complex because they directly affect effective leverage and the liquidation price. A resting limit order, though not yet filled, occupies available margin (available balance), thereby limiting the trader's ability to withstand the unrealized loss on their current position.

7.2.3 Stop-loss and take-profit orders

Stop-loss and take-profit orders are conditional orders that remain dormant until a specific price condition is triggered. Once the market price reaches the trigger price, they convert into market orders or limit orders. Stop-loss orders play a key role in perpetual futures. A stop-loss is a line of defense the trader proactively sets to cut losses before the liquidation price is reached. Yet the execution mechanics of stop-loss orders conceal many pitfalls that are easily overlooked.

Choice of trigger price source: perpetual futures exchanges generally allow traders to choose the "last traded price," the "mark price," or the "index price" as the trigger

source. Using the mark price as the stop-loss trigger source is strongly recommended. The mark price is typically computed from the spot index price (a weighted median of spot prices across multiple exchanges) plus an exponential moving average (EMA) of the funding-rate basis; it is designed to filter out price anomalies at any single exchange and transient market distortions. The choice of the EMA window parameter directly affects how smooth the mark price is: the longer the window, the more stable the mark price but the more it lags true price movements; the shorter the window, the more responsive but the more susceptible to disturbance from transient volatility. (For a detailed treatment of how the mark price is computed and how the parameters are chosen, see the in-depth discussion in Chapter 14.) The last traded price, by contrast, is easily affected by a single large market order or a transient price distortion, triggering unnecessary stops.

Stop-loss market order vs. stop-loss limit order: whether the order converts into a market order or a limit order when the trigger price is reached determines the type of risk the trader faces (see Table 7-4).

Order type	Advantages	Disadvantages and risks	Applicable scenario
Stop-loss market order	Guarantees a fill (certain position closure); timely loss-cutting under normal liquidity	May face enormous slippage in extreme conditions, with an actual closing price far worse than the trigger price	A hard stop where exiting resolutely is mandatory
Stop-loss limit order	Strictly controls the closing price, with no slippage risk	In a fast one-directional move, the limit order may be "skipped" and fail to fill, resulting in eventual liquidation	Relatively calm conditions, where the closing cost is a sensitive concern

Table 7-4. A trade-off comparison of stop-loss market orders and stop-loss limit orders (Data source: compiled by the author)

The interaction of stop-losses and liquidation: many beginners mistakenly believe that setting a stop-loss makes them absolutely safe. This "illusion of stop-loss safety" is not merely a technical misunderstanding; it has deeper roots in behavioral finance. The illusion-of-control theory proposed by Langer (1975) [13] can explain this phenomenon: through the "active behavior" of setting a stop-loss, traders acquire a false sense of control over risk and thereby underestimate the probability that the stop-loss will fail. In addition, anchoring bias causes large numbers of stops to cluster at round-number thresholds (such as BTC $60,000 or ETH $2,000), and this predictable clustering pattern in turn gives stop-loss hunting economic rationale. As shown in Figure 7-5, there is a "safety buffer" between the stop-loss price and the liquidation price. But in extreme conditions (such as the flash crashes of May 19, 2021, April 2024, or October 2025—this book deliberately cites cases from several different periods to show that this type of risk is not an isolated event but a structural feature of perpetual futures markets), enormous slippage may punch through this buffer in an instant, causing the stop-loss

order to fill at a price worse than the liquidation price, so that the trader still faces liquidation or even a negative account balance (bankruptcy).

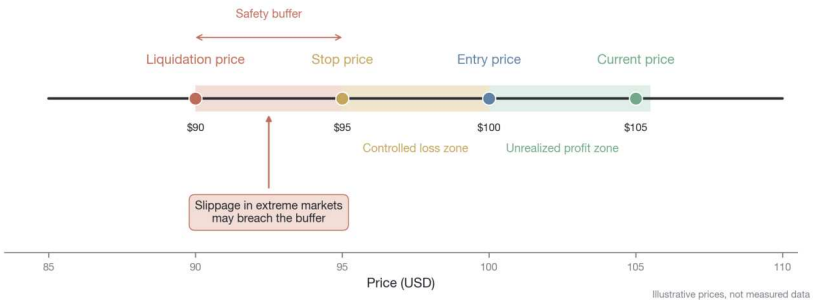


Figure 7-5. The relationship between stop-loss orders and the liquidation price, and the slippage risk under extreme conditions (mechanism illustration: \$90/95/100/105 are illustrative values, not empirical; the long-position safety buffer = stop-loss price − liquidation price, which extreme slippage can punch through)

Trailing stop orders: this is an advanced automated risk-management tool. As shown in Figure 7-6, it has a "ratchet mechanism": the stop line automatically follows as the price moves favorably, but holds unchanged when the price pulls back. This allows the trader to protect gains already secured while leaving room for further profit growth.

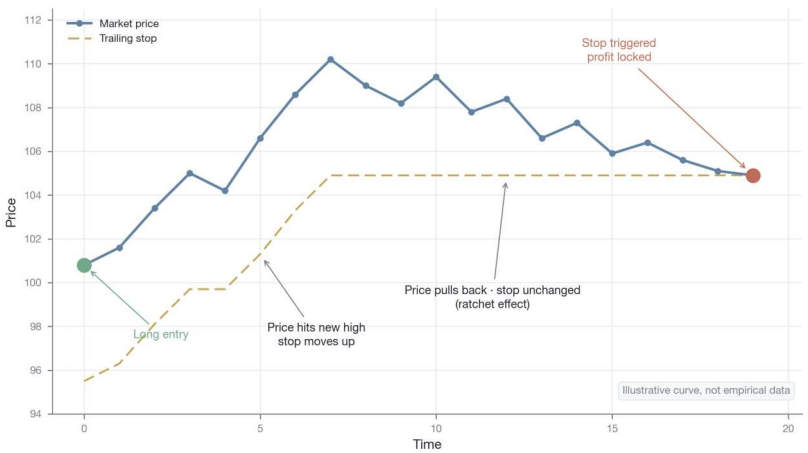


Figure 7-6. The ratchet mechanism of a trailing stop order (long-position example) (ratchet-effect mechanism illustration; the price path uses representative values, not empirical data)

7.2.4 The maker-taker fee model and order selection

Perpetual futures exchanges widely adopt the maker-taker fee model, which is central to understanding the economics of order types. Exchanges incentivize the aggregation of liquidity by charging a higher fee to takers who consume liquidity and a lower fee (or even a rebate) to makers who provide liquidity.

This fee structure strongly shapes a trader's order-selection strategy. As shown in Table 7-5, the difference between maker and taker fees varies enormously across exchanges and VIP tiers. In the fiercely competitive 2025–2026 period, decentralized perpetual futures exchanges such as Hyperliquid have even pushed base fees down to maker 0.015% / taker 0.045% [14].

Order type	Market-microstructure role	Certainty trade-off	Fee attribute	Hidden cost/risk
Market order	Liquidity consumer	Price uncertain, execution certain	Pays the taker fee	Slippage cost; risk amplified under high leverage
Limit order	Liquidity provider	Price certain, execution uncertain	Pays the maker fee (typically lower)	Adverse-selection risk; execution risk
Stop-loss order	Conditional liquidity consumer	Converts to a market/limit order once the condition triggers	Typically pays the taker fee	Liquidation risk from slippage punching through the buffer

Table 7-5. A matrix of the microstructural effects of the basic order types (Data source: constructed by the author)

The core message of Table 7-5 is that the three basic order types display a clear gradient in their certainty trade-offs and hidden costs. The market order offers the highest execution certainty, but its hidden cost (slippage and risk amplification under high leverage) is also the most severe; the limit order provides certainty on the price dimension but exposes the trader to execution risk and adverse-selection risk; the stop-loss order layers the uncertainty of a conditional trigger onto the slippage risk of a market order and may produce unexpected adverse outcomes in extreme conditions.

Within this microstructural ecosystem, limit orders, market orders, and stop-loss orders form a dynamic feedback loop (Figure 7-7). Limit orders constitute the depth structure of the order book, market orders drive price movement by consuming that depth, and stop-loss orders convert into market orders once triggered at a specific price level. When market orders push the price to trigger a large number of stop-loss orders, those stops convert into new market orders that further consume limit-order depth—and this positive-feedback mechanism is precisely the microfoundation of the violent volatility of perpetual futures markets.

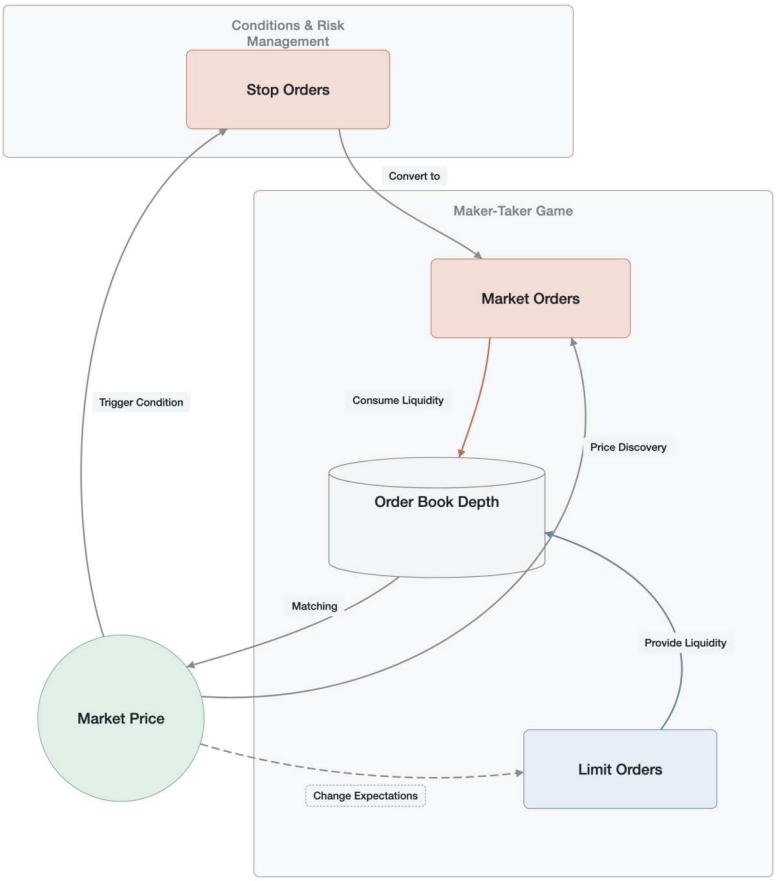


Figure 7-7. The dynamic feedback loop of order flow and price discovery (mechanism feedback-loop illustration, not empirical data)

Having grasped this game mechanism, a trader understands that providing liquidity requires bearing adverse-selection risk and earning corresponding compensation, while execution certainty comes at the cost of higher fees and slippage. In perpetual futures trading, advanced traders rarely enter and exit using market orders alone; instead, they seek the optimal solution between certainty and cost through advanced order types and algorithmic execution. The next section analyzes these in detail.

7.3 Advanced order types

On top of the basic order types, advanced order types allow traders to refine the expression of their strategy along dimensions such as time, size, and execution conditions. As the crypto derivatives market becomes more institutionalized, simple buy and sell instructions can no longer meet complex execution needs. Advanced order types allow traders to express their trading strategies with precision along four dimensions—time, size, execution constraints, and conditional triggers—thereby gaining an edge or lower-

ing costs within a complex microstructure.

7.3.1 Time-control orders

Time-control orders determine the lifespan of an order within the matching engine. They are essential tools for high-frequency trading and arbitrage strategies.

Fill-or-kill (FOK) is an extremely strict order type. It requires the order to fill immediately and in full, or else the entire order is canceled; partial fills are never permitted. FOK has an important role in calendar-spread arbitrage or funding-rate arbitrage. When an arbitrageur spots a price gap between two markets, they need to establish equal positions on both sides simultaneously. Using an ordinary market order might fill one side while the other fills only partially (leg-execution risk), leaving exposure to a one-sided position. FOK ensures the certainty of "all or nothing." Immediate-or-cancel (IOC) requires the order to execute any fillable portion immediately and then cancel the entire unfilled remainder. Unlike FOK, IOC permits partial fills. It is commonly used for "sweeping" operations, in which the trader is willing to consume all favorably priced liquidity on the current order book but unwilling to leave the unfilled portion resting as a limit order, so as not to expose their trading intent. In addition, good-til-time/good-til-block (GTT/GTB): on on-chain order-book decentralized exchanges, GTT often evolves into GTB (good-til-a-specified-block-height). Because blockchain networks are subject to latency and reorganization risk, a trader who does not want their limit order to be unexpectedly executed after market conditions have changed drastically (say, several blocks later) attaches a GTB constraint.

7.3.2 Size-hiding orders

When an institutional investor needs to execute a large order that far exceeds the current order-book depth, they face significant information-leakage risk. Posting a huge limit order directly would immediately alert other market participants (especially high-frequency market makers) to this enormous selling or buying pressure, prompting them to front-run or withdraw quotes, so that the large order becomes difficult to fill or fills at a badly unfavorable price.

The iceberg order was created precisely to solve this pain point. As the name suggests, an iceberg order exposes only a small fraction of its total size to the public order book while hiding the vast majority. Once the visible portion fills completely, the matching engine automatically slices a new visible quantity from the hidden portion and replenishes it onto the order book, until the total order is fully executed.

On a centralized exchange, the logic of the iceberg order is maintained by the centralized matching engine in the background. On an on-chain perpetual futures DEX, however, implementing a genuine iceberg order faces an important technical challenge: there is an inherent tension between the transparency of the blockchain and the need to hide size. All on-chain state is public by default. At present, some cutting-edge DEXs are beginning to explore using zero-knowledge proofs or trusted execution environments to build on-chain dark pools, allowing traders to match large liquidity without broadcasting their true intent to the entire network.

7.3.3 Execution-constraint orders

These two constraint-type orders appear extremely frequently in the daily operations of perpetual futures traders; they target fee optimization and position safety, respectively.

Post-Only (maker-only): for a limit order with the Post-Only attribute attached, the system guarantees that it will never match and fill immediately against an existing order on the book. If the order would immediately match an existing order upon submission (for example, a buy limit priced above the current best ask), the system cancels the order outright rather than converting it into a market order for execution. The core value of Post-Only is to strictly lock in the maker fee. As discussed in Section 7.2, the maker fee is typically far lower than the taker fee (or even negative). Market makers and algorithmic traders use Post-Only to ensure that they only ever provide liquidity and never accidentally consume it, thereby precisely controlling trading costs.

Reduce-Only: this is an order attribute unique to the derivatives market. A Reduce-Only order can only be used to reduce the current position; it never increases a position or establishes a new position in the opposite direction. This feature has an important role when setting take-profit and stop-loss orders. Suppose a trader holds a long position and has set an ordinary sell limit order as a take-profit; if they subsequently close the long manually but fail to cancel the take-profit order, then when the price reaches the take-profit level the sell order will be triggered, unexpectedly establishing a naked short position. If the order has the Reduce-Only attribute attached, then when the long position is zero the system will automatically cancel or refuse to execute the order, entirely eliminating this risk of unintended position-building.

7.3.4 Algorithmic orders: TWAP and VWAP

For truly enormous orders, even an iceberg order may be identified by a clever algorithm because of its frequent replenishment actions (known as “iceberg detection”). At that point, an institution must turn to algorithmic execution strategies, of which the most classic are TWAP and VWAP. Figure 7-8 contrasts the core logical differences among three large-order execution strategies—TWAP, VWAP, and the iceberg order—in terms of execution cadence and information concealment; each is discussed in turn below.

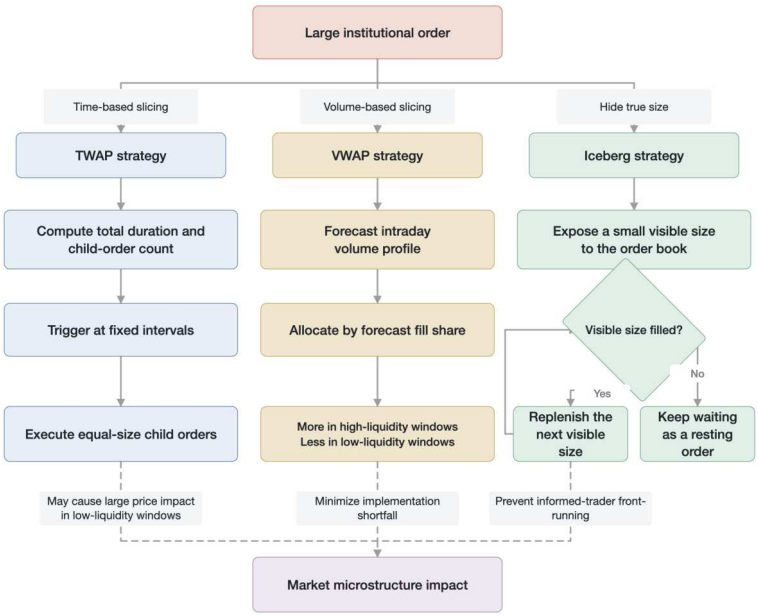


Figure 7-8. A logical comparison of the mainstream large-order algorithmic execution strategies (TWAP, VWAP, and the iceberg order) (conceptual logical-comparison illustration, not empirical data)

The TWAP algorithm divides a large order evenly into many equal-sized child orders over a specified total time and sends them to the market at fixed time intervals. TWAP's advantage lies in its simplicity and robustness; it requires no prediction of market behavior at all. However, liquidity in the crypto market is highly uneven across the intraday cycle. Blindly executing a fixed-quantity TWAP during a low-liquidity window—such as the middle of the night in Asia—will cause considerable price impact.

The VWAP algorithm is more intelligent: it aims to keep its execution cadence aligned with the market's actual volume distribution. In practice, institutional traders typically use adaptive VWAP—re-estimating the volume curve of the remaining window in real time during execution rather than relying entirely on historical data. This distinction is especially important on event-driven days (such as major regulatory announcements or black-swan events), when the historical "twin-peak" volume pattern may fail entirely and the execution performance of a static VWAP deteriorates markedly. As shown in Figure 7-9, according to research by Talos (2025) and Jasiak and Zhong (2024) on the intraday trading characteristics of Bitcoin, crypto-market liquidity exhibits a pronounced "twin-peak" feature: it reaches its highest peak at 14:00–15:00 UTC (the U.S. stock market open), a secondary peak at 08:00–10:00 (the European open), and its trough at 00:00–04:00 (the middle of the night in Asia) [15][16].

The VWAP algorithm predicts this intraday volume curve, allocating more execution quantity during windows predicted to have high liquidity (such as the U.S. open) and less during low-liquidity windows. Research by Genet (2025) further shows that using a deep-learning framework to optimize the VWAP execution objective directly can signif-

icantly outperform the traditional benchmark VWAP strategy, effectively minimizing implementation shortfall—the difference between the ideal decision price and the actual execution price [17].

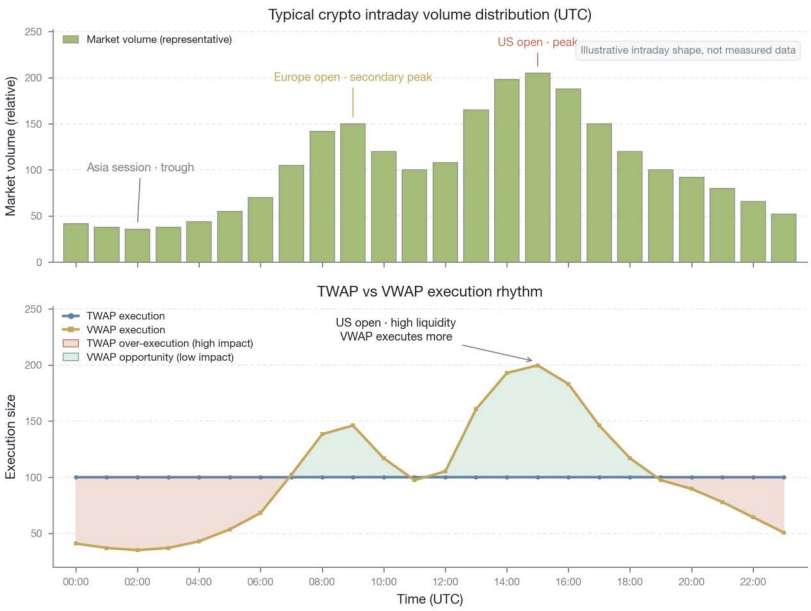


Figure 7-9. A comparison of the typical intraday volume distribution in crypto with the execution cadences of TWAP/VWAP (representative intraday volume-profile illustration: the curves are stylized values, not empirical; the twin peaks are per Talos 2025 [15] and Jasiak & Zhong 2024 [16], and DL-VWAP is per Genet 2025 [17])

Table 7-6 systematically compares TWAP and VWAP along four dimensions: core logic, advantages, disadvantages, and applicable scenario. The core trade-off between the two algorithms is that TWAP trades execution simplicity for a disregard of changes in market structure, whereas VWAP trades reliance on the historical volume pattern for a better execution price.

Algorithm type	Core logic	Advantages	Disadvantages and risks	Applicable scenario
TWAP	Splits the order linearly and evenly over time	Simple logic, no reliance on historical-data prediction; provides some smoothing against extreme "wicks"	Ignores intraday liquidity variation; forced execution during low-liquidity windows leads to costly slippage	Assets lacking a reliable volume-prediction model; tasks with extremely high time-certainty requirements
VWAP	Splits the order in proportion to predicted market volume	Execution trajectory tracks market depth, minimizing impact on market microstructure and implementation shortfall	Highly dependent on the accuracy of the volume-curve prediction; prediction fails when the market undergoes a structural break	Mainstream assets with clear cyclical liquidity patterns (such as BTC/ETH); large position-building highly sensitive to slippage

Table 7-6. A matrix of algorithmic execution strategies for advanced order types (Data source: constructed by the author)

As Table 7-6 shows, TWAP is suited to long-tail assets that lack a reliable volume-prediction model, whereas VWAP performs better in large position-building scenarios for mainstream assets (such as BTC/ETH). In actual execution, institutional traders typically switch between the two algorithms according to the liquidity characteristics of the underlying asset and the execution time window.

7.3.5 Conditional orders

Beyond conventional price triggers (such as stop-loss/take-profit), modern perpetual futures platforms have also spawned conditional orders based on more complex microstructure data. These orders allow traders to push strategy logic directly down to the matching-engine level.

A funding-rate-based conditional order triggers when the funding rate deviates beyond a certain threshold. For example, an algorithmic trader can set: "when the BTC funding rate exceeds 0.05% (longs extremely crowded), automatically trigger a market sell to go short." This allows the trader to capture fleeting rate-arbitrage opportunities. Trading strategies executed automatically on algorithmic conditions may, in several jurisdictions, be classified as algorithmic trading, thereby triggering compliance obligations such as registration, risk-control systems, and trade reporting (for example, the definition of algorithmic trading under the EU's MiFID II and the related rules of the U.S. SEC). A conditional order based on open interest is another option, as OI is a key indicator of market momentum. When open interest (OI) surges in a short period, it often foreshadows imminent violent volatility. A conditional order based on changes in OI can help a trader intervene automatically at the early stage of a trend breakout. There is also the conditional order based on the liquidation heatmap, a highly advanced form of strategy expression. A trader who anticipates that a dense cluster of highly

leveraged long liquidation lines exists at a certain price level can set a conditional order to automatically provide liquidity (buy) when that region is triggered, thereby acquiring a position at a lower price amid a large-scale liquidation cascade.

The evolution of advanced order types is, in essence, an escalation of the game between traders and market microstructure. From the market order's direct impact on the order book, to the limit order's passive waiting to fill, to VWAP's adaptation to the market's liquidity cadence, the richer the available order toolkit, the finer the granularity of strategy expression—and the greater the probability of surviving in the zero-sum game of the derivatives market.

7.4 The strategic dimensions of perpetual futures

In traditional spot trading, the order decision is generally a relatively simple two-dimensional optimization problem: the trader mainly weighs price (at what price to execute) against signal leakage (which order type to choose to control information exposure). In perpetual futures, however—the most dominant derivative instrument in the crypto market—the order decision gains three additional strategic dimensions: leverage, liquidation, and time. These three dimensions make the order decision in perpetual futures an order of magnitude more complex than in the spot market, and they also give "a good order strategy" an influence on a perpetual futures trader's profit and loss far exceeding that in the spot market. For ineffective traders who suffer systematic losses in the market over the long run, the losses often stem not from errors in directional judgment but from their unconscious exposure along these three key dimensions. Behavioral-finance research further shows that even when traders understand the three dimensions of leverage, liquidation, and time at a cognitive level, they still make erroneous decisions because of systematic cognitive biases: overconfidence bias causes traders to systematically underestimate liquidation risk ("I won't get liquidated"), the disposition effect causes them to hold losing positions too long and close winning positions too early, and mental accounting causes them to treat each position under cross-margin mode as an independent bet while ignoring portfolio-level risk contagion. Figure 7-10 uses a three-dimensional coordinate system to show how these three dimensions layer onto the traditional price-signal two-dimensional framework: the horizontal axis represents leverage (from 1× to maximum leverage), the vertical axis represents liquidation distance (from safe to dangerous), and the time axis represents the funding-rate settlement cycle. Every order has a definite coordinate in this three-dimensional space, and that coordinate determines the order's overall degree of risk exposure.

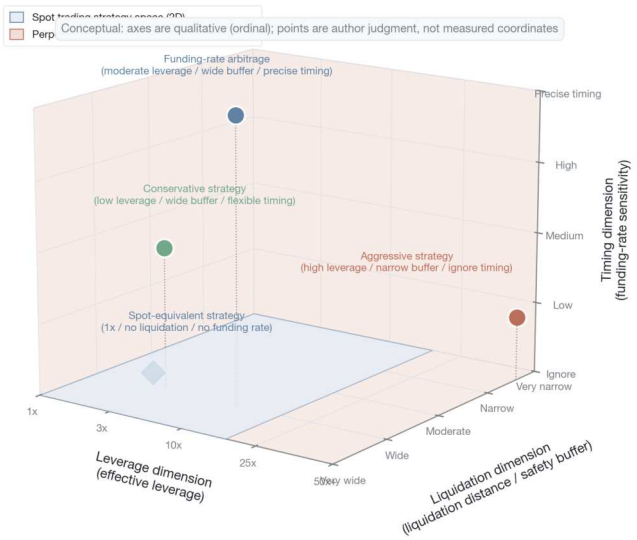


Figure 7-10. The three additional dimensions of perpetual futures order strategy (conceptual three-dimensional strategy-space illustration; the positioning points reflect the author's judgment, not measured coordinates)

7.4.1 The leverage dimension

In the spot market, executing a buy order merely converts cash into an asset, and the risk exposure is linear and bounded. In perpetual futures, however, each order execution not only changes the direction and size of a position but also directly changes the account's effective leverage. Effective leverage is defined as the ratio of notional position value to total account margin (including unrealized profit and loss), computed as:

> Effective leverage = total notional position value / total account equity (margin + unrealized P&L)

When a trader submits a new add-on order on top of an existing position, that order in effect "compresses their risk tolerance." Consider a concrete numerical case to illustrate this dynamic. Take an account with $10,000 of margin, initially holding no position. Executing a $50,000 notional Ethereum perpetual long order takes the account's effective leverage from 0× to 5×. At this leverage level, assuming a maintenance margin rate of 0.5% (charged on notional position value; in practice exchanges generally use a tiered, dynamic maintenance margin rate—Binance's, for example, ranges from 0.4% for small positions to 5% for large ones—so the 0.5% here is merely a pedagogical simplification), the liquidation price sits roughly 19.5% below the entry price. If the trader then executes another $50,000 buy order, the total notional value reaches $100,000 and effective leverage jumps to 10×. At this point the liquidation price rises sharply, narrowing to only about 9.5% below the average entry price. A single add-on order has compressed the "risk tolerance" by half.

> Core insight: in perpetual futures, every add-on order "compresses your risk tolerance," and every reduce order "expands your risk tolerance." The "next order," therefore, is not merely a trading decision but a profound risk-management decision.

This dynamic variation in the leverage dimension requires traders to implement strict margin-allocation strategies at the order level. Two common strategic patterns embody different risk-management philosophies:

Incremental position-building (the leveraged version of dollar-cost averaging) requires the trader to execute the target total position in tranches. After each add-on order fills, the new effective-leverage level is systematically assessed. Once effective leverage approaches the trader's preset upper limit of risk tolerance, further adding stops. The core advantage of this strategy is that the trader retains, at every step, the option "not to keep adding," rather than exposing all margin to a single entry price at once.

Pyramiding, by contrast, requires the trader to add to the position using a sequence of orders of decreasing size after the market trend is confirmed. For example, the first order is 40% of the target position, the second 30%, the third 20%, and the last 10%. This order strategy can increase profit exposure while preventing the average entry price from converging too quickly toward the current market price, thereby keeping effective leverage within a relatively safe range. Its essence is to trade "decreasing marginal commitment" for "increasing trend confirmation."

7.4.2 The liquidation dimension

The liquidation dimension is the most distinctive part of the perpetual futures order decision. In the traditional spot market, the "worst outcome" an investor faces is that the asset value goes to zero, but never a liability. In the high-leverage environment of perpetual futures, the "worst outcome" is not only complete forfeiture of margin but, under extreme market volatility and liquidity depletion, may even produce a bankruptcy liability (though most platforms absorb this portion of the risk through an insurance-fund mechanism). The liquidation price is precisely the rigid mathematical threshold that triggers this "worst outcome." Taking a long position as an example, the approximate formula for its liquidation price is:

> Liquidation price (long) $\approx$ average entry price $\times$ (1 $-$ 1/effective leverage + maintenance margin rate)
> Liquidation price (short) $\approx$ average entry price $\times$ (1 + 1/effective leverage $-$ maintenance margin rate)

The approximate formulas above apply to a single position under isolated-margin mode. Under cross-margin mode, the liquidation price is determined by total account equity rather than the margin of a single position, and the formulas above no longer apply. The approximate formulas can also produce large errors under high leverage (50$\times$ to 100$\times$), because they ignore the effects of cumulative funding, trading fees, and fluctuations in unrealized P&L. In actual trading, a trader should rely on the exchange's real-time computed liquidation price rather than estimate it from a simplified formula.

Within this dimension, the most critical order decision is the setting of the stop-loss order and its interaction with the liquidation price. When submitting any opening order, a trader must simultaneously compute and plan the spatial relationship between the liquidation price and the stop-loss price. This relationship has three basic states:

If the stop-loss trigger price is set beyond the liquidation price (that is, for a long, the stop-loss price is below the liquidation price), the stop-loss order is logically entirely ineffective. This is because, before the market price reaches the stop-loss price, the exchange's risk engine has already forcibly taken over the position and executed the liquidation procedure. Surprisingly, in actual trading, such "ineffective stop-loss" settings—arising from an insufficient understanding of the liquidation mechanism—are not rare. Viewed from another dimension, if the stop-loss order is set before the liquidation price with an ample buffer, it can provide effective protection under normal market conditions. However, as discussed earlier (Section 7.2.3), in extreme conditions a stop-loss market order may encounter enormous execution slippage due to oracle latency or an instantaneous drain of order-book liquidity, causing the final fill price to be even worse than the theoretical liquidation price. The third state is a stop-loss price set flush against the liquidation price, with an extremely narrow buffer. This configuration pursues extreme capital efficiency, but it essentially outsources the trader's own risk management to the exchange's liquidation engine. In a transparent on-chain environment, such an extremely narrow buffer is prone to becoming the target of deliberate price pushing—a risk Section 7.5 analyzes in detail.

The distance between the "liquidation price and the stop-loss price" constitutes the "safety buffer" that the trader actively chooses. The wider the buffer, the higher the trader's probability of surviving violent volatility; but this also means the trader must accept lower effective leverage or a larger single stop-loss, thereby compressing the potential return on capital. This trade-off has no "correct answer," only "an answer matched to the trader's risk preference."

7.4.3 The time dimension

Perpetual futures can track the spot price closely without expiry delivery because of the funding-rate mechanism. This mechanism typically settles once every 8 hours (on some application chains the frequency is higher: Hyperliquid settles once every hour, while dYdX v4 adopts a continuous-accrual model in which the funding rate accrues proportionally within each block without waiting for a fixed settlement point). (For a detailed analysis of the funding-rate formula, the clamp mechanism, and the game-theoretic structure, see Chapter 10.) At the moment of settlement, funds transfer directly between traders holding positions: when the perpetual futures price is above the spot price, longs pay funding to shorts, and vice versa. This periodic flow of funds creates a strategic dimension entirely absent from the spot market—the choice of order timing around the settlement moment.

In the traditional spot market, the "timing" of order execution mainly affects slippage and market-impact cost, a matter of execution-level optimization. In perpetual futures, however, execution timing directly determines whether the trader pays or receives funding—a source of profit and loss independent of price direction. This gives

rise to a series of specific order strategies:

The rate-sniping strategy is the most direct application of the time dimension. When market sentiment is extremely one-sided, causing the expected funding rate to be abnormally high (for example, extremely crowded longs producing an extremely high positive rate), a trader may submit a short market order or an aggressive limit order minutes or even seconds before settlement. After crossing the settlement moment and collecting the high funding rate, they close the position quickly. The core of this strategy is to weigh the net benefit between "the funding collected" and "the slippage and fees generated by the two orders." On platforms that settle hourly, such as Hyperliquid, the frequency and precision with which this strategy can be executed far exceed those of the traditional 8-hour settlement cycle. In practice, however, the strategy faces severe execution constraints. First, the order book during a high-rate settlement window is often extremely thin (other participants are also competing for the same direction), so the adverse-selection cost rises significantly. Second, even within a 1-hour settlement cycle, a 1% adverse move in the underlying asset means a 10% margin loss at 10× leverage, whereas the corresponding rate gain is typically only 0.01%–0.05%—directional risk can be dozens of times the rate gain. Third, some exchanges have adopted time-weighted rate calculations (prorating by the ratio of holding time to the settlement cycle), greatly reducing the effectiveness of last-moment sniping. Consider a simplified numerical case to illustrate the break-even condition: assume the expected funding rate is 0.05% and the trader uses 10× leverage to establish a $1 million notional short position ($100,000 of margin); the single-period rate gain is $500 ($1 million × 0.05%). If the round-trip trade (the two legs of opening and closing) incurs fees of 0.05% × 2 legs = $1,000, plus estimated slippage of 0.02% × 2 legs = $400, then the trading cost alone reaches $1,400 and the strategy makes a net loss of $900. This strategy has positive expected value only when the rate is extremely high (typically above 0.1%) and trading costs are extremely low, and its profitability has a strong dependence on the macro cycle: in a sustained one-directional trend, the funding rate may remain in the same direction for a long time, invalidating the mean-reversion assumption.

The avoidance-reduce strategy exploits the time dimension from a defensive angle. If a trader holds a large trending long position and currently faces an extremely high positive funding rate, they may choose to submit a reduce order before settlement to avoid paying the expensive funding, and re-establish the position after settlement. However, this operation of moving in and out of the order book frequently not only increases trading costs (two sets of fees plus potential slippage) but may also expose the trader's intent, making them a source of signal leakage.

> Core insight of the time dimension: the funding-rate mechanism of perpetual futures makes "order timing" an independent strategic variable. An originally perfect trend-following order, if it merely chooses the wrong execution timing and pays a high funding rate for several consecutive days, may have its final profit-and-loss outcome completely reversed. Chapter 10 analyzes the game-theoretic structure of the funding rate in depth.

7.4.4 The choice of margin mode

Beyond the three dynamic dimensions above, perpetual futures order strategy is also profoundly constrained by the underlying account-level institutional framework. Before submitting an order, a trader must choose the margin mode in which the order sits, and this choice determines the boundary of risk contagion and constitutes the institutional framework for all subsequent order behavior (a comparison of the two modes appears in Table 7-7).

Under isolated-margin mode, each position has its own isolated margin. The loss and eventual liquidation of one position will absolutely not affect the other positions in the account or the available balance. The philosophy of this mode is to "isolate the risk exposure of each position," achieving strict risk isolation. For high-risk, high-leverage one-directional speculative orders, isolated mode is the preferred choice, because it strictly confines the "worst outcome" to the margin allocated to that position. Its cost, however, is low capital efficiency: funds cannot be shared across different positions, and even if one position has a large amount of unused margin, it cannot be used to support another position that is taking a loss.

Under cross-margin mode, all positions in the account share a single margin pool. The unrealized profit of a winning position can serve as collateral to "support" a losing position, greatly improving capital efficiency. The philosophy of this mode is to "bring all positions into a unified risk pool to improve capital efficiency." However, cross-margin has strong procyclicality. Ali (2025), in a study of the crypto-market liquidation cascade of October 2025, notes that the event cleared $19 billion of open interest in 36 hours, and that the procyclicality of cross-margin was a key systemic-risk amplifier [9]. Its mechanism is as follows: when a macro shock causes all assets to fall together, the collateral value of all positions in a cross-margin account shrinks simultaneously, systematically lowering the margin health of the entire account and ultimately causing even originally safe positions to be liquidated by association. This chain of intra-account liquidation contagion does not occur under isolated mode (though market-level and liquidity-level contagion still exist).

The risk profile of cross-margin, however, has a pronounced "two-sidedness," and one must avoid simplistically labeling it a "risk amplifier." Under normal market volatility and with asset correlations near their historical means, cross-margin actually acts as a stabilizer: the unrealized profit of winning positions provides a buffer for losing positions, lowering the probability that a single position is liquidated by temporary volatility. Only when the market undergoes a regime switch—when asset correlations suddenly jump from normal levels to near 1 (that is, "all assets fall at once")—does cross-margin flip into an amplifier: the collateral value of all positions shrinks in unison, the profit buffer vanishes, and the liquidation probability jumps nonlinearly. This critical feature of the regime switch is key to understanding the systemic risk of cross-margin.

Looking further, the systemic shock produced by liquidation propagates in turn through three levels: intra-account contagion (under cross-margin mode, the liquidation of a single position consumes shared margin and drags down other positions) → market contagion (the forced market orders produced by liquidation strike the order book, pushing the price further away and triggering the liquidation of more accounts) →

liquidity contagion (after observing the signal of a liquidation cascade, market makers withdraw liquidity on a large scale, order-book depth plunges, and the price impact of subsequent liquidations intensifies). This contagion path is not unique to crypto markets. Adrian and Shin (2010), in a study of the traditional financial system, had already revealed the systemic-risk mechanism of procyclical leverage and margin spirals [18]. The cross-margin problem in crypto markets is, in essence, an accelerated version of this known systemic-risk mechanism in a 24/7, high-leverage, digital-native environment.

Dimension	Isolated margin	Cross margin
Risk isolation	Strict isolation; liquidating one position does not affect others	No isolation; losses contaminate the entire account
Capital efficiency	Low; funds locked in a fragmented way	High; funds shared and utilized
Liquidation trigger	Triggered when a single position's margin is insufficient	Triggered when total account equity is insufficient
Procyclical risk	Low; risk is isolated	High; risk amplified in a market-wide sell-off
Applicable scenario	High-leverage speculation, standalone strategies	Hedged portfolios, low-leverage multi-asset allocation
Mental model	"Each trade is an independent bet"	"All trades are one portfolio"

Table 7-7. A comparison of isolated margin and cross margin (Data source: compiled by the author)

The choice of margin mode, therefore, is by no means a simple technical setting; it is a fundamental risk-management decision that affects all subsequent order behavior. It determines "whether, and how, the loss from one bad order contaminates the entire portfolio."

7.4.5 A unified decision framework

By combining the traditional order-selection dimensions (price, signal) with the three extended dimensions unique to perpetual futures (leverage, liquidation, time) and the margin-mode institutional framework, we can construct a unified order-decision framework. Under this framework, the formation of every perpetual futures order must pass through a multi-objective optimization process comprising seven decision nodes.

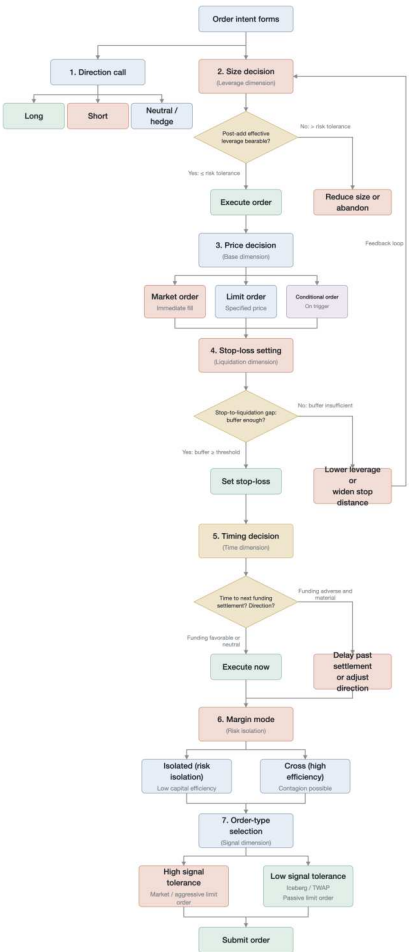


Figure 7-11. A unified decision tree for perpetual futures order strategy (conceptual decision-tree illustration, not empirical data)

As shown in Figure 7-11, after forming a preliminary trading intent, a mature trader must pass in turn through the following seven nodes:

The trader's starting point is directional judgment—deciding, based on fundamental analysis or technical signals, whether to go long, go short, or construct a neutral hedge. This is the step every trader performs, and the only step most ineffective traders perform. Building on this, the trader must convert the directional judgment into a sizing decision, calculating whether effective leverage after adding remains within their risk tolerance. If it exceeds the tolerance, they must reduce the order size or forgo the trade, thereby converting the subjective judgment of "I favor this direction" into the quantified decision of "how much risk I am willing to bear." Next comes the price decision: choosing immediate execution (a market order), waiting at a specified price (a limit order), or a conditional trigger (a stop-loss/take-profit order). This dimension exists in the traditional spot market as well, but in perpetual futures the price decision is tightly coupled with leverage, and different entry prices imply different effective leverage and

liquidation distances.

At the same time, the stop-loss setting constitutes a key feedback loop in the decision tree: the trader must assess whether the safety buffer between the stop-loss price and the liquidation price is sufficient to withstand slippage in extreme volatility, and if the buffer is insufficient, they must fall back to lower the leverage or widen the stop-loss distance. Then the timing decision requires the trader to examine the time remaining until the next funding-rate settlement and the expected direction of the rate, and decide whether to execute immediately or defer until after settlement. For a large position, the funding cost may be a non-negligible factor in profit and loss.

Having established the risk framework, the trader must still make two structural choices. The margin mode determines whether this order is placed in the capital-efficiency-seeking cross-margin pool or under the risk-isolation-seeking isolated mode, thereby defining the order's risk-contagion boundary. The choice of order type then requires the trader, based on their own tolerance for information leakage, to make the final trade-off between execution certainty and signal concealment: for a trader with high signal tolerance, a market order or an aggressive limit order is a reasonable choice; for a large trader who needs to hide intent, an iceberg order, TWAP, or a passive limit order is the better strategy.

> Every perpetual futures order is a multi-objective optimization across these seven decision nodes. This is precisely the microfoundation of why perpetual futures trading is "simple to operate but complex to decide." Ineffective traders who suffer systematic losses in the market over the long run typically stop at only the first node of the decision tree—"directional judgment"—while entirely neglecting the six critical decision nodes that follow. This insufficient analysis is precisely the microstructural operationalization of the systematic losses of the "ineffective traders" analyzed in Chapter 4.

7.5 The commitment property of on-chain orders

Having understood the essence of the order as a signal and the three-dimensional extension that perpetual futures confer on orders, we must now turn our attention to the execution architecture itself. Traditional financial theory generally treats the execution architecture as a neutral background variable: whether at the New York Stock Exchange or on Nasdaq, the economic meaning of a limit order is assumed to be the same. In decentralized finance, however, the physical constraints of the underlying blockchain reshape the economic semantics of an order.

In the two starkly different environments of the centralized exchange and the on-chain central limit order book, the same order faces different processing flows. The core difference is this: on an on-chain CLOB, an order changes from a "private expression of intent" on a CEX into a public, costly, on-chain commitment observable by everyone. This commitment property is the ultimate concretization, at the order level, of the "transparency paradox" of Chapter 1 and the "transparency-fragility paradox" of Chapter 4, and it changes the trader's strategic choices. Taking a single limit order as an example, Figure 7-12 contrasts the "private expression of intent" on a CEX with the "public commitment" on an on-chain CLOB along three dimensions: visibility, cancella-

tion cost, and confirmation latency (the mechanism of each path is detailed in Sections 7.5.1 and 7.5.2).

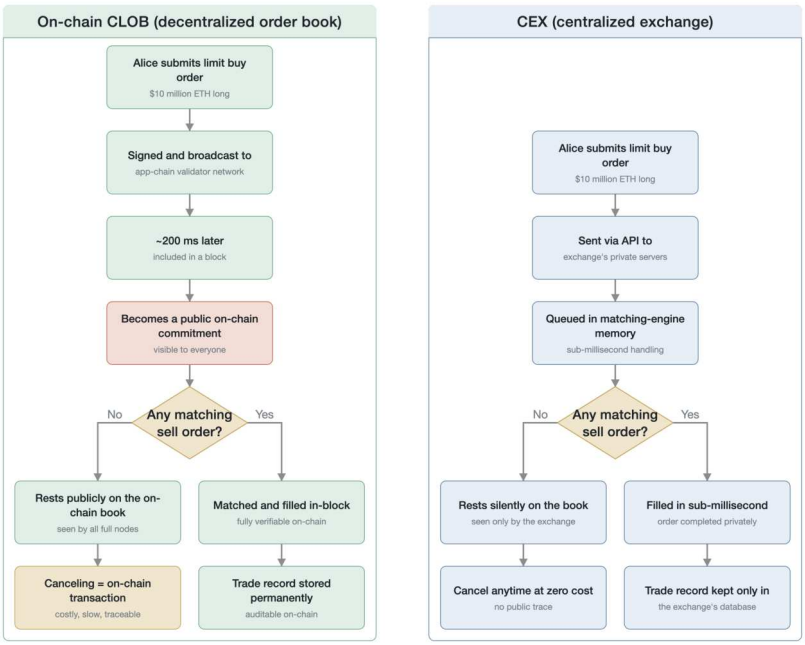


Figure 7-12. The different processing paths of an order on a CEX versus an on-chain CLOB (mechanism-path comparison illustration, not empirical data)

7.5.1 The privacy of CEX orders

In a traditional CEX architecture, when a trader submits a limit order, the order is sent through an API or web front end directly to the exchange's internal private matching engine. Within the matching engine's memory, this order is visible only to the exchange itself and to specific authorized counterparties (such as certain market makers who enjoy a special data feed). If the order is not executed immediately, it rests at a price level on the order book, entirely invisible to the outside world.

For the trader, the CEX environment offers a high degree of strategic flexibility. If market conditions change slightly, the trader can cancel this order at any time, at sub-millisecond speed and zero cost. After the order is canceled, no external observer can prove that the order ever existed; it leaves no public record and no on-chain trace.

Advantage: strategic-privacy protection relative to the public on-chain ledger (though this privacy does not extend to the exchange itself). The trader's intent, capital base, and degree of patience cannot be obtained and exploited in advance by external market participants. This privacy allows large traders to use advanced strategies such as iceberg orders relatively safely to hide their true intent, and allows market makers to instantly withdraw all quotes when unfavorable information appears in order to avoid adverse selection.

Disadvantage: unverifiability and trust cost. The opaque nature of a CEX means that an ordinary trader cannot verify "whether the order book you see is real." The exchange is fully capable of displaying fake market depth, or of using its exclusive visibility into the global order flow to grant itself preferential treatment (such as front-running customer orders). Under this architecture, "the trader's strategic privacy is protected, but at the same time the integrity of the market data cannot be independently verified." This is the structural paradox of the CEX order model.

7.5.2 The public-commitment property of on-chain orders

When the same limit order is submitted to a CLOB protocol built on a general-purpose public chain or an application chain, its nature undergoes an important transformation. The order must be cryptographically signed and broadcast as a transaction to the validator network. From the moment it enters the mempool or is packed into a block, this limit order becomes an on-chain observable public commitment.

Everyone in the world running a full node can clearly see that "someone has committed to buying quantity Q of an asset at price P." More importantly, this commitment cannot be canceled "quietly." Canceling the order likewise requires sending an on-chain transaction, which not only incurs a gas fee (or is constrained by network bandwidth) but also requires waiting at least one block of consensus confirmation. During the time spent waiting for the cancellation to confirm, the original order is still "alive" and can still be filled by anyone.

This "public commitment" property brings mechanism effects that are both positive and negative. The positive meaning: verifiable, genuine liquidity. The on-chain order book achieves verifiability at levels L3 to L4 of the verifiability ladder of Chapter 5. External observers can independently audit every drop of liquidity in the order book, ensuring that no fake depth fabricated by the exchange exists. Everyone sees the same real, fair market state; with respect to the single dimension of fake depth, information asymmetry is significantly reduced at the infrastructure level. This transparency is a structural advantage that current CEX architectures generally cannot provide (without introducing additional cryptographic proofs). The negative meaning: strategy exposure and cancellation cost. Your limit order tells the whole world "I have definite demand at this price." MEV searchers, competing market makers, and other strategic traders can all exploit this public information at no cost to adjust their own strategies. At the same time, because cancellation carries latency and cost, a market maker facing sudden unfavorable information cannot instantly withdraw all quotes as on a CEX, and this significantly increases the adverse-selection risk they face (the transmission by which they widen spreads and pass the cost on to liquidity consumers is detailed in Section 7.6.3).

7.5.3 The MEV exposure surface of orders

In a transparent on-chain environment, an order is no longer merely a number in a matching engine but an object that various algorithms may exploit. Different types of orders, because of the differences in the information they leak to the market and in their execution mechanisms, face starkly different MEV exposure surfaces.

According to data analysis by EigenPhi (2025), the monthly amount extracted by sandwich attacks on the Ethereum mainnet has fallen from nearly $10 million in late 2024 to about $2.5 million by October 2025 [19]. This decline may be partly attributable to the spread of private RPCs and MEV-resistant routing (such as Flashbots Protect and MEV Blocker), as well as to competition among searchers compressing profit margins [20]. For large orders on an on-chain CLOB, however, MEV remains one of the largest implicit execution costs.

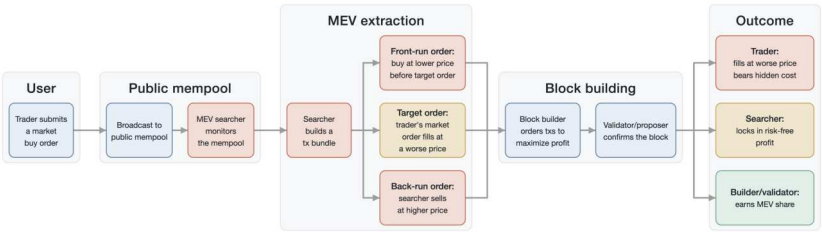


Figure 7-13. The on-chain path of a speculator's market order and MEV intervention (on-chain MEV-intervention mechanism-path illustration, not empirical data; the monthly sandwich extraction of about $10M → about $2.5M is discussed in the text, per EigenPhi [19] and Flashbots [20])

In a perpetual futures environment, the MEV exposure surface of an order is not limited to the traditional sandwich attack; it also includes the deliberate triggering of stop-losses. When a stop-loss order's trigger price is publicly visible on-chain, an attacker can precisely calculate the expected-value relationship between "the capital needed to push the price to the trigger" and "the profit obtainable after triggering." Only when this expected value is positive—as noted at the end of this paragraph, this generally holds only under the narrow conditions of shallow order-book depth at the target level and a high concentration of stop-loss/liquidation positions—does deliberately triggering a stop-loss become a rational attack strategy. Deliberately pushing the price to trigger others' stop-losses, however, is explicitly characterized as market manipulation under the major regulatory frameworks. U.S. SEC Rule 10b-5 renders illegal any fraudulent conduct intended to artificially affect market prices; Article 12 of the EU's Market Abuse Regulation (MAR) defines "manipulative trading" to include trading conduct aimed at creating artificial price levels; and the CFTC's anti-manipulation authority derives mainly from Section 6(c)(1) of the Commodity Exchange Act and Rule 180.1, which is modeled on SEC Rule 10b-5. The CFTC's 2023 enforcement action against Ooki DAO, by contrast, was not based on anti-manipulation grounds but rather established the precedent that a DAO can bear legal liability as an "unincorporated association," as well as the precedent of violations such as operating unregistered illegal leveraged trading; the significance of that case for the legal-personhood status of decentralized protocols is independent of the anti-manipulation rules themselves. As crypto-asset-specific regulations such as the EU's MiCA framework are implemented, the legal risk faced by participants who carry out such strategies—even on decentralized platforms—is steadily increasing. This book's analysis of the stop-loss-hunting strategy here is therefore purely for the academic purpose of understanding market microstructure, and readers should fully recognize the substantial legal risks involved

in engaging in such conduct. From the standpoint of economic analysis, the feasibility of deliberately triggering a stop-loss depends on the relationship between the attack cost and the expected return: the attacker must deploy enough capital to push the price into the stop-loss trigger region (the cost depends on the order-book depth at the target level), while the expected return depends on the total volume of stop-loss orders clustered in the trigger region and the size of the reverse position the attacker has established in advance. Only under specific conditions (namely, shallow order-book depth near the target level but a high concentration of stop-loss/liquidation positions) does such an attack have positive expected value.

Order type	Degree of MEV exposure	Exposure mechanism and attack vector	Mitigation
Large market order (public mempool)	Very high	The trading intent is fully public before filling, making it vulnerable to sandwich attacks and front-running. Searchers can precisely calculate the slippage tolerance and maximize the value extracted.	Use a private transaction pool (such as MEV Blocker); set strict slippage protection.
Stop-loss market order (on-chain trigger)	High	The stop-loss trigger price is publicly visible. In a transparent environment, this amounts to telling the attacker, "push the price here and you trigger my forced sell."	Use a "stop-loss + limit order" combination, or split a large stop into multiple small trigger orders.
Limit order (on-chain order book)	Medium	A long-resting limit order exposes the trader's strategic intent and may be used by counterparties to infer their liquidation price or overall capital strength.	Shorten the resting time (use GTT orders), or use conditional triggers where the protocol supports them.
Iceberg order (on-chain implementation)	Depends on implementation	A simple smart-contract implementation may expose the iceberg order's total size through its state-change pattern, causing the concealment to fail.	Rely on native protocol-level support from the application chain or on zero-knowledge-proof technology.

Table 7-8. The MEV exposure matrix of different order types (Data source: constructed by the author)

After Table 7-8 summarizes the qualitative risk level of each order type, Figure 7-14

goes further, using a heatmap to quantify the degree of exposure of eight common order types under five MEV attack vectors:

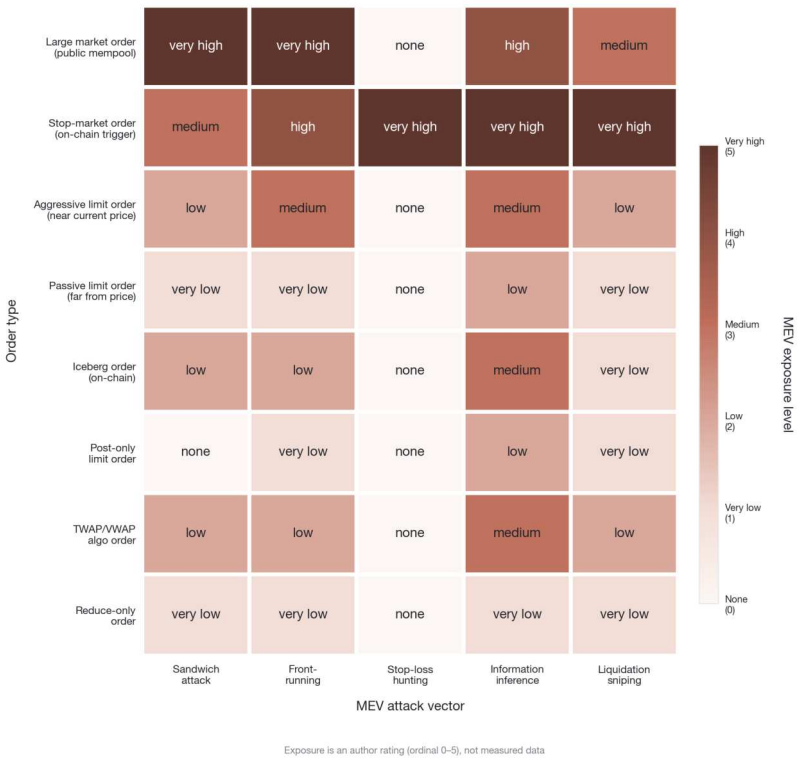


Figure 7-14. The degree of exposure of different order types under five MEV attack vectors (author's qualitative rating matrix; the 0–5 scale reflects the author's judgment, not empirical data; the MEV taxonomy is per Flashbots 2024 [20] and EigenPhi 2025 [19])

Figure 7-14 reveals a key pattern unique to the perpetual environment: although the stop-loss market order has only "medium" exposure on the sandwich-attack dimension, it reaches "very high" on both the deliberate-stop-triggering and information-inference dimensions—a risk-amplification effect unique to the perpetual futures environment. The "very high" exposure of large market orders on the sandwich and front-running dimensions, and the low exposure of Post-Only limit orders and Reduce-Only reduce orders across attack vectors owing to their passive nature, can all be read directly from the heatmap.

> Core insight: on an on-chain CLOB, the public visibility of the stop-loss order is directly related to the "transparency-fragility paradox" of Chapter 4. When your stop-loss price is seen by everyone and you use high leverage, "pushing the price to trigger your stop-loss" can become a positive-expected-value attack strategy for a market maker or an MEV bot. This is a systemic risk unique to perpetual futures in the transparent on-chain environment. At the same time, because deliberately triggering a stop-loss constitutes market manipulation under most regulatory frameworks, participants in such conduct

> also face steadily increasing legal risk.

7.5.4 The order-protection mechanisms of application chains

To address the negative effects of the on-chain order commitment property, a new generation of derivatives application chains (such as Hyperliquid and dYdX v4) has undertaken extensive innovation at the protocol layer, seeking a new balance point between "verifiability" and "strategic privacy." These application chains abandon the global-state model of general-purpose public chains such as Ethereum, opting instead for a highly customized architecture designed for order-book trading, thereby providing three core classes of order-protection mechanism.

The private mempool is the first class of protective measure these application chains adopt. On a general-purpose public chain, a transaction must be broadcast in a public P2P network before it is packed, giving MEV searchers a "submit-to-confirm" time window to observe and exploit pending orders. Many application chains, by contrast, adopt a private mempool or validator-level direct routing, encrypting the user's order or sending it directly to the current block proposer, so that it is not disclosed publicly before it is formally packed into a block and executed. This eliminates the MEV exposure in the "submit-to-confirm" time window and deprives the traditional sandwich attack of its target. The cost of this design, however, is that trust is transferred from the "trustless public network" to the validator nodes, and this transfer of trust introduces a non-negligible centralization risk. Under a private-mempool architecture, validators have complete visibility into and ordering power over the user's order flow, functionally no different from the operator of a CEX matching engine. A malicious or compromised validator can carry out a variety of concrete attacks: selectively delaying a particular user's orders to create an execution environment favorable to itself, front-running high-value orders, censoring (refusing to pack) transactions from particular addresses, and colluding with external parties to leak order-flow information. Take Hyperliquid as an example: its validator set is relatively small (about 4–16 nodes) and its admission mechanism is restricted, so compromising or bribing a single validator suffices to gain an informational advantage over all the order flow that validator processes. Mitigations currently under exploration include: threshold encryption (encrypting an order and distributing it to multiple validators, decryptable only when more than a threshold number of validators cooperate), commit-reveal schemes (the trader first submits a hash commitment of the order and reveals the order contents only after the block is confirmed), and periodic ordering-fairness audits. To assess the current state candidly: mainstream application-chain DEXs are far less decentralized than general-purpose public chains such as Ethereum, and their security model still relies largely on trust in a small number of validators rather than on cryptographic guarantees. From the standpoint of regulatory classification, a private mempool is functionally equivalent to a dark pool or a systematic internalizer in traditional finance, which are subject to extensive transparency and fairness regulation under the SEC and MiFID II frameworks—a regulatory equivalence that has not yet received sufficient attention in the current discussion of DeFi regulation.

Building on this, protocol-level deterministic ordering provides a complementary safe-

guard. General-purpose public chains typically use a priority-fee auction mechanism: whoever pays more gas has their transaction ordered first, which creates the economic basis for front-running, as an attacker need only pay slightly more gas than the target transaction to ensure their own transaction is executed first. Application chains, by contrast, typically enforce first-in-first-out or other deterministic ordering rules at the consensus level: as long as the order in which orders arrive at the validators is determined, their execution order is locked, thereby eliminating the possibility of prioritizing an order by paying a higher fee. Deterministic ordering, however, cannot fully eliminate the validator's internal ordering advantage; validators can still selectively delay or reorder the transactions they receive.

Going further, an application chain, by virtue of its full-stack control (from the consensus layer to the application layer), can also provide native support at the protocol layer for advanced orders such as iceberg orders and implement selective opacity. For example, it can record on-chain only the currently visible portion of an order while hiding the total size in encrypted state. This "selective opacity" design is precisely a response to the profound recognition that "full transparency is not net-positive in a high-leverage environment." It responds to the "transparency-leverage-safety trilemma" analyzed in Chapter 6: in some scenarios, moderate opacity can actually improve the overall safety and efficiency of the market. Zero-knowledge-proof technology offers a more elegant solution in this direction: "prove compliance without exposing the details" (see the analysis of the prospects for zero-knowledge proofs in order protection in Chapter 30).

The evolution of on-chain orders from "private intent" to "public commitment" is not merely a technological shift but an architectural transformation of market microstructure. It requires traders to relearn the "on-chain dialect" of this language—to enjoy the transparency and fairness that decentralization brings while learning to use the new infrastructure to protect their own strategies from strategy-exposure risk. And the application chains' ongoing exploration between "verifiability" and "strategic privacy" is defining the boundaries of this new architecture.

7.6 The execution-architecture spectrum and the reshaping of order semantics

In traditional financial markets, the lifecycle and execution logic of an order are built on highly centralized, low-latency physical infrastructure. The matching engine runs on dedicated hardware, order submission, matching, and cancellation complete at the sub-millisecond level, and the trader hardly needs to care about the physical characteristics of the underlying infrastructure. When perpetual futures trading migrates to an on-chain environment, however, the physical constraints of the underlying blockchain—the discreteness of block time, the latency of the consensus mechanism, the volatility of gas fees, and the costliness of state storage—reshape the semantics of an order. An order whose semantics are clear and whose behavior is predictable on a centralized exchange may exhibit a starkly different lifecycle, cost structure, and risk profile in an on-chain environment.

This section systematically examines the differences in the core properties of the three

main execution architectures (the centralized exchange, the fully on-chain DEX, and the application-chain DEX), and analyzes one by one how the physical constraints of blockchain reshape the economic meaning of an order along five key dimensions: order lifecycle, cancellation mechanism, conditional triggering, matching logic, and state management.

7.6.1 A comparison of the three execution architectures

Three execution architectures currently exist in the crypto derivatives market: the centralized exchange, the fully on-chain decentralized exchange, and the application-chain-based decentralized exchange. These three architectures exhibit a starkly different feature spectrum in latency, throughput, order cost, transparency, and censorship resistance, forming a continuum from "extreme performance but complete opacity" to "complete transparency but limited performance."

On a centralized exchange, the matching engine runs on a single high-performance server or a tightly coupled cluster, and order submission, matching, and cancellation typically complete at the sub-millisecond level. The matching engines of top-tier CEXs such as Binance and Bybit can reach an order-processing throughput on the order of a million orders per second, and order submission and cancellation incur no direct fees. This architecture offers extreme execution efficiency, but at the cost of entirely opaque operation: the trader cannot verify whether the matching engine has fairly executed the price-time priority principle, nor audit whether the exchange engages in improper conduct such as internal front-running or selective execution.

The fully on-chain DEX (such as protocols based on the Ethereum mainnet or a general-purpose Layer 2) goes to the other extreme. The submission, matching, and settlement of every order must, as an on-chain transaction, pass through the consensus confirmation of the entire network. The Ethereum mainnet's block time is about 12 seconds, and even an Optimistic Rollup such as Arbitrum requires a soft-confirmation time of about 250 milliseconds. This brings extremely high transparency and censorship resistance (anyone can independently verify the execution process of every transaction), but it also introduces significant latency and high, volatile gas costs. During periods of network congestion, a simple order submission may require paying tens of dollars in gas fees.

The application-chain DEX (such as dYdX v4 and Hyperliquid) seeks a new balance point between the two. By building a sovereign blockchain purpose-optimized for trading, with a customized consensus mechanism and in-memory order-book state management, this class of architecture can achieve sub-second latency and extremely high throughput. Hyperliquid's block time is about 200 milliseconds, and its official documentation claims it can process about 200,000 order operations per second (about 200,000 OPS) [14] (this figure is self-reported from Hyperliquid's official documentation and represents theoretical peak processing capacity rather than sustained throughput under stress; the JELLY incident of March 2025 exposed the platform's operational constraints under extreme market conditions, indicating that sustained processing capacity under actual stress may be significantly lower than the theoretical peak). dYdX v4, built on the Cosmos SDK, has constructed an independent application chain that

maintains the order book in the validators' memory rather than in on-chain state, thereby greatly improving performance while preserving non-custodial and verifiable properties [21]. Figure 7-15 uses a radar chart to compare these three architectures along six dimensions—latency, throughput, order cost, transparency, censorship resistance, and degree of decentralization: the CEX excels on the performance dimensions while its transparency/censorship resistance is nearly zero, the fully on-chain DEX is the reverse, and the application-chain DEX has the most balanced profile.

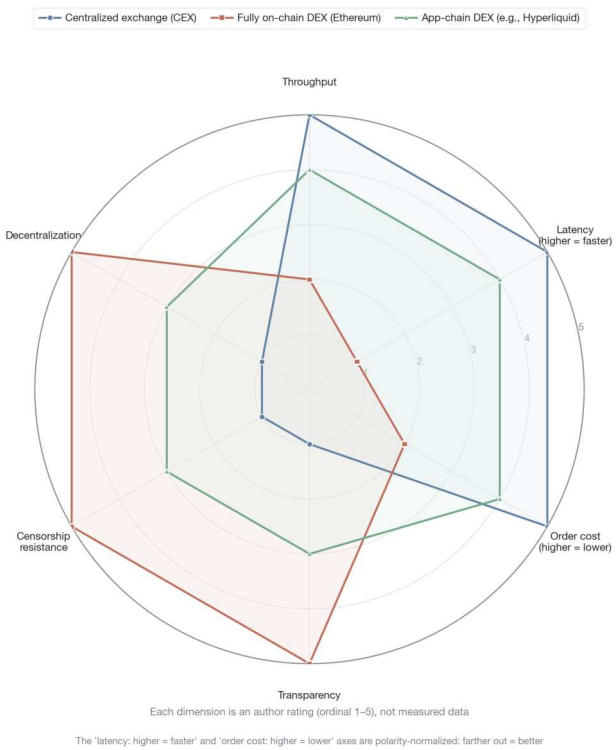


Figure 7-15. A comparison of the core properties of the three execution architectures (author's relative-rating radar; the 1–5 scale reflects the author's judgment, not empirical data; six dimensions, including degree of decentralization, consistent with the text; the latency/order-cost axes have been polarity-normalized so that farther out = better; platforms as of 2025, per Hyperliquid [14] and dYdX [21])

Property	Centralized exchange	Fully on-chain DEX	Application-chain DEX
Latency	< 1 ms	100 ms – 12 s	~200–1,000 ms
Throughput	> 1,000,000 TPS	~15–2,000 TPS	About 200,000 OPS (200,000 operations/second)
Order cost	Zero	High and unstable (gas fee)	Zero or extremely low
Order visibility	Private (visible only to the exchange)	Public (visible network-wide)	Configurable (short-term private/long-term public)
Matching mechanism	Continuous matching	Batch matching	High-frequency batch matching
Censorship resistance	Low	High	Medium
State-storage location	Centralized database	On-chain global state	Memory (short-term) + on-chain (long-term)

Table 7-9. A comparison of the key physical metrics of the three execution architectures (Note: "TPS" in the table refers to order-operation processing capacity, including operations such as order submission, cancellation, and matching, which differs from the transaction-level TPS of the underlying blockchain. The data for the application-chain DEX are the theoretical peak self-reported in official documentation, and sustained throughput under stress may be significantly lower than this value. The "censorship resistance: medium" rating is based on the following dimensions: the size of the validator set, the openness of the validator-admission mechanism, and whether users have a feasible alternative submission path when censored.) (Data source: compiled by the author)

Table 7-9 converts the qualitative comparison of the radar chart into concrete quantitative metrics. Most noteworthy is the order-of-magnitude difference: the CEX's latency is under 1 millisecond, whereas the fully on-chain DEX's latency may exceed 12 seconds—a gap spanning four orders of magnitude. This order-of-magnitude difference is not merely a matter of "fast versus slow"; it fundamentally determines the space of feasible order strategies under each architecture: the speed-sensitive high-frequency market-making and latency-arbitrage strategies that can be implemented at sub-millisecond latency lose essentially all economic viability at second-scale latency.

7.6.2 The order lifecycle

In traditional markets and on centralized exchanges, "good-til-canceled" is the most common order-lifecycle type. A trader can submit a GTC order, which will rest on the order book until it fills completely or the trader actively cancels it. The marginal cost of maintaining an order record in a centralized database is nearly zero, so there is no technical obstacle to implementing GTC.

In a blockchain environment, however, maintaining long-term on-chain state consumes costly storage resources. Every unfilled limit order occupies the disk space of full nodes, and if orders were allowed to exist permanently, the blockchain state would expand without limit, ultimately making node operating costs unbearable. Almost all on-chain order-book protocols have therefore abandoned GTC in its true sense, replacing it with order types that have a well-defined lifecycle.

The design of dYdX v4 provides a highly representative example. Its architecture explicitly distinguishes "short-term orders" from "stateful orders," which use starkly different mechanisms for lifecycle management [22]:

Short-term orders are aimed mainly at high-frequency market makers; their lifecycle is measured in block height, with a maximum survival time of only the current block height plus 20 blocks (about 20–30 seconds). Orders of this kind reside only in the validators' memory and are not written to on-chain state. If the network restarts or halts, these ephemeral orders are simply discarded. This feature introduces a noteworthy asymmetric risk in the chain-halt-and-restart scenario: short-term maker orders (a market maker's bid and ask quotes) are cleared on restart, while long-term stateful taker orders (such as stop-loss and take-profit orders) survive. If the external market price moved violently during the halt, the order book after restart will face a situation of severely insufficient depth while conditional orders can still be triggered—which may lead to a "restart flash crash" in which stop-loss orders fill at enormous slippage on an order book with almost no counterparty liquidity. Whether protocol designers should implement a cooling-off period or an auction-style reopening mechanism to mitigate this risk is a design question worth exploring in depth. This design appears "unreliable," but it is in fact a protective mechanism for market makers: after violent market volatility, stale market-making quotes that remain valid would expose the market maker to high adverse-selection risk. By letting orders expire automatically within an extremely short time, market makers can ensure that their quotes always reflect the latest market state.

Stateful orders are aimed at ordinary retail traders and institutional investors; their lifecycle is measured by timestamp (Good-Til-Time), with a maximum validity of up to 95 days. Orders of this kind are formally written to on-chain state and can be restored after a validator restart, ensuring the durability of long-term trading intent. Conditional orders (such as stop-loss and take-profit orders) also fall within the category of stateful orders.

> Core insight: the shift from GTC to GTT/GTB means that market makers can no longer "submit an order and then ignore its state" as on a CEX. They must develop dedicated infrastructure to continuously monitor the remaining lifecycle of on-chain orders and refresh them in time before expiry, which significantly increases strategic complexity and infrastructure cost.

Hyperliquid takes a more aggressive route. By virtue of the application chain's high throughput and lower state-management cost, Hyperliquid has once again made GTC viable in an on-chain environment, its design philosophy contrasting sharply with that of dYdX [14]. This difference itself reflects the different trade-offs that application-chain designers make between state-management cost and user experience.

Time-in-force type	Meaning	CEX support	Fully on-chain DEX	Application-chain DEX	Typical users
GTC	Valid permanently until filled or canceled	Standard	Generally unsupported	Partially supported	Retail traders, hedgers
GTB	Expires after a specified block height	Not applicable	Partially supported	Standard (short-term orders)	Market makers, high-frequency traders
GTBT	Expires after a specified timestamp	Not applicable	Partially supported	Standard (stateful orders)	Speculators, hedgers
IOC	Immediate-or-cancel	Standard	Complex to implement	Standard	Arbitrageurs
FOK	Fill-or-kill	Standard	Complex to implement	Standard	Arbitrageurs

Table 7-10. A comparison of how the time-in-force types are implemented across the different execution architectures (Data source: compiled by the author)

Table 7-10 reveals a key trend: the application-chain DEX has approached the completeness of the CEX in its support for time-in-force types, and the introduction of GTB and GTBT provides differentiated lifecycle-management tools for market makers and ordinary traders, respectively. The "complex to implement" label on IOC and FOK for the fully on-chain DEX, however, reveals a fundamental constraint: these two order types require immediate feedback (filled or canceled), and the consensus mechanism of a blockchain inherently cannot provide sub-second confirmation feedback—which constitutes a structural bottleneck in the functional completeness of the fully on-chain CLOB.

7.6.3 The order-cancellation mechanism

The semantics of the cancellation operation undergo a significant change in an on-chain environment. On a centralized exchange, cancellation is a free and instantaneous operation. High-frequency market makers frequently use the strategy of "placing large numbers of orders and quickly canceling them before they fill" to probe market liquidity and manage inventory risk. By some estimates, the cancel-to-fill ratio of market makers on top-tier CEXs typically exceeds 95% (this figure is a widely cited industry rule of thumb; the specific ratio varies significantly by trading pair, exchange, and time period—see the SEC's statistics on cancellation rates in the U.S. equity market and the publicly disclosed market-maker-activity reports of various exchanges), meaning that of every 100 orders submitted, more than 95 are ultimately canceled rather than filled.

In a fully on-chain architecture, however, cancellation is no longer free. Every cancellation instruction is an on-chain transaction that must be broadcast, queued, packed, and charged gas. During periods of Ethereum network congestion, the gas cost of canceling an order may be as high as tens or even hundreds of dollars. More seriously, because of the confirmation latency of on-chain transactions, there is a non-negligible time window

between the broadcast of a cancellation instruction and its ultimate effect. Within this window, the order the trader intended to cancel is still "active" and may be targeted for a fill by other traders (especially MEV searchers and arbitrage bots).

> The asymmetry of latency and cost in cancellation significantly weakens a market maker's ability to respond to adverse selection. This portion of extra risk is ultimately passed on to all liquidity consumers in the form of wider bid-ask spreads, becoming an implicit cost of on-chain trading.

To mitigate this problem, application-chain architectures adopt differentiated approaches. For short-term orders, dYdX v4 propagates cancellation instructions rapidly among validators via a gossip protocol, handling them on a "best-effort" basis [22]. This mechanism does not guarantee the absolute success of a cancellation (under network congestion or in the presence of a malicious validator, the cancellation instruction may fail to take effect in time), but it greatly reduces latency and cost, allowing market makers to adjust their quotes at a frequency close to that of a CEX. For stateful orders, cancellation still requires a formal on-chain transaction, but thanks to the application chain's high-frequency block production (such as Hyperliquid's 200-millisecond block time), the latency is compressed to within 1–2 blocks. Figure 7-16 uses a sequence diagram to compare the complete cancellation flow under the three architectures and the processing steps and time overhead of each stage.

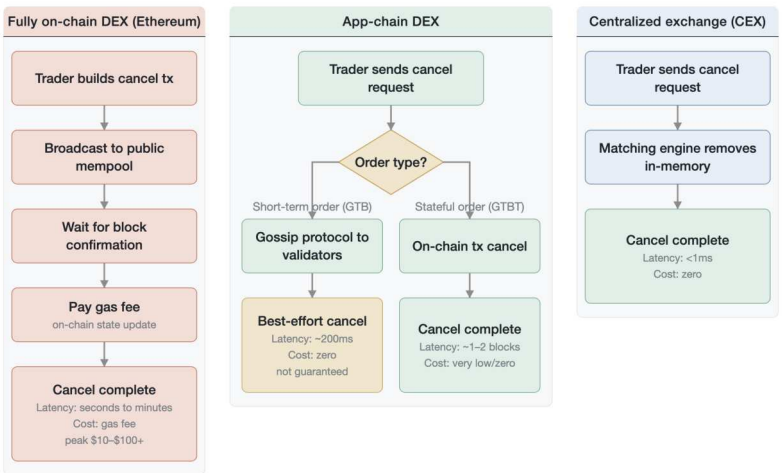


Figure 7-16. The evolution of the order-cancellation flow under the three execution architectures (cancellation-flow mechanism illustration, not empirical; the short-term-order gossip "best-effort" is per dYdX [22], and the fully on-chain gas of $10–$100+ is a peak-period magnitude)

Figure 7-16 shows the qualitative differences in the cancellation flow, and Figure 7-17 quantifies this difference along the two dimensions of latency and cost on a logarithmic scale (see the figure caption for specific magnitudes).

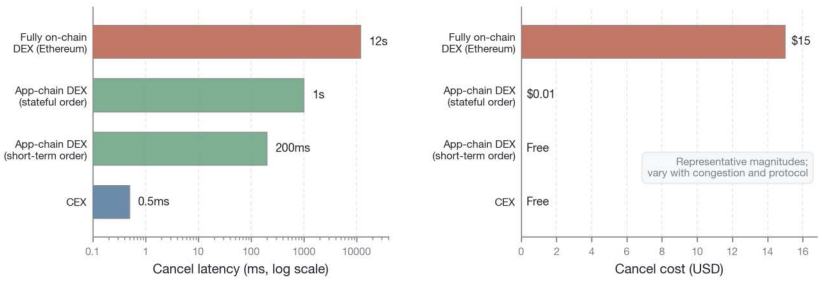


Figure 7-17. A quantitative comparison of cancellation latency and cancellation cost (representative-magnitude comparison illustration: 12s/$15, 1s/$0.01, 200ms/free, 0.5ms/free are order-of-magnitude representative values, not empirical; mapping to the text and to [22][14])

The quantitative comparison in Figure 7-17 reveals a conclusion with profound implications for market makers: on a fully on-chain DEX, the dual pressure of cancellation latency and cancellation cost makes the "place many orders, cancel quickly" style of market making common on a CEX all but infeasible. Although the application-chain DEX achieves a significant improvement on the latency dimension (compressing it from seconds to hundreds of milliseconds), it still lags the CEX's sub-millisecond level by two to three orders of magnitude. This gap directly dictates that on-chain market makers must use wider quote spreads and fewer quote layers to compensate for the structural disadvantage of being unable to cancel quickly.

7.6.4 Conditional-order triggering

The execution logic of conditional orders (such as stop-loss and take-profit orders) undergoes an important change in an on-chain environment. On a CEX, a conditional order is triggered directly by the internal matching engine according to the last traded price generated in real time, and the entire process from the price touching the trigger threshold to the order entering the matching queue completes at the microsecond level. In on-chain derivatives protocols, however, a smart contract cannot directly obtain external market prices, and the triggering of a conditional order generally relies on a price feed provided by an external oracle.

This reliance on the oracle introduces two key physical constraints: discontinuity and inherent latency.

Oracle price updates are discrete rather than continuous, and oracles differ fundamentally across execution architectures in their update frequency and mechanism. In DeFi protocols on general-purpose public chains, conditional orders typically rely on third-party oracle networks such as Chainlink. Chainlink, for example, generally pushes a new price only after the price deviation reaches a certain threshold (such as 0.5%) or a specific time interval elapses (such as the heartbeat period, which varies by feed and chain—for mainstream pairs on the Ethereum mainnet it is often about 1 hour, and for some it can reach 24 hours) [23]. This means that in a general-purpose-public-chain environment, the price changes the on-chain system perceives exhibit a pronounced stair-step character. Mainstream application-chain DEXs, however, adopt a starkly

different oracle architecture. dYdX v4 uses a custom oracle module: validator nodes aggregate price data from the APIs of multiple centralized exchanges and reach price consensus during the consensus process of each block, achieving one update per block (about 1–2 seconds). Hyperliquid runs an internal oracle network in which validators publish a price about once every 3 seconds (first taking a cross-exchange weighted median, then aggregating among validators by stake weight); although its update frequency is faster than that of third-party oracles on general-purpose public chains, it is still significantly slower than its roughly 200-millisecond block production. The update frequency of these custom high-frequency oracles is far higher than Chainlink's, so the conditional-order trigger latency of an application-chain DEX is far lower than that of a general-purpose-public-chain DeFi protocol. Chainlink's parameters (a 0.5% deviation threshold and a 1-minute heartbeat) apply only to describing DeFi protocols on general-purpose public chains and should not be generalized as the oracle characteristics of all on-chain derivatives platforms.

Despite the order-of-magnitude difference in oracle update frequency, all on-chain oracles face a more serious security threat: the oracle-manipulation attack. An attacker can use a flash loan to temporarily manipulate the price source on which the oracle relies within a single transaction, artificially creating an extreme price deviation and thereby triggering stop-losses and liquidations en masse. The Mango Markets incident of October 2022 is a representative case of this attack vector: the attacker manipulated the oracle price of Mango Markets, artificially inflating the price of the MNGO token and then using the fake collateral value to borrow protocol assets, causing a loss exceeding $110 million (over $110 million per the CFTC/DOJ figure; about $116 million per the SEC) [24]. In a perpetual futures system that relies on an oracle to trigger stop-losses and liquidations, a manipulated oracle price can trigger thousands of stop-losses and liquidations simultaneously, causing a catastrophic market impact. The main mitigations include: adopting a TWAP (time-weighted average price) oracle to smooth out instantaneous manipulation; aggregating multiple independent price sources and discarding outliers; and activating a circuit breaker when the price deviation exceeds a threshold, suspending liquidation and stop-loss triggering.

When the market swings violently, the compounding effect of discontinuity and latency may produce adverse consequences, but the severity varies by oracle architecture. In general-purpose-public-chain DeFi protocols, oracle update latency can reach seconds or even tens of seconds, and the lag in stop-loss triggering is especially pronounced. On application-chain DEXs, thanks to custom high-frequency oracles, the trigger latency is compressed to about 200 milliseconds to 1 second, but it still cannot be entirely eliminated. Consider a concrete scenario: with ETH at $2,000, a stop-loss order is set with a trigger price of $1,900. If the market plunges to $1,800 in an extremely short time, then on a general-purpose public chain the oracle may update only once, to $1,850, so the stop-loss triggers at $1,850, by which point the actual market price may already be far below $1,850; on an application chain such as Hyperliquid, the oracle updates at a second scale (about 3 seconds), still faster than the tens-of-seconds third-party oracles on general-purpose public chains, so the trigger lag is relatively small but not eliminated. Furthermore, after a conditional order is triggered, it generally must wait for the next block before it can enter the matching engine as a market order or

a limit order. On dYdX v4, conditional orders are checked and triggered during the EndBlocker stage of block processing (the final processing step of the current block), and the resulting market/limit order does not participate in matching until the next block—an inherent latency of at least one block from trigger to execution [22].

> In extreme conditions, the price may swing more than 5% within 1 second. A 1-second latency is enough to let a stop-loss order intended to cap a loss at 10% ultimately fill at a loss of 20% or more. This is an implicit risk that on-chain traders must understand and manage.

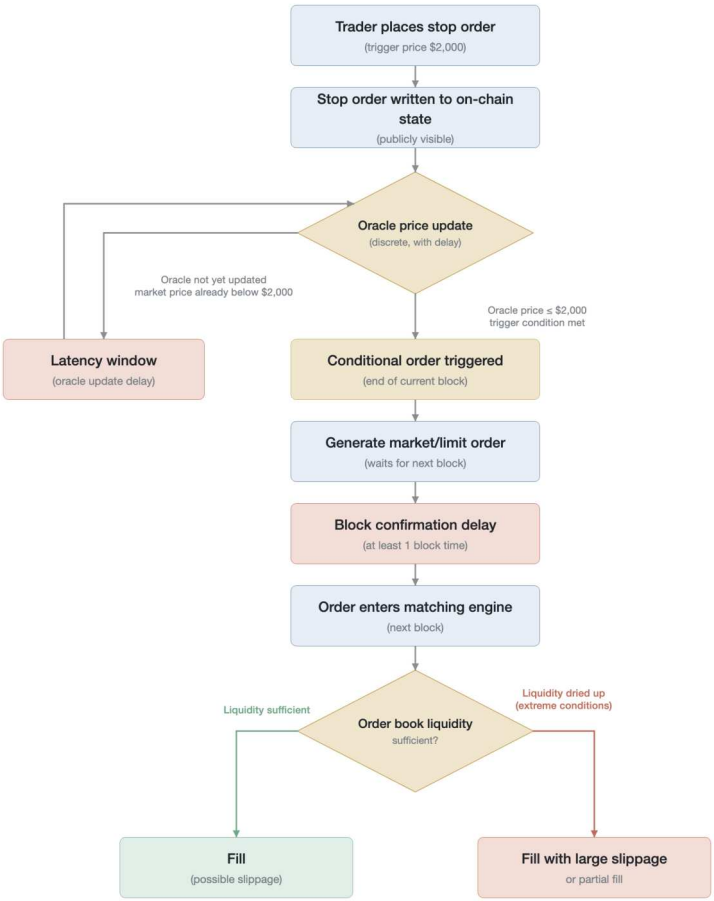


Figure 7-18. The trigger and execution path of an on-chain stop-loss order (on-chain stop-loss trigger-execution mechanism illustration: the $2,000 trigger price is an illustrative value, not empirical; the "at least 1 block" is per the dYdX EndBlocker [22]; the 5%/10%/20% slippage tiers are discussed in the text)

Figure 7-18 uses a timeline to show, in full, all the stages of an on-chain stop-loss order from dormancy to final fill. The time interval between each stage may introduce

slippage, and in extreme conditions the compounding of these intervals may cause a deviation of hundreds or even thousands of basis points between the final fill price and the trigger price.

7.6.5 The order-matching mechanism

On a centralized exchange, the matching engine uses a continuous-time limit-order-book model, matching orders at the instant they are received according to the "price priority, time priority" principle. Between two orders at the same price, the one that arrives a fraction of a microsecond earlier gains priority to fill. This continuous-matching mechanism updates the order-book state in real time as each new order arrives, and the price-discovery process appears as a smooth, continuous trajectory.

The discrete-time nature of blockchain fundamentally changes this matching paradigm. Because state updates proceed in units of blocks, the minimum time resolution is no longer the microsecond but the duration of one block. On the Ethereum mainnet this interval is about 12 seconds; on an application chain such as Hyperliquid, about 200 milliseconds. All orders arriving within the same block are treated as "submitted simultaneously" on the time dimension, and priority cannot be established by arrival time. Budish, Cramton, and Shim (2015) note that such discretization can, in theory, transform speed-based competition into price-based competition [25], and the block-time nature of blockchain bears a structural resemblance to that conception.

The matching implementations of current on-chain perpetual futures order books fall mainly into three modes, each representing a different response strategy to the discrete-block constraint.

dYdX v4 adopts a proposer-matching mode: the order-book state is maintained in the validators' memory, the matching process is completed off-chain by the proposer during the block-proposal stage, and the matching result is then submitted as part of the block to the other validators for consensus verification [22]. This design keeps the matching logic itself following the price-time priority principle, but the arrangement order of orders within a block is determined by the proposer rather than strictly fixed by arrival time.

Hyperliquid adopts a validator-embedded-matching mode: its validator nodes simultaneously play the role of the matching engine, orders enter the matching queue at sub-second latency after arriving at a validator, and through high-frequency block production (about 200 ms/block) and a first-in-first-out ordering rule, the matching behavior approaches the CEX's continuous model on the time-granularity dimension [14]. However, "first-in-first-out" is an ill-defined concept in a distributed system: when multiple validators receive orders submitted by different traders at nearly the same time, the definition of "arrived first" depends on the local timestamp of the leader validator (the block proposer) rather than on a globally verifiable objective order. In the absence of a verifiable ordering proof, the credibility of a FIFO claim ultimately depends on trust in the honest behavior of the current leader validator—which is no fundamentally different from the trust model of a CEX matching engine.

Injective adopts a frequent-batch-auction mode, whose design logic is fundamentally

different from the previous two. Within each block's time window (about 1 second), all submitted orders are sealed—neither disclosed publicly nor entered into a visible order book. At the end of the block, the protocol performs a uniform clearing of all crossing buy and sell orders within that window: the system computes a single clearing price that maximizes traded volume, and all filled orders execute at that price [26]. When the resting quantities of buyers and sellers at the clearing price are asymmetric, the smaller side fills in full and the larger side fills proportionally in part. This mechanism makes the submission order of orders irrelevant to the fill price, thereby transforming the competitive dimension from "who is faster" to "whose price is better," and eliminating at the protocol layer the economic basis for ordering- and speed-based front-running within a block (though cross-block and cross-venue informational front-running, and the potential peeking at sealed orders by validators, are not covered by this mechanism). The frequent-batch-auction theoretical framework proposed by Budish, Cramton, and Shim (2015) [25] is precisely the academic precursor of this design (a derivatives batch-auction mechanism is functionally equivalent to a traditional exchange's call auction and may, in jurisdictions such as the United States, trigger exchange-registration requirements and related compliance obligations). This theory originally targeted the high-frequency-trading arms race in equity markets, and its core assumptions do not necessarily hold in full in the crypto perpetual futures market; the boundaries of its applicability deserve attention when it is cited across domains. Specifically, the model of Budish et al. assumes the following conditions, all of which face challenges in the on-chain perpetual futures market: (1) homogeneous competition among a large number of market makers, whereas the number of active market makers on an on-chain DEX is in fact far smaller than on a traditional securities exchange; (2) discrete arrival of fundamental information through public news, whereas the information sources of the crypto market are dispersed and include the distinctive dimension of oracle latency; and (3) arbitrage activity occurring mainly among different trading venues for a homogeneous asset, whereas cross-chain arbitrage in the crypto market involves additional complexities such as bridging latency and settlement finality.

The three modes exhibit a gradient of difference in matching fairness. In the dYdX v4 architecture, because the block proposer has discretion over the ordering of transactions within a block, the execution order of orders within the same block is not fully deterministic, which affords the proposer a certain ordering advantage. Hyperliquid, by embedding the matching logic in the consensus-protocol layer and adopting a first-in-first-out rule, narrows this ordering-discretion space, but validators still have the ability to fine-tune the orders they receive. Injective's frequent batch auction, through its uniform clearing price and sealed-order mechanism, provides the strongest guarantee of ordering neutrality among the three, and the economic value of ordering power approaches zero under the uniform-clearing mechanism. The cost of this design is a sacrifice of matching time-granularity: traders cannot obtain immediate fill feedback within the auction window, and all orders must wait until the window closes to be processed uniformly.

Another layer of impact that discrete block time has on the matching mechanism is reflected in the order book's update frequency. On a CEX, every minute change in the order book can be observed in real time, and market makers can adjust their

quotes at the microsecond level. In an on-chain environment, however, the order-book state updates only at block boundaries, and between two blocks the order book is "frozen." This means on-chain market makers cannot precisely manage inventory risk and spreads through high-frequency fine-tuning as on a CEX, but must submit their quoting strategy all at once within each block window. This constraint directly affects the depth structure and spread levels of the on-chain order book, which the next chapter analyzes further from the perspective of order-book microstructure.

7.6.6 State-management constraints

The costly state-storage resource of blockchain imposes strict physical constraints on order-book design. On a CEX, a trader can maintain thousands of active limit orders simultaneously at no extra cost. In an on-chain environment, however, every active order occupies precious state space, and the scarcity of storage resources gives rise to a series of design limitations absent on a CEX.

The most direct constraint is a limit on the number of orders. To prevent state bloat and denial-of-service attacks, on-chain protocols must set clear boundaries on trader behavior. Paradex, for example, explicitly stipulates that each trader can maintain at most 100 unfilled resting orders in a single market [27]. On platforms with a higher market share, Hyperliquid sets an upper limit on the number of active orders per address in a single market (the specific value is dynamically adjusted by market and account tier), and dYdX v4 likewise sets an upper limit on the number of active orders per subaccount (typically on the order of hundreds). This hard limit forces market makers to manage their order flow more precisely, abandoning the strategy—common on a CEX—of densely posting orders across dozens of price levels, in favor of a more concentrated and efficient quoting approach.

In addition, the on-chain order book faces a distinctive "dust attack" threat. Usman (2025), in Hacken's security research, describes this attack vector in detail [28]: an attacker uses transactions of an extremely small amount (such as 1 wei, the smallest divisible unit) to partially fill a target order. Because the smart contract must perform complex rounding and state updates when handling a partial fill, this tiny fill may cause a variety of problems: loss of computational precision, invalidation of the remaining order amount when it falls below the minimum trade unit, or even—in a poorly designed protocol—the state of the original large order being erroneously marked as "completed," nullifying the limit-order provider's liquidity. The dust attack is merely the tip of the iceberg of the smart-contract security surface facing on-chain order books. Broader attack vectors also include: reentrancy attacks (repeatedly triggering state updates within an order-fill callback), integer overflow/underflow (producing unexpected results in high-precision price and quantity computations), and state-bloat denial-of-service (creating large numbers of tiny orders to push the contract's storage and computation costs beyond the gas limit). Protocol developers should consult the on-chain-order-book audit reports published by professional security-audit firms (such as Trail of Bits, OpenZeppelin, and Hacken) to systematically assess and defend against these attack surfaces.

To defend against attacks that exploit state-management vulnerabilities, protocol de-

velopers must introduce additional defensive mechanisms: mandating a minimum order size (such as $10 equivalent in tokens), setting a minimum fill ratio (such as at least 1% of the order's remaining amount), automatically zeroing out an order when its remaining amount falls below a specific threshold, or charging an extra processing fee for partial fills of extremely small amounts. This logic, added to adapt to the on-chain physical environment, further increases the complexity and audit difficulty of the smart contract.

> State bloat is the core challenge of the on-chain order book. All mitigations (order expiry, quantity caps, minimum size) weaken, to varying degrees, the competitiveness of the on-chain order book relative to the CEX. Application chains mitigate this problem through higher throughput and lower state cost, but cannot eliminate it entirely.

The table below summarizes the behavioral differences of the core order types across the three execution architectures, providing traders with an intuitive reference framework for choosing order strategies in different environments:

Order type/property	Centralized exchange	Fully on-chain DEX	Application-chain DEX
Limit-order lifecycle	GTC, valid permanently	Must use GTT/GTD	Short-term GTB / long-term GTBT or GTC
Limit-order cancellation	Instant and free	Requires sending a transaction, paying gas, awaiting confirmation	Short-term best-effort free / long-term requires an on-chain transaction
Market order	Instant fill, low slippage	High slippage + MEV sandwich risk	Medium slippage, lower MEV risk
Post-Only	Standard support	Requires smart-contract logic checks	Standard support
IOC	Standard support	Complex to implement, high gas cost	Standard support
FOK	Standard support	Complex to implement	Standard support
Stop-loss market order	Based on internal fill price, instant trigger	Based on oracle, at least 1-block latency	Based on oracle, at least 1-block latency
Trailing stop	Standard support	Extremely complex to implement, generally unsupported	Partially supported
TWAP	Via API or third-party tools	Generally unsupported	Some native support
Iceberg order	Standard support	On-chain transparency renders it ineffective	Possibly supported (depends on the order-visibility design)
Maximum active orders	No practical limit	Constrained by gas and state storage	Clear upper limit (such as 100 per market)

Table 7-11. The behavioral differences of the core order types across the three execution architectures (Data source: compiled by the author)

The core conclusion of Table 7-11 is that the application-chain DEX has approached the functional completeness of the CEX for the vast majority of order types, but a

gap remains in advanced functions such as trailing stops, TWAP, and iceberg orders. Constrained by gas cost and state-storage limits, the fully on-chain DEX has difficulty implementing order types that require immediate feedback, such as IOC and FOK. This functional difference directly affects the platform choices of different types of traders: high-frequency market makers tend toward the application-chain DEX, whereas traders with a strict need for censorship resistance may still have to choose a fully on-chain protocol.

7.7 The order strategies of market participants

Orders do not exist as isolated technical instructions; they are the tools with which different market participants express intent, compete for advantage, and manage risk within a shared trading grammar. In the ecosystem of the decentralized perpetual futures market, market makers, arbitrageurs, speculators, hedgers, and liquidators each play a key role, bear different risks, and pursue different return objectives. The impact of an order radiates outward along four dimensions: funding-rate transmission, the liquidation chain, microstructure signals, and on-chain MEV risk exposure. Understanding the order preferences of different participants not only helps illuminate the operating mechanics of market microstructure but also makes it possible to infer, in reverse, the current market state and the dominant forces.

This section analyzes the characteristics of five core classes of market participant, examines how they use different order types and strategies to achieve their business objectives, and finally shows how the dominant participants in a market can be identified through order-flow analysis.

7.7.1 The order strategies of market makers

Market makers are the primary providers of liquidity on decentralized exchanges and a core component of order-book depth and market continuity. Their core business model is to earn the bid-ask spread and the liquidity incentives provided by the exchange (such as maker-fee rebates or token rewards) by simultaneously providing quotes on both the bid and ask sides. Because market makers generally remain market-neutral and do not actively take directional risk, their order strategies depend heavily on low latency and the ability to adjust at high frequency.

In an on-chain environment, a market maker's most core tools are the limit order (especially with the Post-Only attribute) and the short-term order. The Post-Only attribute (see Section 7.3.3) is decisive for market makers: it not only guarantees that they always enjoy the maker-side fee benefit (on many DEXs the maker fee is zero or even negative) but also avoids the unexpected slippage that arises when a rapid market price movement causes a limit order to "cross" and become a market order.

Market makers densely place limit orders at multiple price levels on the order book, forming a multilayered quote structure to build market depth. When the net position deviates from neutral, a market maker widens the spread on one side, thickens the resting orders on the other side, or uses small aggressive orders to return quickly to neutral. To cope with rapid market-price fluctuations, market makers need to cancel

old orders and place new ones frequently. As discussed in the previous section, on DEXs with an application-chain architecture (such as dYdX v4), market makers make heavy use of short-term orders that expire based on block height [22]. Such orders exist only in memory and expire automatically within tens of seconds, significantly reducing the risk of "stale orders being targeted for attack" due to network latency or congestion.

In the perpetual futures market, market makers also face a distinctive dual inventory risk: not only the directional risk from movements in the underlying asset's price but also the position-carrying-cost risk from the funding rate. If a market maker holds a large long position and the current funding rate is negative (longs pay shorts), the continuously paid funding will erode market-making profit. A more common practice, therefore, is "perpetual market-making + spot hedging"—providing liquidity in the perpetual futures market while establishing an opposite-direction position in the spot market, thereby isolating directional risk. This hedging strategy, however, faces several non-negligible execution risks in practice. The first is basis risk: the basis between the perpetual futures price and the spot price fluctuates continuously, so the hedge is not perfect. The second is cross-platform margin lock-up: when the perpetual leg is in profit while the spot leg is at a loss (or vice versa), the funds on the profitable leg cannot be reallocated in real time to the losing leg to top up margin, potentially causing one leg's position to be liquidated while the "hedge" on the other leg fails. Most serious is the operational risk during a crisis: an exchange outage, a withdrawal delay, or a cross-chain-bridge failure may prevent the spot-leg position from being adjusted in time, leaving the perpetual position unprotected at precisely the moment hedging protection is most needed. In addition, market makers make wide use of algorithmic strategies such as time-weighted average price or iceberg orders, splitting large liquidity into multiple small orders to hide their true position intent.

7.7.2 The order strategies of arbitrageurs

Arbitrageurs are a key force maintaining cross-market price consistency, playing the role of "price calibrator" in market microstructure. By capturing minute price differences of the same asset across different exchanges (such as between a CEX and a DEX, or between different DEXs), they buy low and sell high to obtain risk-free or low-risk profit. Arbitrage opportunities are typically fleeting, so speed and execution certainty are the arbitrageur's core requirements.

An arbitrageur's preferred tools are the market order and limit orders with the immediate-or-cancel or fill-or-kill attribute. The market order offers the strongest fill certainty: regardless of the current order-book state, a market order immediately takes the available liquidity. An IOC order takes the liquidity within the fillable depth and then automatically cancels the unfilled remainder, avoiding the one-sided-exposure risk the arbitrageur would face if one leg fills while the other does not. An FOK order is stricter, requiring the order to fill in full or else cancel entirely, guaranteeing the atomicity of the arbitrage strategy.

In the perpetual futures market, arbitrageurs enjoy a rich variety of opportunity sources: basis arbitrage between perpetual futures and spot, cross-exchange price-difference arbitrage, and a cross-chain combination of price-difference and funding-rate-difference

arbitrage. When an arbitrage bot detects that the price on an on-chain DEX lags that on a CEX, it immediately sends an IOC order to the DEX. If network latency means the price has already moved unfavorably by the time the order arrives, the IOC attribute ensures that the portion unable to fill immediately is automatically canceled, preventing a failed arbitrage from turning into one-sided exposure.

The on-chain environment, however, has rewritten the arbitrageur's profit model. Gas cost becomes a non-negligible fixed cost: every arbitrage transaction incurs a gas fee, which means only arbitrage opportunities whose price gap exceeds the gas cost are worth executing. More seriously, MEV front-running may cause an arbitrageur to pay gas yet fail to fill, or to fill after the price has already been pushed to a worse level. On-chain arbitrageurs are therefore increasingly turning to private transaction channels to protect their arbitrage intent. The two must be distinguished. MEV-protection infrastructure on Ethereum (such as Flashbots Protect and MEV-Share) applies only to Ethereum and its Layer 2 ecosystem, and its trust model is built on the competitive mechanism of the block-builder market. On application-chain DEXs (such as Hyperliquid and dYdX v4), the relevant MEV-protection mechanisms are the validator private mempool and protocol-level deterministic ordering, whose trust model is entirely different, relying on trust in the honest behavior of a small number of validators rather than on market competition (see the detailed analysis in Section 7.5.4). Conflating the two would mislead practitioners.

7.7.3 The order strategies of speculators

Speculators, often called directional traders or retail traders, are the primary contributors of market volatility and trading volume. Based on analysis of market fundamentals, technicals, or macroeconomic events, they actively take directional risk in the hope of obtaining leveraged gains from a price rise or fall. Speculators are the most active and diverse group of participants, and their core constraints include not only structural factors (informational disadvantage, slippage cost, the chain effect of stop-loss triggering, and on-chain MEV exposure) but also behavioral factors: revenge trading (rushing to "win it back" after a loss by increasing position size and leverage), FOMO-driven chasing (buying in with a market order during a sharp price surge for fear of missing out), and the endowment-effect-driven refusal to close losing positions (treating an "unrealized loss" as a "loss not yet confirmed" rather than an economic loss that has already occurred). These behavioral biases systematically undermine the speculator's risk-management discipline.

The order strategies of speculators are the most varied, but their core demands can be summarized in two keywords: "entry" and "risk control." To capture market trends quickly, speculators frequently use market orders to take liquidity directly (as takers). Mature speculators, however, are more inclined to use conditional orders—especially stop-loss and take-profit orders—to achieve automated risk management.

In on-chain perpetual futures trading, because they cannot monitor the market 24 hours a day, speculators rely heavily on stateful conditional orders to protect their leveraged positions. For example, when going long on Ethereum, a speculator simultaneously sets a stop-loss order with a trigger price below the current market price. Once the

oracle price falls below that threshold, the stop-loss order is activated and converts into a market order, forcibly closing the position to prevent losses from widening. This "set-and-forget" strategy is generally implemented on an application-chain DEX through a months-long GTBT stateful order.

The special risks speculators face in an on-chain environment cannot be ignored. They are the primary source of MEV's "extractable-value order flow": their market orders and trigger orders are visible on-chain and more easily fall prey to sandwich attacks. The final fill price of an on-chain market order depends not only on order-book depth but also on the outcome of the game among MEV bots in the same block. Typical countermeasures include: using MEV-resistant routing (such as submitting transactions through a private mempool), setting reasonable slippage-protection parameters, and choosing an application-chain DEX with a protocol-level MEV-protection mechanism.

7.7.4 The order strategies of hedgers

Hedgers participate in the perpetual futures market not to obtain absolute returns but to transfer or hedge their risk exposure in the spot market or other lines of business. Typical hedgers include miners (hedging the price risk of mining revenue), token project teams (hedging the value volatility of treasury assets), cross-chain-bridge operators (hedging the exchange-rate risk of liquidity pools), and institutional investors holding large spot positions.

Hedgers typically trade in large sizes and are highly sensitive to the execution price but have relatively low requirements for execution speed. Hedgers therefore rarely use market orders to strike the order book directly, because a large market order would produce considerable market impact and slippage. Instead, they are more inclined to use limit orders (GTC or GTBT) and algorithmic orders (such as TWAP and VWAP).

By splitting a large hedging need into smaller child orders and executing them gradually over a longer time window, a hedger can minimize the impact of their trading on the market price. The splitting logic of TWAP and VWAP was detailed in Section 7.3.4 (the former splits evenly by time, the latter weights by the historical volume distribution) and is not repeated here.

In the perpetual futures market, the two most common structures for hedgers are: spot long + perpetual short (the standard hedging structure for miners and institutions) and the funding-rate trade (when the funding rate is high, shorting perpetual futures while going long spot to earn the rate differential). In addition, the Reduce-Only attribute is especially important for hedgers: it ensures that an order can only reduce an existing position and will not accidentally increase speculative exposure, preventing the creation of new risk exposure due to an operational error when adjusting the hedge ratio.

7.7.5 The order strategies of liquidators

Liquidators (or the liquidation mechanism) are key participants in the decentralized leveraged-trading ecosystem and the core executor of the solvency-safeguard mechanism. When a trader's margin ratio falls below the maintenance-margin requirement, the liquidation mechanism forcibly takes over and closes the defaulting position, thereby

protecting the protocol from bad-debt risk. The liquidation architecture of current on-chain derivatives protocols, however, comes in two starkly different modes that must be distinguished.

The first is the external-liquidator-competition mode, seen mainly in DeFi lending protocols (such as Aave and Compound) and early derivatives protocols (such as dYdX v3). In this mode, liquidators are independent third-party participants who must monitor the health factor of all on-chain accounts in real time and, the moment they find an account whose margin ratio has fallen below the liquidation threshold, call the protocol's liquidation function as fast as possible. This liquidation operation is, in effect, equivalent to a forced market order. In return, the liquidator generally receives a portion of the liquidation penalty as a profit incentive. The behavioral pattern of liquidators closely resembles that of arbitrageurs: both are highly time-sensitive participants who pursue execution certainty.

The second is the protocol-built-in-liquidation-engine mode, adopted by current mainstream application-chain DEXs. Hyperliquid uses an internal liquidation engine: when an account's margin ratio falls below the threshold, the protocol automatically assigns the defaulting position to a backstop liquidity pool for handling, without the participation of external liquidators bidding. dYdX v4's liquidation is likewise executed automatically by the protocol during block processing: validators check the margin status of all accounts during each block's EndBlocker stage and automatically initiate liquidation for accounts that fail to meet the maintenance-margin requirement. In this mode, the concept of the "liquidator" is internalized as the protocol's automated risk-management module rather than an external competitor.

Among protocols on general-purpose public chains that use the external-liquidator-competition mode (such as Ethereum-based DeFi lending protocols), competition among liquidators is especially fierce during violent volatility. Large numbers of accounts hit their liquidation lines simultaneously, and liquidation opportunities emerge and vanish within an extremely short time. To be the first to complete a liquidation, liquidators participate in MEV bidding, raising the gas fee to ensure their liquidation instruction is packed first. On Ethereum, the competition mainly takes the form of fierce gas bidding; on high-throughput chains such as Solana, it takes the form more of a latency race and infrastructure competition. On application-chain DEXs that use a protocol-built-in liquidation engine, the liquidation process is executed automatically by the protocol, eliminating competition and gas bidding among external liquidators; but its processing capacity may face a throughput bottleneck in extreme conditions: when the number of accounts that need liquidation within a single block exceeds the engine's processing limit, some liquidations may be deferred to subsequent blocks.

Whatever liquidation architecture is used, forced liquidation raises several unresolved legal issues. Under the consumer-protection framework of traditional finance, forced position closure generally requires specific contractual authorization, prior notice, and a grace period. The automated liquidation of on-chain protocols lacks a notice mechanism and a grace period, and the fairness of the liquidation penalty likewise lacks regulatory review. As frameworks such as MiCA gradually bring crypto-asset service providers within the scope of consumer-protection obligations, the legal status of forced

liquidation (including the boundary of the protocol's fiduciary duty to users and the reasonableness of the liquidation penalty) is becoming a frontier topic in regulatory discussion.

7.7.6 Participant identification based on order flow

In a fully transparent on-chain order book, the behavioral patterns of different participants leave clear and traceable digital footprints. Unlike the opaque environment of a CEX, the on-chain environment allows anyone to infer the current market's dominant participants and their intent by analyzing the characteristics of the order flow. This analytical capability is a distinctive advantage brought by the transparency of on-chain finance.

Through a combined analysis of the type, survival time, cancellation frequency, and fill pattern of orders, as well as their correlation with market-price movements, the following typical behavioral patterns can be identified:

- Market-maker activity signal: a dense, symmetric distribution of limit orders paired with an extremely high cancellation rate and frequent two-sided quoting generally means that market makers are actively providing liquidity and that the market is in a relatively stable state.
- Arbitrageur execution signal: a large order sweeping through multiple price levels at once, appearing simultaneously in opposite directions across two markets, is generally the signal of an arbitrageur executing a two-leg hedge strategy.
- Retail activity signal: a continuous inflow of small market orders suggests that the activity of retail speculators is rising.
- Event-driven shock: the sudden appearance of a large market order eating through several price levels is generally related to a major news event or a large on-chain liquidation.
- Stop-loss cascade signal: successive market impacts as stops are triggered reflect the coupling effect between the speculators' stop-loss chain and the liquidation engine.
- Risk-perception signal: a sudden widening of the bid-ask spread means that market makers have perceived an increase in risk and adjusted their quotes accordingly.
- Institutional execution signal: slow, tranched filling over a long period is a typical feature of a hedger or institution using an algorithm such as TWAP to execute a large order.

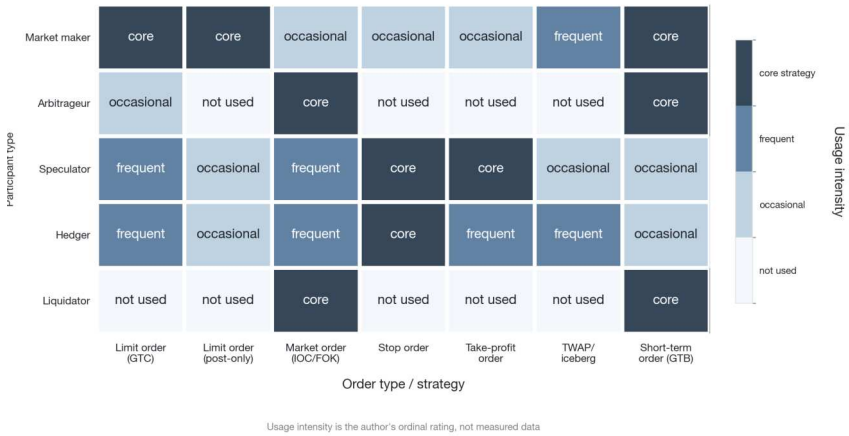


Figure 7-19. A heatmap of the order-strategy usage of different market participants (author's qualitative rating heatmap; "not used/occasionally/frequently/core" reflect the author's judgment, not empirical data; consistent with Table 7-12)

As Figure 7-19 shows: market makers' orders are highly concentrated in Post-Only limit orders and short-term orders; arbitrageurs concentrate on IOC and FOK (reflecting a need for execution atomicity); speculators' tool usage is the most dispersed (covering market orders, stop-loss orders, take-profit orders, and more, reflecting both strategic diversity and an emphasis on risk management); and liquidators concentrate on the equivalent of market orders and short-term trigger orders, reflecting their passive execution mode driven by protocol rules.

Table 7-12 systematically summarizes the order strategies of the five classes of participant along five dimensions: core objective, typical order types, reason for preference, main risks, and on-chain-specific countermeasures:

Participant	Core objective	Typical order types	Reason for preference	Main risks and costs	On-chain-specific countermeasures
Market maker	Earn the spread and maker rebates	Post-Only limit orders, layered quoting, short-term orders	Ensure maker status, build multilayered depth	Adverse selection, inventory skew, cancellation-latency cost	Widen spreads, shorten order lifespan, perpetual + spot hedging
Arbitrageur	Capture risk-free spreads	Market orders, IOC, FOK	Ensure fill certainty and two-leg consistency	Slippage, naked single leg from an unfilled order	Account for gas and MEV costs, use private transaction channels
Speculator	Directional leveraged returns	Market orders, stop-loss orders, take-profit orders, trailing stops	Quickly express a directional view and manage risk	Sandwich attacks, stop-loss cascades, slippage amplification	MEV-resistant routing, reasonable slippage protection, choose application-chain DEXs
Hedger	Reduce portfolio risk exposure	Limit orders, TWAP, VWAP, Reduce-Only	Reduce market impact and prevent operational errors	Long execution cycle, missing the optimal timing	Tranched execution, cross-market hedging, funding-rate trades
Liquidator	Obtain liquidation rewards	Liquidation-function calls (equivalent to market orders), IOC	Driven by system rules, seizing execution priority	Gas cost of losing the competition, on-chain congestion	Gas bidding, latency racing, private liquidation channels

Table 7-12. A comparison matrix of the order strategies of the five classes of market participant (Data source: constructed by the author)

By understanding "who says what with which tool," we can better understand the microstructure behind the decentralized perpetual futures market. Each class of participant's order choice is not random but a rational trade-off around cost, certainty, and risk exposure under their specific constraints. In the decentralized derivatives market, every order instruction is not only a carrier of value transfer but also a direct expression of each party's interests and strategic maneuvering. A systematic analysis of order flow not only helps traders devise more targeted counter-strategies but also provides important data support for DEX protocols to optimize their fee structures, improve their matching mechanisms, and guard against malicious attacks.

7.8 A three-dimensional framework for risk analysis

In the crypto derivatives market, the risk and cost of an order are not static attributes but dynamic variables determined by the execution environment. As discussed earlier, the slippage of a simple market order may differ by hundreds of times between a normal market and an extreme crisis, and the MEV risk it faces differs starkly between a centralized exchange and a fully on-chain decentralized exchange. To comprehensively assess the potential risks in trade execution, we need to go beyond single-dimensional consideration. This section integrates the three dimensions of order type, market state, and execution architecture into a unified analytical framework, providing traders with a practical tool for decision-making in complex market environments.

7.8.1 Defining the three dimensions

The formation of trading risk is a multivariable, interactive process. In this three-dimensional analytical framework, each dimension represents a major driver of risk, and the cross-combination of the three determines the ultimate risk level.

Dimension one: order type (microstructure role). The order type determines the role the trader plays in market microstructure: the limit order, as a liquidity provider, has as its core risks adverse selection and execution uncertainty—just as Copeland and Galai (1983) argued that "a limit order is equivalent to giving the market a free option" [4] (the full mechanism of this causality appears in Section 7.1.3); the market order, as a consumer, mainly bears market-impact cost and slippage risk; and the stop-loss market order is the most distinctive, as it may encounter a liquidity vacuum when triggered and even become a catalyst for a liquidation cascade (see Section 7.2).

Dimension two: market state (macro liquidity environment). The market state reflects the macro liquidity environment and is generally divided into three types—normal, high-volatility, and extreme-crisis: under a normal state, order-book depth is ample and spreads are tight; as the state evolves toward high volatility and then extreme crisis, liquidity becomes thin and fragile, and risk exhibits nonlinear, more-than-proportional amplification. As cited in Section 7.1.4, Kaiko Research (2025) measured that under extreme conditions order-book depth plunged by more than 90% and market-order slippage spiked from a few basis points to 5% or even 10% [10].

Dimension three: execution architecture (underlying infrastructure). The differences in execution architecture determine the transmission path and ultimate impact of risk: the CEX has extremely low latency (sub-millisecond) and zero order cost, but its opacity may conceal counterparty risk; the fully on-chain protocol has the highest transparency and censorship resistance, yet faces significant MEV risk due to high gas costs, slow block confirmation, and a public mempool; and the application-chain DEX (such as Hyperliquid and dYdX v4) seeks a balance between the two through a sovereign, customized blockchain [21] (the full comparison of the three architectures appears in Section 7.6.1).

These three dimensions together constitute a three-dimensional risk space, as shown in Figure 7-20. In this three-dimensional coordinate system, the X-axis is the order type, the Y-axis is the market state, and the Z-axis is the execution architecture, and

any order can be located by its three-axis coordinate to determine its risk level. The color gradient in the figure indicates that risk does not grow linearly as the coordinate moves, but may exhibit a more-than-proportional amplification effect when all three dimensions trend toward the extreme simultaneously. Two caveats apply. First, the "more-than-proportional" pattern rests mainly on the slippage-depth pairing observed in a few extreme events such as October 2025; it still lacks a linear-superposition baseline rigorously defined within the framework, so it should be treated as an empirical observation awaiting testing against more cross-event evidence rather than a proven law. Second, the three dimensions are not fully orthogonal—the execution architecture already endogenizes most of the risk drivers of order type and market state, so the framework more closely resembles a hierarchical structure of "architecture as the main factor, with order type and market state as moderating factors," and the three-dimensional coordinate system is merely an approximation for ease of visualization. For example, the risk level of a stop-loss market order executed through a fully on-chain DEX in an extreme crisis far exceeds the simple sum of the three single-dimensional risks.

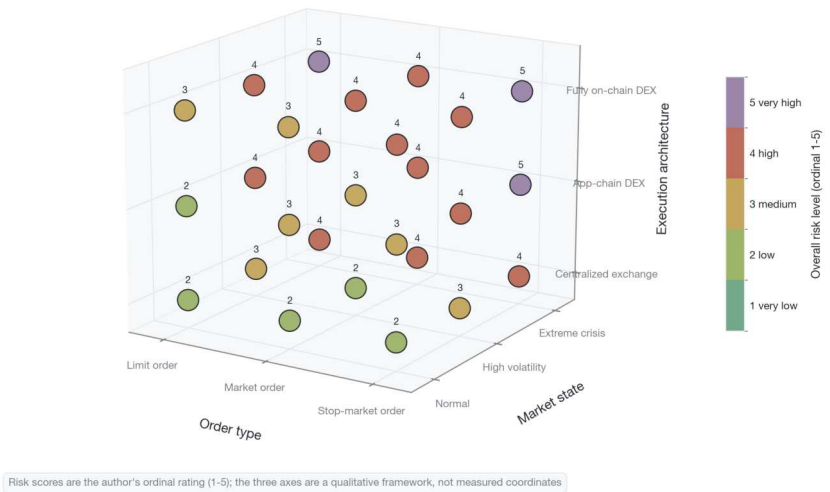


Figure 7-20. A three-dimensional analytical framework for trading risk (author's composite risk-rating 3D scatter; the ordinal 1–5 reflects the author's composite judgment, not empirical data; sources: Ali 2025 [9], Copeland & Galai 1983 [4], EigenPhi [19], dYdX Foundation [21][22])

7.8.2 Applying the three-dimensional risk matrix

To show more intuitively how these three dimensions interact and affect the ultimate execution risk, we can map the three-dimensional space onto a concrete risk matrix. Table 7-13 details the specific risk characteristics that different order types face under various market states and execution architectures.

Order type	Market state	Centralized exchange	Fully on-chain DEX	Application-chain DEX
Limit order	Normal	Low adverse-selection risk	Medium adverse-selection risk	Low-to-medium adverse-selection risk
	High volatility	High adverse-selection risk	Very high adverse selection + costly cancellation	High adverse selection + cancellation latency
	Extreme crisis	Order may fail to fill	Fails to fill + cancellation fails due to network congestion	May fail to fill
Market order	Normal	Low slippage	Medium slippage + potential MEV risk	Low-to-medium slippage, lower MEV risk
	High volatility	Medium slippage	High slippage + high MEV exposure	Medium slippage
	Extreme crisis	Very high slippage	Extreme slippage	High slippage
Stop-loss market order	Normal	Low trigger slippage	Medium trigger slippage + oracle latency	Low-to-medium + oracle latency
	High volatility	Medium trigger slippage	High trigger slippage + oracle latency	Medium + oracle latency
	Extreme crisis	High trigger slippage + may cause a cascade	Extreme slippage + risk of trigger failure	High trigger slippage + may cause a cascade

Table 7-13. The three-dimensional risk matrix from the user's perspective (Note: order-of-magnitude reference for the qualitative risk levels in the table: "low slippage": <10 bps; "medium slippage": 10–50 bps; "high slippage": 50–200 bps; "very high/extreme slippage": >200 bps. "Low adverse-selection risk": market-maker spread <5 bps; "high adverse-selection risk": spread widening to >20 bps. These values are the typical order of magnitude for mainstream trading pairs under the corresponding market state; the specific values vary by underlying, time period, and platform.) (Data source: compiled by the author)

Analyzing this matrix, we can identify several key regularities.

Under a normal market state, the CEX has a clear advantage in execution efficiency and cost, especially for market orders and stop-loss orders, whose low slippage and instant triggering are difficult for on-chain protocols to match. On a fully on-chain DEX, however, even in the normal state, a market order must bear medium slippage and potential MEV risk, and a limit order must pay a gas fee when canceled. The application-chain DEX performs close to the CEX in the normal state, but oracle latency remains an inherent constraint on conditional-order triggering.

When the market enters a high-volatility state, the risk exposure of on-chain protocols increases sharply. High cancellation costs and network congestion make it difficult for market makers to adjust their quotes quickly, resulting in very high adverse-selection

risk. At the same time, MEV searchers are more active in this environment, and market orders face a very high probability of being sandwich-attacked. The application-chain DEX shows a degree of resilience in this state: its lower cancellation cost and faster confirmation speed mitigate these risks to some extent, but it still cannot entirely eliminate the stop-loss-triggering lag caused by oracle latency.

Most noteworthy is the performance under an extreme-crisis state. At this point, all execution architectures face challenges, but the form the risk takes differs. On a CEX, the slippage of a market order may reach a very high level, and the mass triggering of stop-loss orders may become a catalyst for a liquidation cascade—precisely the concrete manifestation of the procyclicality of cross-margin in the October 2025 liquidation cascade described by Ali (2025) [9]. On a fully on-chain DEX, the situation can become extremely severe: a market order may face enormous slippage, and a stop-loss order, owing to oracle-update latency and network congestion, may even face the risk of trigger failure. This nonlinear widening of the "intent-execution gap" during a crisis is a systemic risk that traders must fully recognize. Especially worth guarding against is the correlated failure mode in which all protective layers fail simultaneously: in an extreme crisis, a stop-loss order fails to fill at an acceptable price because of a liquidity vacuum, the liquidation engine is delayed because its processing queue is overloaded, and the oracle price feed lags the actual market price because of latency in the underlying data source itself. When these three lines of defense fail at once, a trader's position can persist at a level far beyond the liquidation price, producing bad debt that exceeds the margin. At this point the insurance fund is called upon to absorb the bad debt; and if the insurance fund is exhausted, the protocol generally activates the auto-deleveraging (ADL) mechanism, forcibly closing the corresponding positions of the profitable side to digest the loss. This comprehensive failure of defense-in-depth is rare, but it is precisely the extreme tail scenario that systemic-risk assessment must consider.

7.8.3 Practical implications for risk management

The three-dimensional risk framework is not merely a theoretical model; it can also provide concrete guidance for a trader's actual operations. Based on the analysis above, we can draw the following practical implications for risk management.

There is no universally "optimal" execution architecture, only the architecture "best suited" to a specific scenario. Traders should choose an execution platform according to their own strategic needs, order size, and tolerance for different types of risk. Under a normal market environment, a high-frequency trader pursuing extremely low slippage may lean toward a CEX; whereas for a trader who highly values self-custody of funds and censorship resistance, a fully on-chain protocol is the preferred choice. The application-chain DEX offers a highly attractive option for those traders who wish to find a compromise among speed, cost, and decentralization. This judgment echoes the core argument of Chapter 5 regarding the "impossible trinity": every architecture choice is a specific trade-off among performance, decentralization, and security.

One must recognize the nonlinear character of risk growth. As discussed in Section 7.8.1, from a normal market to high volatility and then to extreme crisis, slippage and adverse-selection risk amplify nonlinearly and more than proportionally, rather

than increasing proportionally. When devising a strategy, therefore, a trader cannot backtest based only on normal-market parameters but must incorporate the nonlinear risk under extreme conditions. As discussed in Chapter 1 with the concept of "risk-surface transformation," high-frequency variable shocks are the theoretical basis for the nonlinear amplification of risk.

At the operational level, the following principles deserve special attention. For large trades, whatever platform they are executed on, it is strongly recommended to use algorithmic orders such as time-weighted average price or volume-weighted average price for splitting, so as to effectively reduce market impact and signal leakage. Stop-loss orders must be used with extreme care; never "set and forget." When using stop-loss orders in an on-chain environment, one must fully understand the inherent lag brought by oracle latency and block confirmation. Traders should treat MEV loss as an implicit trading cost and include it when assessing a strategy's profit and loss. By diversifying asset allocation across the CEX and on-chain protocols, one can simultaneously hedge the trust risk of centralized platforms and the technical risk of decentralized protocols, achieving more robust overall risk management. However, the diversification strategy itself also introduces new risks: cross-platform rebalancing operations carry latency (especially during a crisis, when cross-chain transfers and exchange withdrawals may lag severely), capital efficiency declines because funds are dispersed, and in extreme conditions multiple platforms may experience liquidity depletion or technical failure simultaneously, weakening the expected protective effect of diversification.

The three-dimensional framework above analyzes the static risk characteristics of a single order in a specific environment. In an actual market, however, risk evolves dynamically: thousands upon thousands of orders pour into the market as a continuous flow, and the orders of different participants intertwine, superimpose, and hedge one another. To understand this dynamic process, one must rise from the static analysis of a single order to the dynamic examination of order flow.

7.9 Order-flow analysis

If a single order is a "word" in the language of trading, then a continuous order flow is a sequence composed of these orders. The information content of order flow and its compositional structure fundamentally determine a market's price-discovery efficiency, the quality of its liquidity, and the distribution of profit among different participants. This section transitions from the static analysis of a single order to the dynamic examination of order flow, establishing a conceptual bridge to understanding the market's information ecosystem and laying the groundwork for the next chapter's discussion of the order book that order flow aggregates into.

7.9.1 The composition of order flow

The order flow in a market is not a homogeneous whole but a composite structure formed by the confluence of different sources. Each source represents the trading intent of a specific participant, carries a different type of information, and produces a distinctive impact on market microstructure. We can classify these order flows into

four main types, as shown in Figure 7-21.

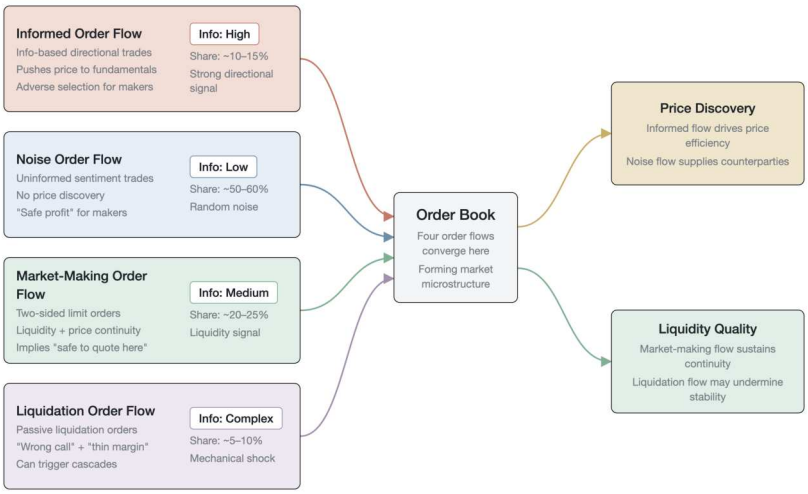


Figure 7-21. The four sources of order flow and their information content (illustration of the information content of the four classes of order flow: the shares of about 10–15% / 50–60% / 20–25% / 5–10% are the author's order-of-magnitude estimates, not empirical, and are affected by the macro cycle; per Easley et al. 2012 VPIN [29], Copeland & Galai 1983 [4], Ali 2025 [9])

Figure 7-21 divides order flow into four classes—informed, noise, market-making, and liquidation—in descending order of information content; the characteristics of each are analyzed in turn below.

Informed order flow is initiated by traders who possess information not yet fully priced into the market. These traders make directional bets based on their informational advantage, and their order flow has extremely high information content. The PIN/VPIN literature of market microstructure (Easley et al.) estimates that in a typical mature market, informed order flow accounts for about 10% to 15% of total order flow, but its influence on price far exceeds its share by count [29]. Informed order flow is the core force driving price toward fundamental value, promoting the market's price-discovery efficiency. For the market makers who provide liquidity, however, informed order flow is the direct source of adverse-selection risk (the mechanism by which market makers accordingly widen the bid-ask spread to compensate for potential losses appears in Section 7.1.3).

Noise order flow comes from participants who lack an informational advantage and trade on the basis of emotion, liquidity needs, or mistaken beliefs. This class of order flow has extremely low information content, generally manifesting as random buying and selling that does not exert a long-term directional push on price. Yet noise order flow plays an important role in the market ecosystem: it provides market makers with a relatively safe counterparty and is the primary source of their spread profit, thereby indirectly supporting the market's overall liquidity. In the crypto derivatives market, because of the heavy participation of retail speculators, the share of noise order flow is generally higher than in traditional equity markets; the author's preliminary order-of-magnitude estimate, based on public data, is about 50% to 60% of total order flow.

This ratio, however, carries a high degree of methodological uncertainty and depends heavily on the macro cycle: in a bull market, retail speculative activity surges and the noise-flow share may rise significantly; in a bear market, retail participation ebbs and the share contracts sharply. In addition, the original study of the PIN model by Easley et al. (2012) [29] targeted the U.S. equity market, and directly transplanting its informed-flow ratio to the crypto perpetual futures market requires caution, because the two markets differ fundamentally in participant composition, information-propagation speed, and regulatory environment. Although the precise figure is hard to determine, the higher share of noise flow remains an important reason for the relatively high market-making profit margin in this market.

Market-making order flow consists mainly of the two-sided limit orders submitted by market makers, estimated to account for about 20% to 25% of total order flow in mainstream trading pairs (based on the publicly disclosed market-share data of market makers on top exchanges). This class of order flow does not itself contain strong directional information; its core function is to provide liquidity and maintain price continuity. If market-making order flow conveys any information, it is that "market makers regard providing liquidity at the current price level as relatively safe." When market makers perceive an increase in market risk (for example, anticipating an imminent surge of informed order flow), they quickly withdraw or adjust their quotes, and the sudden widening of the bid-ask spread and the sharp drop in order-book depth themselves constitute a strong signal of a change in market state.

Liquidation order flow is a class distinctive to, and extremely important in, the crypto derivatives market. It consists of passive orders forcibly triggered by the protocol's risk engine when a trader's margin is insufficient, generally entering the market in the form of market orders. The information content of liquidation order flow is quite complex: it may merely reflect a single trader's poor leverage management (a purely mechanical trigger, with no directional information), or it may hint at a reversal of the market trend that has caught large numbers of traders on the wrong side. A liquidation cascade, moreover, is generally not confined to a single market: the forced selling pressure produced by a large-scale liquidation of BTC perpetual futures spills over to related assets such as ETH and SOL, triggering a cross-asset "correlation spike" and in turn setting off liquidation cascades in those assets themselves, forming a cross-market systemic contagion. More critically, large-scale liquidation order flow produces a strong mechanical price impact and may even trigger a chain of liquidations (a liquidation cascade), undermining the stability of market liquidity [9]. During the normal operation of a typical perpetual futures market, liquidation order flow is estimated to account for about 5% to 10% of total order flow (extrapolated from the frequency of on-chain-observable liquidation events), but in extreme conditions this ratio may spike sharply, becoming the core force dominating the market's short-term movement [9].

7.9.2 Order-flow toxicity and VPIN

Having understood the composition of order flow, the core challenge facing market makers and quantitative traders is how to assess in real time the information content of incoming order flow—the so-called "order-flow toxicity." Highly toxic order flow means

that informed traders are actively entering, and if market makers do not adjust their strategy in time, they will face serious adverse-selection losses.

> VPIN is a classic indicator for measuring order-flow toxicity, proposed by Easley, Lopez de Prado, and O'Hara (2012). Its core intuition is that when buyer-initiated volume and seller-initiated volume exhibit a significant and sustained imbalance over a short period, it very likely means that informed traders are actively building positions using their informational advantage [29].

VPIN is computed as follows: total volume is divided equally into n volume buckets, each containing a fixed volume V. For the i-th bucket, a bulk-classification algorithm divides the volume into buyer-initiated volume V_i^B and seller-initiated volume V_i^S. VPIN is defined as:

$$\text{VPIN} = \frac{\sum_{i=1}^{n} |V_i^B - V_i^S|}{n \cdot V}$$

where $n \cdot V$ is the total volume within the window, so VPIN falls in the interval $[0, 1]$. This volume-based rather than time-based sampling method can capture microstructural anomalies that are hard to detect with traditional time-period-based data. When VPIN approaches 1, it indicates that the order flow is almost entirely one-directional and toxicity is extremely high; when it approaches 0, it indicates that buying and selling are roughly balanced and toxicity is low. In practice, a sharp rise in VPIN often precedes large market swings, so it is widely used as an early-warning indicator.

However, one must be especially cautious when applying VPIN or similar indicators in the perpetual futures market. This is because order-flow imbalance in the perpetual futures market may arise not only from informed traders' directional bets but also from a liquidation cascade. When large numbers of long positions are forcibly closed, an enormous volume of sell market orders suddenly emerges, which appears in the data as a significant seller imbalance—highly similar in feature to "informed traders shorting heavily." In reality, however, these liquidation orders are "forced to sell" and contain no predictive information about future prices; they are mechanical, not informational.

Accurately distinguishing information-based toxic flow from mechanism-based liquidation flow is key to high-frequency trading strategies and market-making strategies in perpetual futures. Research by Bieganowski and Ślepaczuk (2026) further shows that by analyzing the microstructural features of order-flow imbalance (such as the duration of the impact and the accompanying pattern of OI changes), one can distinguish these two different kinds of order flow to some extent [7]. Later chapters further explore how market makers quantify and manage the information content of order flow (see the "toxic-flow spectrum" framework in Chapter 19); this section only establishes the basic recognition.

7.9.3 The distinctive signals of perpetual futures order flow

Compared with the traditional spot market, the order flow of perpetual futures carries some distinctive signals derived from its derivative mechanisms. These signals offer the keen observer valuable clues for predicting short-term price dynamics.

Because perpetual futures allow high-leverage trading, traders' liquidation prices tend to cluster in specific price ranges, forming so-called liquidation-price clusters. In an on-chain environment, these liquidation prices are even publicly visible. When the liquidation prices of large numbers of positions cluster near the current price, a potential zone of concentrated risk forms, as shown in Figure 7-22.

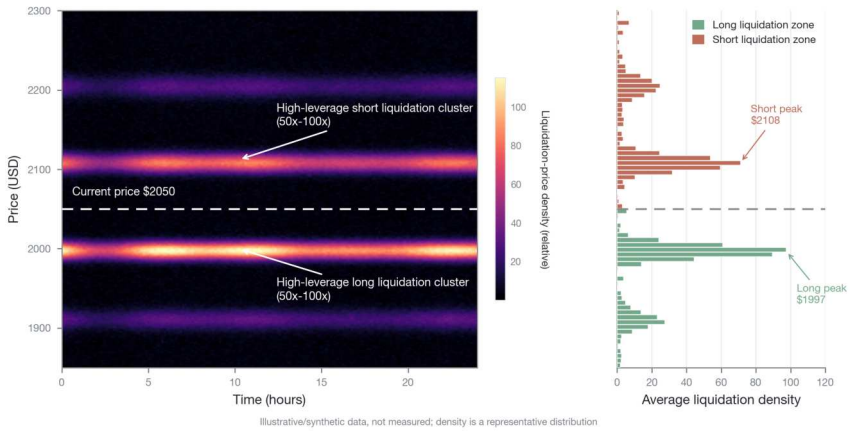


Figure 7-22. A heatmap of liquidation-price clusters (conceptual illustration: $2,050/$1,997/$2,108 and the like are synthetic representative values, not empirical; the distribution pattern is per Ali 2025 [9] and public statistics from CoinGlass/Bybit/Binance)

A liquidation-price cluster is a signal with a strong self-reinforcing property. It signals to the market: "if the price reaches this level, it will trigger a large number of forced market orders." When enough participants observe this signal, speculative capital may strategically push the price toward the cluster (the so-called deliberate triggering of stop-losses or deliberate triggering of liquidations), in the hope of setting off a liquidation cascade and profiting from it. The triggering of a liquidation cascade, however, does not depend solely on deliberate manipulation by large players: as the price approaches a publicly visible liquidation cluster, panic-driven early exits (traders proactively selling to avoid being liquidated) accelerate the price's move toward the liquidation zone, forming a self-fulfilling prophecy. The real-time spread of social media further amplifies this effect—the wide dissemination of liquidation heatmaps makes the liquidation zone common knowledge, accelerating herd behavior and increasingly blurring the boundary between deliberate hunting and the spontaneous cascade caused by collective panic. Methodological clues for distinguishing a deliberately manipulated trigger from a natural swing touching a liquidation cluster include: anomalous order-flow patterns (whether large one-directional orders inconsistent with normal trading activity appear before the trigger), the price-recovery shape (deliberate manipulation is generally accompanied by a rapid V-shaped reversal after the trigger, whereas a trend-driven trigger shows a sustained decline), and the pattern of change in open interest (a deliberate manipulator has generally established a reverse position before the trigger, so the change in OI exhibits a special temporal relationship with the price direction). This fragility, brought by the transparency of the mechanism design, significantly affects the short-term microstructure of the perpetual futures market and is a concrete manifestation, at the order-flow level, of the "transparency-fragility paradox" discussed

in Chapter 4.

The anticipatory trading behavior around funding-rate settlement (whose mechanisms of anchoring to spot and settling every 8 hours appear in Section 7.4.3 and Chapter 10) constitutes another class of distinctive signal. Around the moment of funding-rate settlement, order flow often exhibits an observable systematic pattern, reflecting traders' expectations of the direction and magnitude of the rate. For example, if large amounts of short-position-building order flow suddenly appear just before a funding-rate settlement, this may mean that the market broadly expects the next period's funding rate to be positive (longs pay shorts), so speculators are "positioning in advance" to collect the rate. By analyzing the order-flow dynamics before and after settlement, one can infer the trade-off strategies market participants are making between directional gains and rate gains.

The direction of order flow, paired with changes in open interest, can provide a richer signal than a bare price rise or fall—this is precisely the informational value embedded in changes in open interest. An increase in open interest means new capital is entering the market to establish positions; a decrease means existing positions are being closed. Cross-analyzing the price direction with the change in open interest can distinguish four combination patterns with different market implications.

Price direction	OI change	Market implication	Signal strength
Up	Increase	New long capital actively building positions	Strongly bullish
Up	Decrease	Short covering (shorts closing) driving the rise	Weakly bullish
Down	Increase	New short capital actively building positions	Strongly bearish
Down	Decrease	Longs closing (long stop-losses) driving the decline	Weakly bearish

Table 7-14. Interpreting the signals of price direction and changes in open interest (Data source: compiled by the author)

The most noteworthy contrast in Table 7-14 is the difference between "up + OI increase" and "up + OI decrease": the former indicates that new long capital is actively entering, constituting a strongly bullish signal; the latter merely reflects short covering, with weak upward momentum. Likewise, "down + OI increase" indicates that new short capital is entering and is a strongly bearish signal, whereas "down + OI decrease" merely reflects long stop-losses, with limited downward pressure. This distinction has practical value for judging the persistence of a price movement.

In the 24/7, high-leverage perpetual futures market, this interaction between OI and order flow changes more rapidly and violently. Especially during a liquidation cascade, a sharp drop in OI accompanying a one-directional price plunge is a typical feature of "mechanical volatility" (rather than "informational volatility") (see the detailed analysis

in Chapter 22).

7.9.4 From orders to the order book

In this chapter, we have deconstructed the "words" of the trading language—the order—in detail. We have explored the dual function of an order (trading and signaling), the strategic implications of different order types, the three-dimensional extension that perpetual futures confer on orders, and how the underlying execution architecture reshapes the physical properties of orders. We have seen that even the simplest market order or limit order conceals complex maneuvering and trade-offs behind it.

A market, however, is not composed of isolated orders. As this section has shown, thousands upon thousands of orders carrying different intents from different participants converge to form a continuous order flow. These order flows interact, match, or rest within the matching engine, ultimately aggregating into a real-time-updated data structure—the order book.

The order book is "the real-time record of the market's conversation." Its state at every microsecond is the cumulative result of all prior order submissions, cancellations, fills, and liquidations. The shape, depth, and buy-sell asymmetry of the order book, and its resilience when struck by new orders, embed rich information about the market's current liquidity conditions and future price direction. The next chapter raises the line of sight from the level of "words" and "sentences" to that of "discourse," dissecting the structural dynamics of the order book and establishing a complete analytical framework for understanding the evolution of market microstructure.

7.10 Chapter summary

The order is the most basic building block of market microstructure and the only interface through which a trader interacts with the market. In the highly complex, round-the-clock, and strategically charged environment of the crypto derivatives market, understanding the meaning of an order is a precondition for every participant seeking to build an effective strategy. This chapter has systematically deconstructed the concept of the "order," establishing for the reader—from several core perspectives—an analytical framework for mastering this fundamental toolkit of trading.

The foundation of understanding microstructure lies in recognizing that the order is a signal. While every order performs its surface "trading function," it inevitably also exercises a "signaling function," leaking to the market clues about the submitter's urgency, capital base, and information content. The choice of order type is, in essence, the process by which a trader finds the optimal trade-off position between "execution certainty" and "information leakage"; and it is precisely by reading these signals that market makers manage adverse-selection risk, which constitutes the microfoundation of bid-ask spread formation [4].

The emission of a signal, however, is only the starting point. Between a trader's original intent and the final fill outcome lies an intent-execution gap composed jointly of slippage, latency, MEV, information leakage, and liquidation risk. The width and composition of this gap differ fundamentally across execution architectures: on a centralized exchange,

the gap is generally narrower but contains a hard-to-quantify implicit risk of information exploitation; on an on-chain protocol, the gap may be wider, but every component of it is transparent and auditable. Especially worth guarding against is that, under an extreme-crisis state, this gap undergoes a sharp nonlinear widening [10]—a tail event that risk management must guard against.

Building on this, the distinctive mechanisms of perpetual futures further intensify the complexity of order decisions. Order decisions in the traditional spot market revolve mainly around price, quantity, and time, whereas perpetual futures add three important dimensions: leverage (every order changes the effective leverage and liquidation distance), liquidation (the complex interaction between the stop-loss order and the forced-closure mechanism), and time (the funding-rate settlement cycle creates new strategic timing). This multidimensional complexity is the microfoundation of the perpetual futures market being "simple to operate but complex to decide," and it is an important reason for the systematic losses of many ineffective traders [9].

At the same time, when these orders are submitted to an on-chain environment, their nature undergoes a fundamental transformation: on a centralized exchange, an order is a private expression of intent that can be canceled at any time at zero cost, but on an on-chain central limit order book, every order becomes a "public commitment" observable by the entire world. This transparency is a mechanism with both positive and negative effects: on the one hand it provides verifiable, trustless liquidity; on the other, it fully exposes the trader's strategic intent, making key levels such as the stop-loss order easy targets for sniping [20].

More fundamentally, the physical constraints of blockchain reshape the economic semantics of orders from the ground up: state-storage cost forces permanently valid orders to be replaced by orders with a lifecycle; the consensus mechanism turns cancellation from free and instant into costly and latent; oracle dependence introduces an inherent lag into the triggering of conditional orders; and discrete block time transforms continuous matching into a batch auction. Understanding how these underlying architectures change the behavioral characteristics of orders is the key entry point for understanding the distinctiveness of on-chain finance.

It is thus clear that an order is by no means merely a button on a software interface or a string of code in an API, but a complex carrier of intent, risk, and information. This chapter has completed a systematic dissection of the "words" of the trading language; the next chapter raises the perspective to the level of "sentences" and "narrative," analyzing the core data structure that order flow aggregates into—the order book—and revealing its structural dynamics, its information content, and how it drives the evolution of the entire microstructure ecosystem as the real-time record of the market's conversation.

References

[1] CoinGlass. (2025). 2025 annual crypto derivatives market report. CoinGlass. https://www.coinglass.com/learn/2025-annual-report-en

[2] CoinGecko. (2025). 2025 annual crypto industry report. CoinGecko Research.

https://www.coingecko.com/research/publications/2025-annual-crypto-report

[3] Harris, L. (2003). *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press.

[4] Copeland, T. E., & Galai, D. (1983). Information effects on the bid-ask spread. *The Journal of Finance*, 38(5), 1457–1469. https://doi.org/10.2307/2327580

[5] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1), 71–100. https://doi.org/10.1016/0304-405x(85)90044-3

[6] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335. https://doi.org/10.2307/1913210

[7] Bieganowski, B., & Ślepaczuk, R. (2026). Explainable patterns in cryptocurrency microstructure. *arXiv preprint*, arXiv:2602.00776. https://arxiv.org/abs/2602.00776 https://doi.org/10.48550/arXiv.2602.00776

[8] Ruan, Q., & Streltsov, A. (2022). Perpetual futures contracts and cryptocurrency market quality. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4218907 https://doi.org/10.2139/ssrn.4218907

[9] Ali, Z. (2025). Anatomy of the Oct 10-11, 2025 crypto liquidation cascade: Macroeconomic triggers, market microstructure, and systemic risk lessons. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5611392 https://doi.org/10.2139/ssrn.5611392

[10] Kaiko Research. (2025, October). When October surprise meets crypto liquidity drought. *Kaiko Research*. https://research.kaiko.com/insights/when-october-surprise-meets-crypto-liquidity-drought

[11] Perold, A. F. (1988). The implementation shortfall: Paper versus reality. *The Journal of Portfolio Management*, 14(3), 4–9.

[12] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, 22(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[13] Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32(2), 311–328. https://doi.org/10.1037/0022-3514.32.2.311

[14] Hyperliquid. (n.d.). Overview — The Hyperliquid L1. Hyperliquid Documentation. https://hyperliquid.gitbook.io/hyperliquid-docs

[15] Talos. (2025). VWAP or TWAP for crypto execution? A market impact perspective. Talos Insights. https://www.talos.com/insights/vwap-or-twap-for-crypto-execution-a-market-impact-perspective

[16] Jasiak, J., & Zhong, A. (2024). Intraday and daily dynamics of cryptocurrency. *International Review of Economics & Finance*, 96(Part B), 103658. https://doi.org/10.1016/j.iref.2024.103658 https://www.sciencedirect.com/science/article/pii/S1059056024006506

[17] Genet, R. (2025). Deep learning for VWAP execution in crypto markets: Beyond the volume curve. *arXiv preprint*, arXiv:2502.13722. https://arxiv.org/abs/2502.13722

https://doi.org/10.48550/arXiv.2502.13722

[18] Adrian, T., & Shin, H. S. (2010). Liquidity and leverage. *Journal of Financial Intermediation*, 19(3), 418–437. https://doi.org/10.1016/j.jfi.2008.12.002

[19] EigenPhi. (2025). Exclusive data reveals that sandwich attacks on Ethereum have waned. *Cointelegraph Research*. https://www.tradingview.com/news/cointelegraph:fa12ba092094b:0-exclusive-data-from-eigenphi-reveals-that-sandwich-attacks-on-ethereum-have-waned/

[20] Flashbots. (2024). Publishing Flashbots Protect and MEV-share data. Flashbots Collective. https://collective.flashbots.net/t/publishing-flashbots-protect-and-mev-share-data/3087

[21] dYdX Foundation. (n.d.). Intro to dYdX Chain architecture. dYdX Documentation. https://docs.dydx.exchange

[22] dYdX Foundation. (n.d.). Short-term vs stateful orders. dYdX API Integration Trading Documentation. https://docs.dydx.exchange/api_integration-trading/short_term_vs_stateful

[23] Chainlink. (n.d.). Data feeds — Using data feeds. Chainlink Documentation. https://docs.chain.link/data-feeds

[24] Commodity Futures Trading Commission. (2023, January 9). CFTC charges Avraham Eisenberg with manipulative and deceptive scheme to misappropriate over $110 million from the Mango Markets exchange [Press Release No. 8647-23]. https://www.cftc.gov/PressRoom/PressReleases/8647-23

[25] Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, 130(4), 1547–1621. https://doi.org/10.1093/qje/qjv027

[26] Injective Labs. (n.d.). Injective exchange upgrade: A novel order matching mechanism. Injective Blog. https://injective.com/blog/injective-exchange-upgrade-a-novel-order-matching-mechanism

[27] Paradex. (n.d.). Contract specifications — Maximum number of open orders. Paradex Documentation. https://docs.paradex.trade

[28] Usman, A. (2025). 19 security pitfalls in on-chain order books. *Hacken Insights*. https://hacken.io/insights/order-book-security-vulnerabilities/

[29] Easley, D., Lopez de Prado, M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *The Review of Financial Studies*, 25(5), 1457–1493. https://doi.org/10.1093/rfs/hhs053

The Perpetual Futures Order Book and Market Microstructure

On October 10, 2025, the cryptocurrency market experienced a flash crash of unprecedented scale. Every extreme market event unfolds along a three-stage chain of preconditions, triggers, and amplification mechanisms: a contraction in global risk appetite and a turn in the monetary policy cycle constituted the preconditions for systemic fragility; specific selling pressure served as the trigger; and defects in the market microstructure supplied the amplification mechanism. According to post-mortem reports from FTI Consulting and Amberdata, more than $9.89 billion in leveraged positions was forcibly liquidated within just 14 hours. This forced-liquidation volume represented roughly 6% to 7% of the total open interest in perpetual futures across the market at the time (approximately $9.89 billion against roughly $146.7 billion in open interest) and exceeded the roughly $8.6 billion liquidated during the May 2021 crash, with about 70% of the damage concentrated within 40 minutes [1] (a clarification is warranted: the $9.89 billion figure refers to the core window of roughly 14 hours under the perpetual-futures forced-liquidation measure; if one measures the evaporation of market-wide leverage and open interest, and includes voluntary position closures and position transfers, the entire deleveraging round reached roughly $19 billion over about a day. The two figures differ because of differences in statistical measure and time window [2]). Yet the fundamental reason this crisis escalated from a routine price correction into a systemic collapse lay not in any single price shock but in the underlying architecture of the market microstructure. Comparing three snapshots of this perpetual futures order book (such as BTC-USDT) — 24 hours before the crisis, at the moment it erupted, and 24 hours afterward — systematically reveals the evolution of the liquidity illusion and of systemic fragility.

Comparing three snapshots of the BTC-USDT perpetual futures order book before and after the crisis makes this process clearly observable. In the 24 hours before the crisis, the order book displayed a textbook equilibrium: depth on both the bid and ask sides was ample and approximately symmetric, the best bid-ask spread was as low as 0.02 basis points, and visible liquidity (defined as the cumulative two-sided resting depth within $\pm 2\%$ of the best bid and ask) reached as high as $103.6 million [1]. Even so, bid-side depth already showed a pronounced "depth cliff" near the dense liquidation zone below the prevailing price, as market makers, foreseeing that this region would generate a toxic flow of liquidation orders, had withdrawn in advance. At 20:50 UTC on October 10, after sustained selling pressure had driven the price to the liquidation threshold, the

order book underwent a violent structural change within minutes. At the peak moment of 21:15 UTC, $3.21 billion in positions was forcibly liquidated within 60 seconds (93.5% of it forced selling), visible depth shrank from $103.6 million to just $170,000 (a decline of about 99.8%), and the spread widened from 0.02 basis points to 26.43 basis points (both figures are instantaneous snapshot values rather than time-weighted averages, and the 26.43 basis points was the peak spread recorded at 21:15 UTC) — a widening of roughly 1,321-fold. The market sell orders generated by the liquidation engine and the collective retreat of market makers formed a positive feedback loop of "price decline → liquidations triggered → depth consumed → further price decline" [1]. The post-crisis recovery, by contrast, exhibited pronounced asymmetry: the collapse took less than two hours, whereas restoring liquidity required weeks. A CoinDesk report one month after the crisis showed that order book depth remained far below early-October levels and that market makers had undergone a structural retreat [3].

These three snapshots crystallize the central question of this chapter. Chapter 7 analyzed the single order — the basic vocabulary of the language of trading. The object of this chapter is the order book: the dynamically evolving market microstructure formed by the continuous interaction of millions of orders. The order book does more than perform matching; it is a continuously operating information-aggregation device. The high leverage, liquidation mechanisms, funding rates, and around-the-clock trading environment of perpetual futures leave a distinctive set of "four fingerprints" on the order book. When executing a large perpetual futures long position, the choice between a centralized exchange's continuous matching system and an application-chain-based batch-auction mechanism directly determines whether the microstructure can offer genuine protection under extreme market conditions. Analyzing the structure of the order book is the foundation for understanding this choice.

The chapter proceeds as follows. It first defines the order book as an information machine and proposes a joint diagnostic framework of five signal channels, then traces the dual causes of the depth illusion and the methods for quantifying it. A survey of the five institutional mechanisms unique to perpetual futures shows how each reshapes order book behavior. Turning to topology, the analysis distills a morphological classification of the order book and its mapping to market states, and on that basis proposes the "four fingerprints" framework, which translates those institutional mechanisms into observable, quantifiable diagnostic tools. Introducing the time dimension, it then examines the structural disadvantage in order book resilience. The discussion moves next to the infrastructure layer, tracing the evolution of matching-engine architecture and the institutional innovation of matching mechanisms, and closes by contrasting the information ecosystems of on-chain and off-chain order books.

8.1 The order book as an information machine

Traditional finance textbooks typically define the function of the order book simply as "matching buyers and sellers." This is correct but incomplete. Matching is only the mechanical function of the order book; its deeper and more essential function is information aggregation.

In his classic work *Empirical Market Microstructure*, Joel Hasbrouck notes that the order book is the most fundamental institutional arrangement in market microstructure and that its design directly determines the efficiency of price discovery, the quality of liquidity, and the fairness of trading [4]. The order book is an information machine: through continuous price competition, it transforms the private information, expectations, and risk preferences dispersed across many participants into a set of publicly observable state variables. These variables carry rich signals about the current state of the market and its near-term future.

In perpetual futures markets, this information aggregation is pushed to an extreme. Because these contracts have no expiry date and are anchored to the spot price through the funding rate, the order book becomes the central arena in which longs and shorts engage in high-frequency competition over the funding rate, liquidation risk, and macroeconomic expectations. Every order submitted, modified, or canceled broadcasts information: a market maker's cancellation may hint that liquidity is about to dry up, while an abnormal buildup of depth on one side may signal that an institution is establishing a large one-directional position. Treating the order book as an information machine is precisely the analytical starting point for understanding perpetual futures microstructure.

8.1.1 The structure and rules of the CLOB

The central limit order book (CLOB) is the core infrastructure of modern electronic trading markets. Whether in traditional stock exchanges or in emerging cryptocurrency derivatives platforms, the CLOB holds a dominant position. Understanding how this information machine operates requires first dissecting its physical structure.

Structurally, the order book consists of two sides: the bid side and the ask side. Each side is strictly sorted by price: bid orders are arranged from high price to low, and ask orders from low price to high. This ordering reflects the core objectives of market participants: buyers wish to buy at the lowest possible price, while sellers wish to sell at the highest possible price.

The gap between the best bid and the best ask is called the bid-ask spread. The spread is the price of liquidity — the minimum cost a trader must pay to obtain immediate execution. In markets with ample liquidity, the spread is typically very narrow, whereas when liquidity dries up or the market becomes highly volatile, it widens rapidly. The size of the spread reflects not only the current state of liquidity but also market makers' expectations of near-term volatility and their risk premium.

The quantity of resting orders accumulated at each price level constitutes the depth at that level. Depth reflects the market's capacity to absorb the impact of large orders without a dramatic price move. As Larry Harris defines it in *Trading and Exchanges*, the order book is "a dynamic electronic record of all unexecuted limit orders submitted by market participants, displaying the state of market liquidity in real time" [5]. Depth is not merely a static indicator; it also evolves dynamically as participants' expectations change. The distribution of depth is often nonuniform, typically densest near the best bid and ask and thinning progressively as prices move away from the best quotes.

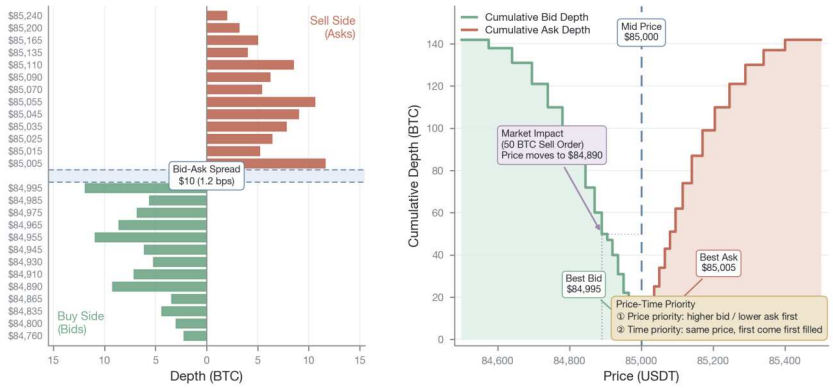


Figure 8-1. Basic structure of the central limit order book and the market impact effect (conceptual structural illustration: the $85,000 mid price, the $10 spread of about 1.2 basis points, and the impact of a 50 BTC sell order down to $84,890 are all illustrative synthetic values, not empirical data; the central limit order book and the price-time priority rule follow Harris 2003 [5])

Order book depth follows an inverted-cone, nonuniform distribution — thick in the middle and thin at both ends. The key insight lies in the cumulative depth curve in the right panel: when order size is small, the slope of the curve is gentle and price impact is limited, but once size exceeds the capacity of the near-touch liquidity, the curve turns sharply upward, and the price slippage induced by each additional unit of size is significantly amplified. This nonlinearity is precisely the microstructural basis for understanding the execution cost of large trades and extreme market events.

With respect to matching priority rules, the vast majority of CLOBs follow the principle of price-time priority, which is key to understanding the microdynamics of the order book. Price priority means that a higher bid receives higher priority among buyers, and a lower ask receives higher priority among sellers. This principle ensures that the market always matches in the direction most favorable to traders. It is the cornerstone of the price discovery mechanism, encouraging participants to offer more competitive quotes and thereby narrowing spreads and improving market efficiency. When prices are equal, time priority applies: the order that reaches the system first receives priority for execution, which is intended to ensure fairness of processing. In the modern high-speed electronic environment, however, time priority has spawned intense speed races and latency arbitrage. High-frequency trading firms invest heavily in optimizing network connectivity and the response speed of matching engines, all to arrive at a given price level a few microseconds ahead of their competitors. This is explored in depth in the later section on the evolution of matching engines. In perpetual futures markets, the widespread use of APIs further intensifies this time-based competition.

With the priority rules in place, consider how a market order consumes the book level by level. When a newly arriving market buy order enters the system, it consumes resting orders starting from the lowest ask, level by level. If the size of the market order exceeds the available depth at the first level, it continues to consume the resting orders at the second level, the third level, and so on, until the order is fully filled or the order book depth is exhausted. The price movement produced by this process is called

market impact.

The magnitude of market impact depends on the ratio of order size to available depth. Kyle's (1985) seminal model predicts a linear price impact function, $P = P_0 + \lambda x$, in which the parameter λ characterizes market depth (depth equals $1/\lambda$; that is, the larger λ is, the shallower and less deep the market, the two being inversely related), so that an order of any size bears a price impact proportional to its size [6]. Subsequent empirical research, however, has systematically revealed the limitations of this linear prediction. Almgren and Chriss (2001) introduced a nonlinear impact function within an optimal-execution framework [7], and Bouchaud, Farmer, and Lillo (2009), analyzing large-scale empirical data, found that the price impact of large orders exhibits pronounced concavity — the so-called square-root law, whereby the magnitude of impact is proportional to the square root of order size rather than to size itself [8]. This implies that although the actual impact of small orders roughly matches the linear prediction of the Kyle model, the impact of large orders grows more slowly as size increases, even as its absolute level still far exceeds a linear extrapolation. In perpetual futures markets, this nonlinear impact effect is especially pronounced. Because of the high leverage present, the impact of a large market order not only consumes liquidity but may also trigger other participants' stop-loss orders or forced liquidations, setting off a chain reaction. This nonlinear price jump induced by liquidity consumption is central to understanding the microstructural mechanics of extreme market events such as flash crashes.

8.1.2 The five signal channels of the order book

If the order book is viewed as an information machine, its various observable parameters constitute the analytical dimensions for extracting market signals. Traditional microstructure analysis often studies the spread or depth in isolation. This chapter systematizes these dimensions into a joint diagnostic framework of five core signal channels (Table 8-1). These five channels do not exist in isolation; together they depict a multidimensional picture of the market.

Signal channel	Observable variables	Core information conveyed	Correspondence with Chapter 7
Spread channel	Absolute and relative values of the bid-ask spread	The market's immediate transaction cost; market makers' pricing of adverse-selection risk; a real-time indicator of market "tension."	Corresponds to Section 7.2.4 (the maker-taker game and the economics of the spread). The spread is not only a transaction cost but also compensation to market makers for the risk of information asymmetry.
Depth channel	Resting order quantity at each price level and the cumulative depth curve	The market's capacity to absorb impact without a significant price move; the thickness of liquidity. An imbalance in depth between the bid and ask sides (order imbalance) is one of the strong predictors of short-term price movement [9].	Corresponds to Section 7.2.2 (limit orders as the supply of liquidity). The thickness of depth directly determines the market's capacity to withstand impact.
Slope channel	The rate at which depth decays with distance from the mid price	Participants' "confidence distribution" regarding price deviations from the current level; the steeper the slope, the more concentrated liquidity is at the current price.	Reflects the market's expectations of extreme price movements. A steep slope means the market regards a large price deviation as highly improbable.
Flow channel	The number of trades and the trading volume per unit of time	The market's "activity" and "urgency"; a high trade rate means the market is reacting intensely to new information.	Corresponds to Section 7.6 (the order-flow information ecosystem). Flow is the most direct manifestation of information entering the market.
Dynamics channel	The update frequency and cancellation rate of quotes	Market makers' "tension"; a rising cancellation rate is an early-warning signal that market makers' confidence is declining and that they are preparing to retreat.	Corresponds to Section 7.1.2 (the signal-strength spectrum). A high cancellation rate often foreshadows an imminent sharp market move.

Table 8-1. The five-signal-channel matrix of the order book as an information machine (Data source: constructed by the author)

These five signal channels are not mutually independent; they exhibit significant cross-

correlations. For example, a rise in the quote cancellation rate in the dynamics channel often precedes a widening of the bid-ask spread in the spread channel, while a one-sided contraction of liquidity in the depth channel typically accompanies an intensifying asymmetry in the rate of depth decay in the slope channel. Consequently, jointly monitoring the five channels — for instance, observing an abnormal climb in the cancellation rate even as the spread remains superficially stable — provides far greater diagnostic value than observing any single channel in isolation. This cross-channel joint-analysis framework is precisely the key methodological foundation that elevates the order book from a simple record of quotes into a systemic risk-warning tool.

Figure 8-2 provides a visual mapping of the core observable variables of the five signal channels and their analytical dimensions, showing how they capture market microstructure information from different angles and converge into a comprehensive assessment of the market's overall state, thereby transcending isolated reliance on any single indicator.

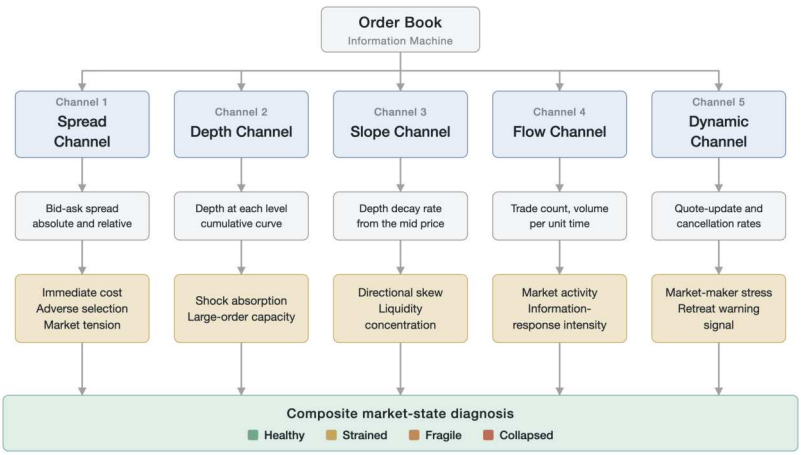


Figure 8-2. The five signal channels of the order book and the integrated diagnostic process (illustration of the integrated diagnostic process across the five signal channels; not empirical data)

The combined patterns across these five channels yield a comprehensive reading of the market state. Analyzing the signals jointly, rather than watching any single indicator in isolation, is what makes systematic diagnosis possible.

For example, in a superficially calm market with a narrow spread and thick depth (channels 1 and 2 healthy) but a sudden surge in the cancellation rate (channel 5 anomalous), the reading is often an early signal that high-frequency market makers have detected toxic order flow and are beginning to retreat. If the slope channel (channel 3) suddenly steepens while the depth channel shows a cliff at a particular price level, this strongly suggests that participants are preparing for a price breakout or a liquidation event. Such subtle structural changes often reveal the market's underlying trajectory earlier than price or volume indicators do.

8.1.3 Signal timeliness and information decay

The signals broadcast by the order book have one key characteristic: their information value decays sharply over time. In perpetual futures markets, this decay is especially rapid.

The depth consumption and price impact caused by a large market order may be filled within seconds — or even hundreds of milliseconds — by high-frequency market makers' quote updates; an abnormal asymmetry in bid and ask depth may return to balance within tens of seconds as arbitrageurs step in. Perpetual futures markets operate around the clock, which means that signals have no cross-day accumulation effect but instead exist in a perpetual cycle of continuous generation and rapid decay. This continuous high-frequency interaction keeps the order book in a dynamic-equilibrium steady state, in which any anomalous signal that deviates from the steady state is quickly smoothed away by market forces.

In the modern crypto market dominated by high-frequency market makers, orders are updated and canceled at extremely high frequency. This extremely short information half-life carries an important methodological implication: the time granularity of any analysis based on order book data must match the half-life of the signal. Using daily or even hourly data to analyze the order book produces a severe mismatch between temporal resolution and signal half-life, causing the analysis to miss almost every meaningful microstructural signal. In the millisecond-scale micro-world, liquidity may have vanished and reformed countless times, and these changes are entirely invisible on minute- or hour-level candlestick charts.

This mismatch in time granularity is precisely why many traditional macro analysts produce inaccurate results in crypto-market analysis. They attempt to explain the market's microdynamics using minute- or even hour-level candlestick charts, overlooking the fact that within these time windows the order book may already have undergone thousands of reconstructions and rebalancings. For instance, a momentary liquidity vacuum triggered by a liquidation may cause the price to plunge and rebound within seconds, leaving only a long lower shadow on the hourly chart. Without descending to the level of tick-by-tick order book data, one cannot understand the true driver of such a price anomaly.

A further distinction is needed: not all price impacts share the same decay characteristics. The microstructure literature decomposes impact into temporary impact (driven by liquidity consumption and market makers' inventory adjustments, which reverts within a relatively short time) and permanent impact (which reflects the incorporation of genuine information into the price and does not revert). In perpetual futures markets, because the liquidation of a leveraged position itself conveys genuine information about market direction (being liquidated means the market's tolerance for positions in that direction has been exhausted), the proportion of permanent impact within total impact may be systematically higher than in traditional spot markets. This distinction has a direct implication for the analysis of signal decay: if an impact is primarily permanent, then waiting for a "reversion" is itself a mistaken expectation.

This explains why, in practice, order book analysis remains the exclusive domain of a small number of professional quantitative institutions. The capacity to acquire, store,

and process millisecond-level or tick-by-tick order book snapshot data constitutes a natural technological and informational barrier. This phenomenon confirms the transparency paradox discussed in Chapter 1: although the crypto market is fully public and transparent at the level of data, the sheer volume of data and the rapid decay of information mean that genuine insight remains in the hands of a small number of institutions commanding top-tier computing resources. These institutions can monitor and parse minute changes across the five signal channels in real time, allowing them to act before the market undergoes a structural shift.

8.1.4 From single signals to state diagnosis

By integrating the five signal channels, we can construct an order book state dashboard that divides the market's microstructural state into four typical, continuous phases: healthy, tense, fragile, and collapsed. It formalizes the "three snapshots" of the introduction and previews, at the order-book level, the liquidity-fragility model of Chapter 20. The specific characteristics of each state across the five channels will be mapped in detail in Section 8.4.4 in conjunction with order book morphology. Here we merely outline the core logic: the transition from "healthy" to "collapsed" is not a sudden event but a traceable, continuous process of deterioration, manifested as a gradual widening of the spread, a contraction of depth from symmetric to one-sided thinness, and a rise in the market-maker cancellation rate from steady climbing to full retreat. The flash crash of October 2025 is a textbook case that evolved precisely along this path.

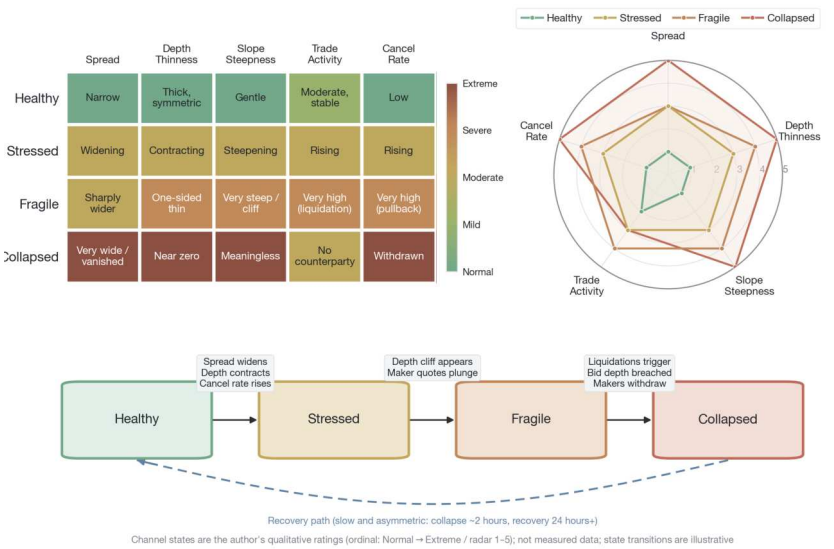


Figure 8-3. The order book state dashboard (the author's state-diagnosis framework: the four states — healthy, tense, fragile, and collapsed — and the 1–5 severity scale are the author's qualitative ordinal ratings, not empirical data; the recovery path of "collapse in about 2 hours / recovery in more than 24 hours" is a representative time scale)

The upper half of Figure 8-3 presents, in matrix form, the typical behavior of the five

signal channels across the four states, while the lower half shows the state transitions from healthy to collapsed and the recovery path. The latter is markedly asymmetric: the deterioration path is short and steep, whereas the recovery path is long and gradual. The correspondence between each state's behavior across the five channels and order book morphology is refined in Section 8.4.4.

This diagnostic capability constitutes the core of modern quantitative trading and risk management. By continuously monitoring the order book's state transitions, traders can dynamically adjust their position sizes and leverage, while exchanges can trigger circuit breakers or liquidity-protection mechanisms in advance. Yet even if we could read these signals perfectly, we would still confront a cognitive bias: is the depth we "see" in the order book truly the depth we can "use"? This question leads to the subject of the next section — the depth illusion.

8.2 The depth illusion

In cryptocurrency markets, and especially in perpetual futures trading, one of the most common errors traders make is to equate the resting order quantity visible on the order book with the trading size the market can genuinely absorb. The depth displayed in the order book misleads along two dimensions, and this is no mere technical detail of microstructure: it is the root cause of severe slippage under stress.

A trader who equates visible depth with usable depth may get by in the everyday low-volatility environment, but at the critical moment when the market faces a shock, that trader will often encounter severe slippage. The high-leverage environment, the absence of an expiry date, and the around-the-clock trading mechanism unique to perpetual futures make this depth illusion far more pronounced than in traditional financial markets (such as equity spot markets or traditional futures markets). During the cryptocurrency flash crash of October 2025, post-mortem data showed that within the first few minutes of the price plunge, "the cumulative depth within 5% of the best bid contracted by more than 90%." This figure directly reflects how the depth illusion manifests under extreme stress. Starting from the underlying microstructural logic, this section examines the dual causes of the phenomenon and proposes a set of actionable quantitative methods.

8.2.1 Flickering liquidity and ghost quotes

When we observe the order book of a modern electronic trading platform, what we see is a constantly flickering, rapidly updating matrix of numbers. Beneath this seemingly thriving appearance of liquidity, however, hides a large volume of what is called flickering liquidity. The quotes of high-frequency market makers and algorithmic traders may exist on the order book for less than a second before being automatically canceled and replaced by the system.

These flickering quotes are counted as valid depth in any static snapshot of the order book at a given instant. In actual trading practice, however, when a market order or an aggressive limit order is submitted to the matching engine, network latency and the market maker's extremely fast cancellation often mean that the quote has already been

canceled before the matching engine processes the order. In other words, the trader "sees" ample depth on the terminal, but by the time it comes to "use" that depth — to submit an order — it is gone. This phenomenon constitutes the first layer of the depth illusion.

Academics have long attended to this phenomenon, usually calling it ghost liquidity or flickering quotes. In studies of traditional equity markets, ample empirical evidence has shown that flickering liquidity accounts for a considerable share of visible depth. In highly electronic markets such as Nasdaq, for example, the prevalence of high-frequency trading has substantially shortened the average lifespan of orders.

In the cryptocurrency perpetual futures market, this phenomenon is not only fully inherited but further intensified by the specific market environment. Market makers in crypto markets face significantly higher adverse-selection risk. In a detailed analysis of crypto-market microstructure, Tiniç et al. (2023) found that the absence of unified regulation and disclosure mechanisms pushes the share of informed trading far above that of traditional markets [10]. This means that when a market maker's passive quote is filled, there is a very high probability that the counterparty is a trader with an information advantage, exposing the market maker to systematic losses.

To cope with this high adverse-selection risk — compounded by the fast and frequent price movements that the crypto market's around-the-clock trading and high leverage produce — the market maker's defensive strategy is high-frequency quote updating. To avoid being "arbitraged" or "sniped" by informed traders, market makers must update their quotes at extremely high frequency, adjusting orders at the millisecond or even microsecond level in response to price signals from other exchanges, changes in on-chain data, or macroeconomic indicators. This directly makes the lifespan of each quote extremely short. Flickering is therefore not only ubiquitous in perpetual futures markets; its scale and frequency far exceed those of traditional markets, making it central to any assessment of liquidity quality in crypto markets.

8.2.2 Conditional depth

If flickering liquidity is a "physical-layer" illusion produced by the microstructural mechanics of high-frequency trading, then a deeper and more destructive illusion lies in the behavioral logic of liquidity providers. Even when certain quotes are "real" (that is, they remain on the order book long enough not to be mere flickering quotes), the existence of that depth is absolutely premised on the implicit condition that market conditions are normal.

In the central limit order book model, the limit order a market maker submits appears on the surface to be a firm commitment to provide liquidity to the market, but in essence it is only a conditional commitment. The market maker's internal logic can be made concrete:

> "As a market maker, I have placed a buy order for 100 ETH at the $2,000 price level, providing liquidity to the market. But note that the effectiveness of this commitment rests on a series of strict implicit preconditions: that the market is currently experiencing no extreme directional volatility, that no signs of an ongoing liquidation cascade have

> been observed, that prices across the major exchanges show no significant divergence, and — most important — that my own inventory risk remains within the manageable range of my internal risk-control model. If any one of these conditions no longer holds and my risk-control threshold is triggered, my algorithm will automatically cancel this order within milliseconds and withdraw from the market."

This internal logic reveals the essence of order book depth: it is absolutely not an unconditional guarantee but a highly conditional commitment. The sole condition for its persistence is that the market state remain within the stable range the market maker's algorithm can tolerate. Once the market state deteriorates — even at the earliest sign of deterioration — these conditional commitments retreat collectively within an extremely short time. This collective retreat causes the seemingly ample visible depth on the order book to vanish rapidly, forming zones of extreme liquidity scarcity.

The discussion in Chapter 2 of "the cost of abolishing the expiry date" clarifies the special nature of this conditional commitment in perpetual futures. Traditional futures contracts have a definite expiry date, so a market maker's inventory risk naturally converges as expiry approaches. The loss of a risk-termination date for perpetual futures market-making, however, means that when providing quotes the market maker effectively bears an open-ended inventory risk. Lacking a risk-release mechanism along the time dimension, market makers become far more sensitive to market volatility. As a result, the liquidity commitments they make in perpetual futures are far more conditional than in traditional markets. Once the market enters a trending move, or volatility rises sharply, market makers — to avoid an indefinite erosion of inventory value — trigger their retreat mechanism at a very low threshold. This structural fragility, arising from institutional design, constitutes the second layer of the depth illusion and is the core reason liquidity dries up during extreme market conditions.

8.2.3 The three-layer decomposition of depth

Given that the depth illusion objectively exists and carries significant destructive power, how can it be identified and quantified in practice? The ILLIQ indicator proposed by Amihud (2002) measures illiquidity through the ratio of the absolute value of the daily return to the trading value, providing a standardized tool for the macro-level measurement of market depth [11]. That indicator, however, is based on daily-frequency data and cannot capture the depth changes at the micro time scale with which this chapter is concerned. To break the cognitive fallacy that "visible equals usable" and to measure the scale of the depth illusion precisely at the order book level, this chapter proposes a three-layer definitional framework of depth (Table 8-2). This layered quantitative framework helps traders and risk managers form more accurate dynamic expectations of the liquidity actually usable under different market states.

Depth layer	Conceptual definition	Measurement and calculation method	Reliability and application scenarios
Nominal depth	The total resting order quantity displayed directly in a static snapshot of the order book at a specific point in time. This is the most intuitive depth an ordinary trader sees on the terminal interface.	$\sum_i q_i$ (where q_i is the resting order quantity at the i-th level, and i ranges over all levels within a set percentage band around the best bid and ask). Typically calculated as the cumulative order size within a certain percentage (such as 1%, 2%, or 5%) of the best bid and ask.	Lowest reliability. It undergoes no filtering and contains a large volume of flickering liquidity and conditional commitments that may be canceled at any time. Suitable only for an extremely stable market environment.
Effective depth	The depth that, after removing high-frequency flickering quotes, is genuinely available with high probability for matching and execution under normal market conditions.	Cannot be read directly from a snapshot. It requires high-frequency historical order-flow data to compute the average lifespan of orders and the actual fill rate, from which a discount coefficient X is calibrated. Effective depth = nominal depth $\times$ X%. The value of X varies significantly by asset class, exchange type, and market state: based on empirical analysis of multiple historical extreme events, the X value for mainstream assets (such as BTC and ETH) on leading exchanges is typically between 60% and 80%, for mid-cap assets between 40% and 60%, and for less liquid long-tail assets possibly as low as 20% to 40%. It must be emphasized that the ex ante measurement of stress depth faces fundamental difficulties; the above ratios are merely empirical estimates based on historical-event retrospection and should not be applied mechanically to future stress scenarios.	Moderate reliability. It reflects the true cost of liquidity in everyday trading, but its premise remains that market conditions are normal, and it cannot withstand extreme shocks.
Stress depth	The depth that actually remains on the order book and is available for execution after market makers retreat collectively when the market faces an extreme shock, volatility spikes, and stress rises sharply.	Can be measured only through ex post retrospective analysis of high-frequency tick-by-tick trade data during historical extreme events (such as flash crashes and liquidation cascades). It measures how much depth truly withstood the test at the moment of crisis.	Highest reliability, but extremely difficult to predict. It is the core indicator for assessing market tail risk and calculating extreme slippage. Stress depth is often only 5% to 15% of nominal depth.

Table 8-2. The three-layer decomposition framework of depth (Data source: compiled by the author)

Through this three-layer division, the scale of the depth illusion can be quantified precisely: the vast gap between nominal depth and stress depth is its absolute magnitude.

Take the crypto-market flash crash of October 2025. The detailed post-mortem report by FTI Consulting (2025) shows that within the first few minutes of the crisis, the nominal depth within 5% of the best bid appeared to be as high as hundreds of millions of dollars, but as liquidation selling pressure surged in and market makers canceled orders instantly, the stress depth that actually supported execution amounted to less than 10% of the pre-crisis nominal depth [2]. This means the reduction from nominal depth to stress depth exceeded 90%. The larger this gap, the more the market's liquidity was built on a highly conditional foundation, and the worse its actual performance under stress relative to how it appeared on the surface. For institutional investors who rely on algorithms to execute large orders, estimating slippage and market impact on the basis of nominal depth will produce execution costs far in excess of expectations.

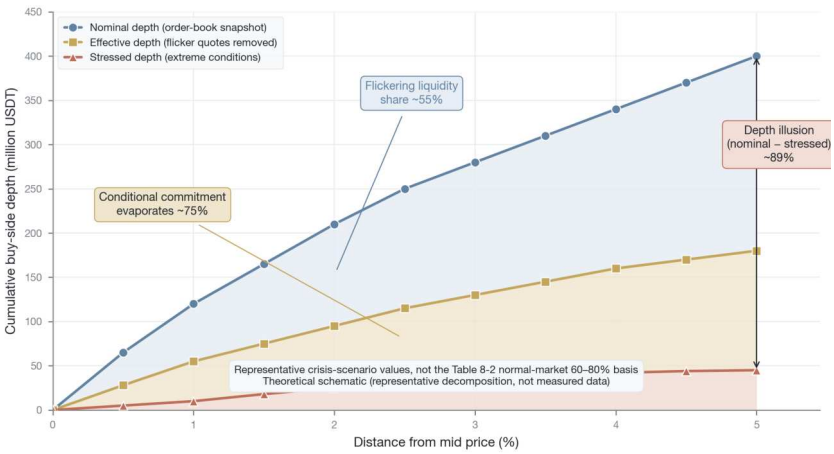


Figure 8-4. A three-layer comparison of nominal depth, effective depth, and stress depth

Note: This figure is a theoretical illustration drawn from the characteristic parameters of the October 2025 flash crash. The flickering, conditional evaporation, and nominal-to-stress depth gap shown are the author's representative decomposition, not empirical data, and are representative values for a crisis scenario rather than the normal-condition measure of Table 8-2 ("effective depth = nominal × 60–80%"); the figure follows FTI Consulting (2025) [2]. Within 5% of the mid price, the reduction from nominal depth to stress depth adopts the same ">90%" measure as Section 8.2 in the main text (the roughly 89% shown is a representative approximation of that reduction), indicating that under extreme market conditions the liquidity a trader can actually rely on is only about one-tenth of the nominal value displayed on the trading terminal. For institutional traders who design large-order execution algorithms, estimating expected slippage on the basis of nominal depth alone will produce execution results that deviate from expectations by an order of magnitude at the moment of crisis.

8.2.4 The amplification mechanisms of the depth illusion

Why is the depth illusion in perpetual futures markets far more severe than in spot markets or in traditional low-leverage markets? The reason is that the underlying

design of perpetual futures embeds two core mechanisms that, under stress, together amplify the depth illusion.

The first amplification mechanism is the anticipation of a liquidation cascade and the resulting defensive order cancellation. In the modern crypto market, market makers possess extremely advanced data-analysis capabilities. On decentralized exchanges (DEXs, such as application chains with an on-chain CLOB architecture), they can precisely calculate the liquidation trigger points across different price ranges by analyzing the fully public on-chain position and liquidation data; on centralized exchanges, they rely on sophisticated machine-learning models and historical experience to estimate the location of dense liquidation zones.

As the price gradually approaches these dense liquidation zones, market makers can clearly foresee the characteristics of the future order flow: "If the price reaches this trigger point, the liquidation engine will automatically take over and submit a large volume of forced sell orders to the market. These orders are highly toxic: they are involuntary, entirely insensitive to price (usually issued as market orders), and, because of the multiplier effect of high leverage, potentially far larger than normal orders."

Faced with this highly certain and enormously destructive anticipation of "toxic order flow," a rational market maker will not remain in place as the counterparty to liquidation orders. Market makers therefore adopt a defensive strategy, sharply reducing — or even entirely withdrawing — their bid quotes near the dense liquidation zone in advance. This anticipation-based collective cancellation forms a liquidity vacuum at specific price levels on the order book, which we call a depth cliff. This preemptive retreat makes the already-thinned depth even more fragile, directly amplifying the destructive power of the depth illusion.

The second amplification mechanism is the compounding effect of high leverage and procyclical retreat. High leverage amplifies not only traders' profits and losses but also the inventory risk market makers face. Market makers inevitably accumulate inventory when providing liquidity. When the market moves in one direction — especially a leveraged one-directional move — the opposing positions market makers hold face accelerating losses.

As discussed in Section 8.2.2 regarding "the cost of abolishing the expiry date," perpetual futures lack a delivery date that would force prices to converge and release risk, so market makers bear an open-ended inventory risk. Layered on top of this is the loss acceleration that leverage brings. Unable to judge how long an adverse trend will last, market makers see their tolerance (threshold) for retreat pressed extremely low.

After analyzing the microstructural data of the flash crash, FTI Consulting (2025) confirmed this point: during the crisis, the speed of market makers' retreat reached the millisecond level, and the scale of the retreat was unprecedented [2]. This reveals a dangerous "procyclicality" in market-maker behavior: precisely when the market is highly volatile and most in need of liquidity to absorb the shock and stabilize prices, market makers instead have the strongest incentive — out of the rational motive of risk avoidance — to withdraw liquidity on a large scale. The superposition of this procyclical retreat and the high-leverage environment exposes the depth illusion to its fullest extent at the most critical moment.

Beyond these two rational mechanisms, the reflexive effect of information transparency constitutes a third amplification channel. In cryptocurrency markets, liquidation data are displayed to the entire market in real time through aggregation platforms such as CoinGlass. This transparency aids price discovery and risk assessment in normal times, but during a crisis it produces a counterproductive accelerating effect: when traders observe on social media and data dashboards that liquidation volume is climbing sharply, these publicly visible liquidation data themselves turn into a panic signal, triggering more participants to close positions voluntarily and thereby generating even larger-scale liquidations. This constitutes a reflexive mechanism: the transparent disclosure of liquidation information in turn accelerates the spread of liquidations, making the crisis self-reinforce far faster than in comparable events under an information-opaque environment.

At the same time, during extreme stress events, arbitrageurs — who ought to perform a stabilizing function — may also fail. Under normal market conditions, arbitrageurs push prices back toward equilibrium by capturing cross-market or cross-asset price deviations. Brunnermeier and Pedersen (2009), however, revealed the mechanism of the liquidity spiral: when the market is highly volatile, arbitrageurs themselves face margin calls and are forced to close out their arbitrage positions to meet funding needs [12]. These forced closures not only fail to narrow the price deviation but instead worsen it, and the force that should anchor prices exits the market at precisely the moment it is most needed. In the high-leverage environment of crypto perpetual futures, arbitrageurs' capital constraints are even tighter, so this "stabilizer failure" is correspondingly more common.

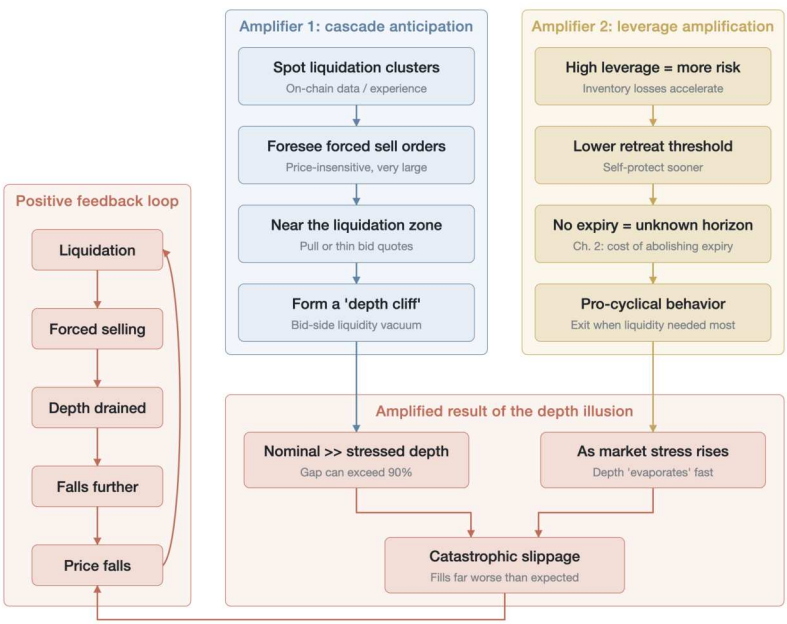


Figure 8-5. The causal chain of the depth-illusion amplification mechanisms (illustration of the mechanism causal chain, not empirical data; this figure focuses on the two rational retreat mechanisms

— liquidation-cascade anticipation and high-leverage procyclicality — while the third amplification channel, the reflexivity of information transparency and the arbitrageur liquidity spiral, is discussed in Section 8.2.4 of the main text, the latter following Brunnermeier & Pedersen 2009 [12]; "2009" is the citation year of that study, not the date of the flash crash)

The left side of the figure shows how market makers' defensive cancellation based on liquidation anticipation forms a depth cliff, and the right side shows the procyclical retreat triggered when high leverage and the absence of an expiry date lower the retreat threshold. The nonlinear interaction of the two (anticipatory cancellation drains the depth buffer at key price levels in advance, while procyclical retreat destroys the residual liquidity when volatility spikes) constitutes a "double vacuum" effect, making the actual severity of the liquidity crisis far exceed what either factor alone would predict.

8.2.5 Case study: the Hyperliquid POPCAT phantom order-wall incident

The preceding analysis focused mainly on the depth illusion caused by market makers' unintentional conditional retreat. The POPCAT incident on the decentralized perpetual futures platform Hyperliquid in November 2025, however, shows that the depth illusion can arise not only passively as an endogenous fragility of the market microstructure but can also be actively constructed and weaponized by a trader to manipulate the market price and transfer risk.

Viewed from a legal perspective, this behavioral pattern (first building a position, then placing a large order with no genuine intent to trade in order to induce others to follow, and subsequently canceling the order for profit) constitutes the classic spoofing conduct of traditional securities law. Under Section 747 of the U.S. Dodd-Frank Act, such conduct is explicitly defined as spoofing — that is, submitting a bid or offer "with the intent to cancel before execution." The 2015 Navinder Sarao case established a precedent for the criminal liability of such conduct, with the court finding that his repeated placement and cancellation of large orders in the S&P 500 futures market constituted market manipulation. On an on-chain decentralized exchange, however, legal enforcement against such conduct faces unique challenges: the manipulator uses anonymous wallet addresses, operates across jurisdictions, and the decentralized protocol itself lacks the monitoring and enforcement obligations of a traditional exchange. This enforcement gap leaves on-chain perpetual futures markets facing a more severe institutional deficit in defending against manipulation than traditional markets.

On November 12, 2025, an anonymous trader withdrew about $3 million in USDC from OKX, dispersed it across 19 wallet addresses, and then built a POPCAT long position on Hyperliquid with total exposure of roughly $26 million to $30 million at about 5x leverage. After completing the position, the trader placed a buy limit order of about $20 million at the $0.21 price level. This order formed a bid-side "order wall" on the order book that far exceeded the normal depth level of that market, transmitting a signal of strong buy-side demand to other market participants [13]. Attracted by this surface depth, other traders followed into long positions, further raising the market's leverage level and long-side concentration.

From a microstructural perspective, the core of the incident is that the $20 million order

wall was, in essence, an artificially manufactured depth illusion. Unlike the flickering liquidity discussed in Section 8.2.1 and the conditional depth analyzed in Section 8.2.2, this order wall was never intended to be filled from the moment it was submitted; its sole function was to manufacture visible depth in order to induce others to follow. Once long positions had accumulated to a sufficient scale, the manipulator withdrew the order wall. The $20 million of nominal depth dropped to zero in an instant, and the high-leverage long positions held by the follower traders who had relied on this visible depth suddenly faced an order book with a severe shortage of real depth. The POPCAT price fell about 43% within minutes (from about $0.21 to about $0.12), triggering about $63 million in platform-wide liquidations. Because the forced sell orders generated by the liquidations far exceeded the true absorptive capacity of the order book, Hyperliquid's community liquidity pool, HLP, ultimately bore about $4.9 million in bad debt [13].

Within Hyperliquid's risk architecture, HLP plays a dual role: it is both the liquidity provider for everyday market-making and the final risk bearer of the liquidation system. When the liquidation engine forcibly closes a position, if the order book lacks sufficient counterparty quotes, the unmatched exposure is first absorbed by the platform's insurance fund. When the insurance fund is insufficient to cover the entire loss, the remaining bad debt is distributed to HLP's liquidity providers through a socialized-loss mechanism. In the POPCAT incident, the insurance fund was rapidly exhausted, and the $4.9 million in bad debt was ultimately borne jointly by HLP's participants in proportion to their contributions. This transmission path (liquidation engine $\rightarrow$ insurance fund $\rightarrow$ HLP socialized allocation) reveals the ultimate destination of risk in a decentralized liquidation system.

From an operational-risk perspective, the incident also exposed multiple defensive deficiencies in Hyperliquid's market monitoring: the dispersed but synchronized position-building across 19 linked wallets (which would typically trigger an alert in the linked-account detection systems of traditional exchanges), the highly concentrated position in a single asset (with total exposure representing a significant share of POPCAT's market-wide open interest), and the $20 million order wall exceeding the normal size distribution. These three anomalies should have triggered, respectively, linked-account, position-concentration, and large-order alerts, yet all failed simultaneously — indicating that decentralized perpetual futures platforms still exhibit a significant institutional lag in the transition from technical architecture to an operational risk-control system.

This incident provides a case of special analytical value for the theoretical framework of the depth illusion. In flickering liquidity and conditional depth, the depth illusion is market makers' spontaneous behavior grounded in their own risk-management logic, intended to protect capital rather than to mislead the market; POPCAT reveals a third form — strategically constructed false depth — in which the manipulator turns the market's default trust in visible depth (the cognitive inertia that "the depth one sees is the depth one can use") into a tool of manipulation. From the perspective of behavioral finance, the $20 million order wall created a strong anchoring effect, causing participants to systematically overestimate the strength of support at that price level and to add leverage accordingly; combined with the availability heuristic (the tendency to rely in decision-making on the most readily available visual information), this explains why followers could gather at large scale in a short time. This strategic

depth illusion is especially dangerous in long-tail-asset perpetual markets: the lower the normal depth, the greater the relative illusion a single large order creates and the stronger its distortion of others' decisions. After the incident, Hyperliquid suspended deposits and withdrawals and launched a risk-control review, becoming a textbook case of the institutional-design challenges that decentralized perpetual platforms face in defending against market manipulation.

8.3 The distinctiveness of the perpetual futures order book

Having dissected the depth illusion, we must further recognize that the perpetual futures order book is not a simple replica of the traditional spot order book. It is deeply reshaped by a series of financial rules unique to the underlying design of perpetual futures. The leverage mechanism, the forced-liquidation system, the funding rate that maintains the price anchor, the dual-track price system (mark price and last traded price), and the vast open interest all give rise, at the microstructural level of the order book, to a series of distinctive phenomena that simply do not exist in spot markets.

This section analyzes systematically how these five core mechanisms shape the dynamic behavior of the order book. This analysis deepens our understanding of market microstructure and lays the empirical and logical groundwork for the "four fingerprints of the perpetual futures order book" framework proposed later in this chapter.

8.3.1 The order book transmission mechanism of the liquidation cascade

In perpetual futures markets, the liquidation order is key to understanding the microstructure of extreme markets. It is a highly unusual order type: it is not actively issued by a human trader or a conventional quantitative strategy based on a market judgment but is executed automatically and compulsorily by the exchange's liquidation engine when a user's margin ratio falls below the maintenance-margin requirement.

This compulsory and special origin gives liquidation orders two microstructural characteristics that carry significant systemic risk. The first is absolute price insensitivity: in order to take over and close the position as quickly as possible to prevent the position from going underwater, the liquidation engine typically throws the position into the market as a market order or an extremely aggressive limit order (such as a bankruptcy-price limit order), executing an immediate fill with no price limit and inflicting a direct and enormous shock on the order book's available depth. Accompanying this is the adverse timing of their appearance: liquidation orders never appear in calm conditions; their triggering is necessarily accompanied by a market that has already moved sharply against the liquidated position. The influx of liquidation orders therefore intensifies the market shock and reinforces the existing trend and volatility.

From the standpoint of the micro-transmission path, the positive feedback loop of the liquidation cascade forms a self-reinforcing chain: "initial decline → liquidation of high-leverage positions → price-insensitive market sell orders consume bid-side depth → accelerated price decline → deeper liquidations triggered → procyclical retreat of market makers → liquidity vacuum." In this loop, the collective retreat of market

makers plays a key accelerating role, pushing the order book from "thin depth" into a "liquidity vacuum" and thereby amplifying a normal price correction into a systemic liquidity collapse.

The flash crash of October 2025 provides empirical validation of this transmission mechanism. According to the post-mortem report by FTI Consulting (2025), after the structural long position that a large fund had built near $86,000 was triggered for liquidation, the bid-ask spread of the BTC/USDT perpetual futures widened within minutes from less than 10 basis points to the thousand-basis-point range recorded on individual platforms at the peak of the cascade (spread snapshots vary enormously across exchanges and across moments in time; the thousand-basis-point level here is an extreme reading on a thin-book platform at the peak of the collapse, which differs in measure and moment from the 26.43 basis points recorded on a mainstream platform at 21:15 UTC cited earlier), with cumulative forced liquidations reaching $9.89 billion over about 14 hours [2][1]. The complete micro-transmission mechanism of the liquidation cascade on the order book (in particular, how market makers form a defensive depth cliff) is analyzed further in the four-fingerprints framework in Section 8.5.1.

The failure of the existing risk-buffer mechanisms to interrupt the cascade during this crisis deserves scrutiny. Mainstream cryptocurrency exchanges typically deploy three lines of defense: an insurance fund to absorb underwater losses, an auto-deleveraging mechanism that hedges and closes profitable-side positions when the insurance fund is insufficient, and position limits and price-protection mechanisms that constrain extreme behavior. In the October 2025 event, the insurance funds of several exchanges were substantially depleted during the peak of the cascade, and some exchanges triggered their auto-deleveraging programs, but the design parameters of these mechanisms (the size of the insurance fund relative to potential liquidation exposure, the trigger thresholds for auto-deleveraging, and the width of the price-protection band) had never been stress-tested against a synchronized liquidation event of this scale. The existence of the safety net gave market participants a sense of institutional security, but the gap between the safety net's actual carrying capacity and the extreme liquidation pressure is precisely the institutional reason the cascade could not be interrupted.

This transmission mechanism bears a surface resemblance to — but a fundamental mechanistic difference from — the "hot potato effect" described by Kirilenko et al. (2017) in their microstructural analysis of the U.S. equity flash crash of May 2010 [14]. In the 2010 U.S. equity flash crash, the "hot potato" referred to the rapid passing of toxic inventory among high-frequency traders — that is, a position repeatedly changing hands among multiple intermediaries, with each transfer accompanied by a further deterioration in price. In the liquidation cascade of crypto perpetual futures, by contrast, the driving force is not the lateral transfer of positions but their one-directional elimination: the forced market orders generated by the liquidation engine are entirely involuntary, absolutely price-insensitive, one-directional selling. Their distinctive destructive power lies precisely in this complete price insensitivity: they do not "wait" for a reasonable price but execute immediately at any available price, punching directly through all available depth in the order book.

The destructive power of the liquidation cascade spreads further through cross-market

contagion paths. Under a unified margin system, the liquidation of a BTC position in an account consumes all the available margin in that account, forcing the trader to simultaneously reduce positions in ETH and other altcoins to free up margin, thereby transmitting the pressure from the BTC market to the entire crypto derivatives market. In addition, liquidation pressure on centralized exchanges spills over to decentralized exchanges through arbitrageurs' cross-platform activity: when the price of BTC perpetual futures on a CEX falls significantly below the DEX quote because of liquidation selling pressure, arbitrageurs execute sell orders on the DEX to capture the spread, replicating the CEX liquidity crisis in the on-chain market. This multilayered contagion network makes the reach of the liquidation cascade extend far beyond the boundary of any single trading pair or single platform.

Figure 8-6 aligns the spread (Panel A), depth (Panel B), and liquidation volume (Panel C) of the October 2025 flash crash along the same time axis.

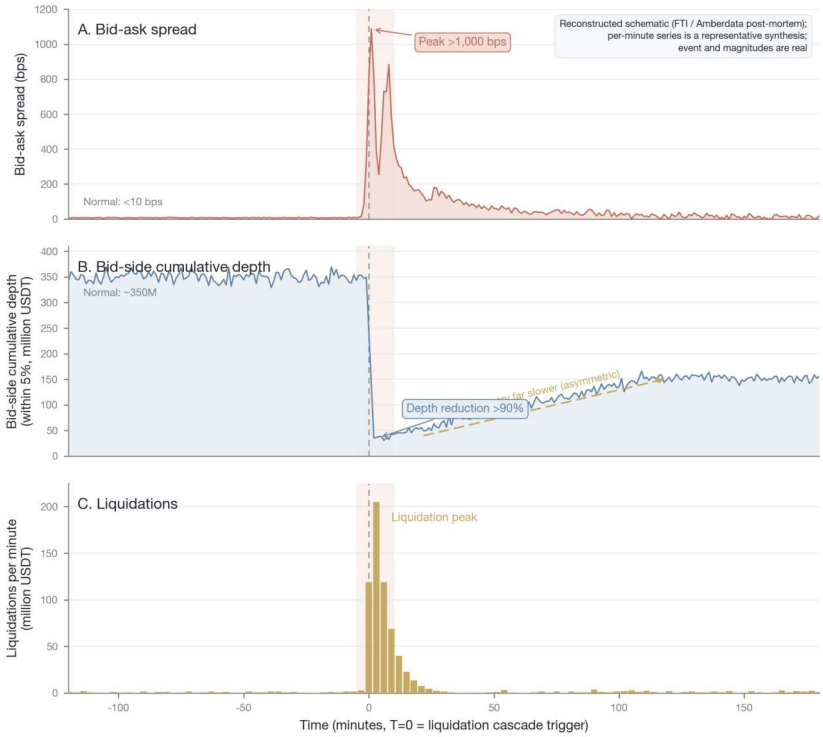


Figure 8-6. The microstructure of the October 2025 flash crash in three panels (a stylized three-panel view of the October 2025 flash crash: the spread peak of >1,000 basis points, bid-side depth of about $350M contracting by >90%, and the liquidation peak of about $200M are representative synthetic values; the event is real, but the minute-by-minute waveform is a post-mortem reconstruction rather than raw empirical measurement; the spread peak is on a thin-book-platform basis, distinct from the 26.43 basis points recorded on a mainstream platform at 21:15 UTC in the main text; following Amberdata 2025 [1] and FTI Consulting 2025 [2])

The temporal alignment of the three panels reveals a key empirical finding: the widening of the spread (Panel A) and the evaporation of depth (Panel B) were almost perfectly

synchronized, indicating that the collective retreat of market makers and the influx of liquidation selling pressure were concurrent rather than sequential processes. The depth recovery in Panel B, in turn, exhibits the asymmetry of "rapid collapse, slow repair" (collapse in minutes, recovery over hours or more), which is analyzed further in Section 8.6.

8.3.2 The cyclical influence of the funding rate

The reason perpetual futures can remain "perpetual" without their price drifting far from spot lies, at its core, in the distinctive funding rate mechanism. This mechanism anchors the perpetual futures price to the spot index through a periodic exchange of fees between longs and shorts. Carried out every 8 hours (or every hour on some high-frequency platforms), however, this compulsory settlement introduces a distinctive cyclical imbalance into the order book's microstructure.

When the funding rate is positive and high (which is relatively common in the crypto market's systematic long bias), longs must pay a fee to shorts. For cost-sensitive traders and arbitrage funds, their collective avoidance behavior — briefly closing positions before the settlement moment (such as UTC 00:00, 08:00, and 16:00) and re-establishing them afterward — has a cyclical structural effect on the order book.

In an empirical study of cryptocurrency perpetual futures, Ruan and Streltsov (2022) found that the spot bid-ask spread and trading activity exhibit a U-shaped pattern within the 8-hour funding rate cycle: in the period approaching funding settlement, informed trading and adverse selection rise (more markedly when the absolute value of the funding rate is large), prompting market makers to widen their quoted spreads [15]. This cyclical variation in liquidity, driven by an institutional mechanism, constitutes the empirical basis for the "depth tide" fingerprint in Section 8.5.2.

8.3.3 Order book fragility under high leverage

Leverage is the core attraction of derivatives markets, but from a microstructural perspective it brings both capital efficiency and systemic risk. To understand the fragility of the perpetual futures order book, one must first understand the highly nonlinear mathematical relationship between the leverage multiple and the "liquidation distance" (the percentage margin by which the current price is separated from the price that triggers forced liquidation).

In the most simplified model, the liquidation distance can be approximated as 1/leverage. The actual liquidation-trigger mechanism, however, is far more complex than this simplified formula. First, the existence of the maintenance-margin ratio further shortens the liquidation distance: for an isolated-margin long (entry price P_0, leverage L, maintenance-margin ratio m, ignoring fees and funding), the exact liquidation distance is $\dfrac{1/L - m}{1 - m}$, whereas liquidation distance $\approx 1/L - m$ is only a first-order approximation under small m (for the short direction the denominator should become $1 + m$). Taking Binance as an example, the maintenance-margin ratio corresponding to the highest leverage tier for BTCUSDT is about 0.4%; by this approximation, the actual liquidation distance at 100x leverage is not 1% but about 0.6% (the exact

expression gives $(0.01 - 0.004)/0.996 \approx 0.60\%$, consistent with the approximation). Note that the formula above does not include opening/closing fees and accumulated funding, both of which would further shorten the actual liquidation distance. Second, liquidation behavior differs fundamentally between isolated-margin and cross-margin modes: under isolated margin, the liquidation distance is determined strictly by the margin of the individual position, whereas under cross margin all available account balance can be used to maintain the position, so the liquidation distance depends on the account-level aggregate risk exposure. In addition, most mainstream exchanges have introduced a tiered margin-ratio system, in which large positions require a higher maintenance-margin ratio than small ones, as well as partial-liquidation mechanisms (reducing position size before full liquidation in an attempt to restore margin adequacy) and auto-deleveraging mechanisms. These institutional-level nuances have a material effect on the actual timing of liquidation triggers and on the transmission speed of the cascade.

This inverse relationship means that as the leverage multiple increases, the margin for error contracts exponentially. For a moderate investor using 2x leverage, the market would need to move about 50% adversely to trigger liquidation, which usually requires an extreme macro black-swan event; at 10x leverage, the liquidation distance shrinks to about 10%, which falls within the range of normal monthly or even weekly volatility in the crypto market.

When the leverage multiple is pushed up to 50x, 100x, or even 125x, however, the situation changes qualitatively. The liquidation distance corresponding to 100x leverage is a mere 1%; at 50x leverage it is only 2%. We define the region where the liquidation distance is less than 2% as the "extreme-risk zone." In a market such as cryptocurrency — where Bitcoin's average daily volatility can reach 3% to 5% (based on the median range of the absolute value of daily BTC/USD returns from 2023 to 2025), and altcoin volatility is higher still — a price movement of 1% to 2% falls within the normal range of the market's everyday volatility.

In the euphoric phase of a bull market, large numbers of retail traders and aggressive capital pour in, using extremely high leverage to build long positions. This produces a microtopological structure with a high degree of systemic risk: the liquidation price lines of a vast number of traders overlap heavily and cluster within a narrow band just 1% to 2% away from the current market price.

This position-distribution feature caused by high leverage is structurally embedded in the fragility of the order book. In such an environment, no macro bad news and no deterioration in fundamentals are needed; the normal fluctuation of order book liquidity, or the ordinary market impact caused by a medium-sized market order, is enough to move the price down by 1%. This tiny 1% drop instantly touches the first dense liquidation cluster, detonating the liquidation-cascade chain reaction described in Section 8.3.1.

Therefore, the extreme fragility that the perpetual futures order book exhibits when facing liquidation pressure is not a technical defect of the matching engine, nor an accidental market failure, but an inevitable product of the institutional design of high leverage, and it directly constrains the order book's resilience under extreme stress.

If 100x leverage means almost certain liquidation under normal volatility, why do so many traders still choose it? The prospect theory of Kahneman and Tversky (1979) offers an explanatory framework [16]: humans exhibit a systematic distortion of probability weighting when evaluating low-probability events, tending to assign excessive subjective weight to low-probability, high-payoff events. The potential payoff under 100x leverage (a 1% price movement leveraging a 100% return on principal) is experienced psychologically as a kind of "lottery ticket," and its extremely asymmetric profit-and-loss structure precisely caters to this bias in the probability-weighting function.

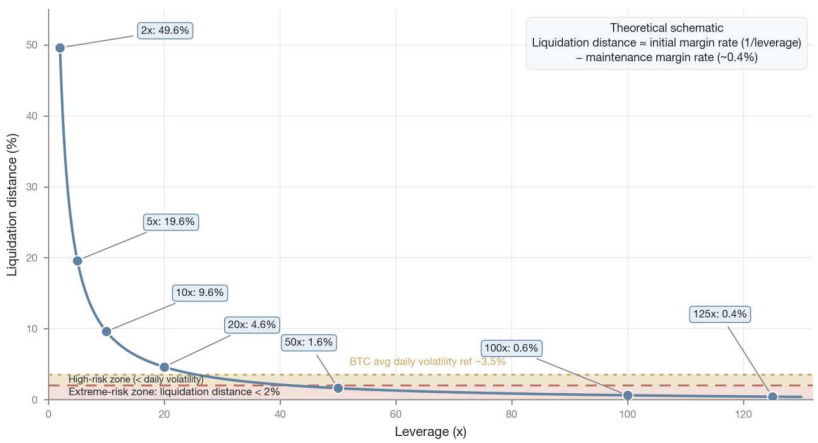


Figure 8-7. The nonlinear relationship between the leverage multiple and the liquidation distance

Note: This figure is a theoretical illustration. The curve is drawn according to "liquidation distance ≈ initial margin ratio (1/leverage) − maintenance-margin ratio (about 0.4%)," consistent with the exact expression $(0.01-0.004)/0.996 \approx 0.60\%$ in this section, and all labeled points are theoretical calculations, not empirical data. The hyperbola shows that once the leverage multiple exceeds 20x, the liquidation distance enters a steeply declining segment, in which the liquidation distance reduced by each additional unit of leverage is disproportionately amplified. The "extreme-risk zone" in the figure (liquidation distance $< 2\%$) and the BTC average-daily-volatility reference line (about 3.5%, taken from the 3%–5% range in the main text) together delimit an operational dividing line: high-leverage positions falling within this zone can be liquidated under everyday price movements, which mathematically explains why liquidation events cluster and erupt in the high-leverage range. The behavioral explanation for the high-leverage preference follows Kahneman and Tversky (1979) [16]; "1979" is the publication year of prospect theory, not a data point.

8.3.4 The divergence signal between mark price and traded price

In traditional spot markets, the price is singular: the price of the most recent matched trade is the current market price. To prevent market manipulation and unfair liquidations caused by anomalous wicks, however, perpetual futures introduce a dual-track price system: the last traded price and the mark price. These two prices play distinct

yet intertwined roles in the order book microstructure.

The last traded price, abbreviated LTP, is the final result of the actual matching of buyers' and sellers' funds on the order book; it reflects the instantaneous supply-demand balance on that specific exchange at the current microsecond. It is directly affected by that exchange's local liquidity depth, the impact of large orders, and market-maker behavior.

The mark price, by contrast, is a synthetic indicator, typically computed from a price index of several leading spot exchanges (such as Binance, Coinbase, and Kraken) plus a funding basis that decays over time. Its core function is to calculate users' unrealized profit and loss and to serve as the sole benchmark price for triggering forced liquidations. The mark price was designed from the outset to filter out short-term anomalous fluctuations on any single exchange's local order book.

The specific construction of the mark price — the weighted or median synthesis of the spot prices of multiple constituent exchanges, the removal of outlier constituents that deviate from the median beyond a threshold, and the exponential-moving-average smoothing of the basis between the perpetual price and the index price, together with the differences across exchanges in algorithm and sampling window — belongs to the domain of price-benchmark governance, which Chapter 14 develops systematically. For the purposes of this chapter, it suffices to grasp the essential point: this multi-source synthesis and smoothing is exactly what lets the mark price screen out short-term anomalies on any single exchange's local book.

Well designed as it is, the mark price mechanism still faces operational risk. The Mango Markets incident of October 2022 provides a cautionary case: the attacker manipulated the oracle quote by artificially pushing up the token price on a constituent exchange with extremely low liquidity, thereby raising the mark price, and on that basis borrowed and withdrew more than $110 million in assets (>$110M, on the CFTC/DOJ/SEC basis) [17]. In addition, under extreme market conditions, the price feeds of constituent exchanges may be delayed by API failures or network congestion, causing the mark price to fail to reflect the true state of the market in time and thereby creating the risk of improper or delayed liquidation.

Under a normal liquidity environment, cross-exchange arbitrageurs quickly eliminate any small spread between the LTP and the mark price, keeping the two closely aligned. Makarov and Schoar (2020) show that cross-exchange price deviations in crypto markets are typically small in normal times but widen significantly during periods of market stress, and that frictions in capital flows are the main reason price deviations persist [18].

A distinction is needed: three layers of price deviation exist in perpetual futures markets. The first layer is the deviation between the traded price and the mark price within the same exchange, which reflects the degree of supply-demand imbalance on that exchange's local order book. The second layer is the deviation of the price of the same contract across different exchanges — the cross-exchange arbitrage opportunity studied by Makarov and Schoar (2020) [18]. The third layer is the basis deviation between the perpetual futures price and the spot price, which is the target the funding rate mechanism seeks to correct. All three layers of deviation are at low levels in a

normal market but widen synchronously and reinforce one another during a crisis.

When the market encounters extreme volatility, however, or when a particular exchange's local liquidity faces exhaustion, these two prices diverge significantly.

This divergence is itself a highly valuable microstructural signal. When the LTP falls significantly below (or rises significantly above) the mark price, and the magnitude of the deviation exceeds the normal arbitrage-cost threshold, it clearly indicates that this exchange's local order book is bearing enormous one-directional selling pressure (or buying pressure) and that market makers and arbitrageurs have stopped providing liquidity because the risk is too high or their capital is exhausted. The local liquidity reservoir has dried up and can no longer effectively maintain the anchor between the price and the global macro fair value.

The degree of divergence between the LTP and the mark price can therefore be regarded as a real-time gauge of the pressure borne by the order book. When constructing an advanced trading algorithm or risk-control system, this deviation indicator should be used as a key supplementary signal to the state dashboard (as described in Section 8.1.4). It also demonstrates, conversely, the importance of using the mark price for liquidation: if, at the instant a liquidity drought causes the LTP to plunge, the system were to use the LTP to trigger liquidations, countless positions that were originally safe would be liquidated unjustly, setting off an even more devastating death spiral. The mark price mechanism is a necessary institutional safeguard for high-leverage derivatives markets.

8.3.5 Open interest as a harbinger of latent order flow

Synthesizing the analysis above, we must extend our view from the instantaneous snapshot of the order book to a macro indicator that measures the market's overall stock of leverage — namely open interest, abbreviated OI. OI represents the total value or total quantity of all derivatives contracts that have been matched and executed but not yet closed and settled at a specific moment.

Within the framework of microstructural analysis, open interest is not merely a static size figure; at the analytical level, its essence is a macro reservoir that foreshadows latent future order flow. The logic can be stated briefly as follows: derivatives trading is a zero-sum game, and every contract currently in an open state (whether long or short) will, at some future moment, be closed out through a trade in the opposite direction. The 10 BTC a trader buys to open a long today will ultimately be closed by an opposing trade — whether the holder actively sells or the liquidation engine or auto-deleveraging mechanism closes the position. In a low-leverage traditional market, these future closing order flows are typically dispersed, rational, and price-sensitive. In the high-leverage environment of perpetual futures, however, the situation is entirely different. As noted earlier, a considerable proportion of the vast OI is composed of high-leverage capital. This means that this portion of latent future order flow is highly likely, when the market encounters headwinds, not to appear as gentle limit orders but to be converted into forced market orders executed automatically by the liquidation engine, potentially far larger than normal in scale. The absolute size of OI, the growth slope of OI, the market's long-short position ratio, and the density of leverage distribution across different price ranges inferred from on-chain data therefore constitute the core

set of indicators for assessing the potential fragility of the current order book under future stress.

An order book that on the surface has low volatility and symmetric, ample bid-ask depth is misleading in its appearance of "health" if lurking behind it are open interest at a historical high, an extremely skewed long-short ratio (for example, a markedly elevated share of retail long positions), and a large accumulation of high-leverage positions clustered just below the current price. Once a tiny price fluctuation triggers the first round of liquidations, the latent closing pressure embedded in the vast OI rapidly converts into a large-scale one-directional forced-liquidation order flow.

In quantitative analysis, the OI-to-market-cap ratio is widely regarded as a core indicator of the degree of leverage accumulation for a specific asset. When this ratio spikes into the historically high zone, it becomes a highly salient early warning: leverage within the system has reached a historical extreme, and any tiny disturbance may trigger a violent deleveraging process. The introduction of this indicator elevates static order book analysis to the level of dynamic systemic-risk assessment, and it takes on diagnostic value that directly determines the efficacy of the analysis when the superposed effect of the four fingerprints is discussed later.

OI as a measure of leverage accumulation, however, has several inherent limitations. First, the statistical measure of OI includes a large volume of hedged positions (for example, in a delta-neutral strategy both the long and short legs are counted in OI, yet their net directional risk is zero). Second, different exchanges differ in how they compute OI (whether by number of contracts or by dollar value, whether expired-but-unsettled positions are included, and so on), so a simple aggregation across exchanges may result in double counting. A rise in the OI-to-market-cap ratio therefore does not necessarily indicate the accumulation of directional leverage. The net long-short ratio (that is, net long-short ratio = (long OI−short OI)/total OI, the share of the difference between long OI and short OI in total OI) and the estimated leverage ratios computed by third-party data platforms, as complementary indicators, can more precisely reflect the actual degree of directional-risk accumulation in the market.

8.4 Topological features of the order book

The order book is not merely a matching engine but a complex information machine. Cont, Stoikov, and Talreja (2010), through statistical modeling of the limit order book, demonstrated that the order book's shape features (rather than merely the absolute value of depth) have significant predictive power for short-term price movements [19]. The order book therefore possesses not only a "size" that reflects market capacity (that is, depth) but also a "shape" that reflects the market microstructure (that is, topological features). The shape of the order book — including its symmetry, slope, and convexity/concavity — carries rich information about the balance of directional forces and the composition of participants. Combined with the signal channels described earlier, these topological features constitute a mapping from micro-morphology to macro market states.

8.4.1 Symmetry and asymmetry

In analyzing the order book's topological structure, symmetry is one of the most intuitive and important features — the core dimension for quantifying the balance of buying and selling forces. In a fully "healthy" market in a random-walk state, the depth distribution on the two sides of the order book is typically approximately symmetric. This symmetry indicates that market makers and liquidity providers assess the risk of a price move in either direction as balanced, and that the forces of buyers and sellers are roughly equal.

When the market develops a directional preference or expectation, however, the symmetry of the order book breaks. The emergence of asymmetry is a clear signal that a significant change is occurring in the market microstructure. For example, when bid-side depth significantly exceeds ask-side depth, it usually means that buying power is dominant and that the market may face upward pressure. From the market maker's standpoint, however, this asymmetry may carry a different message: buyers may be exhibiting an overly aggressive tendency, while concentrated sell orders not yet revealed may exist at higher price levels above. Conversely, if bid-side depth thins while ask-side depth remains thick, this often indicates that selling power is dominant, or that market makers have become more cautious on the bid side.

In perpetual futures markets, asymmetry carries a special meaning that spot markets lack. Because of high leverage and the forced-liquidation mechanism, order book asymmetry often carries key information about liquidation risk. If bid-side depth suddenly thins sharply at a specific price level below the current price, this very likely means that market makers foresee a large cluster of long-liquidation trigger prices near that level. To avoid becoming the counterparty to toxic order flow in a liquidation cascade, market makers selectively withdraw bid orders in that region. In perpetual futures, therefore, asymmetry is not merely a reflection of directional pressure but an early-warning signal of the market's endogenous fragility.

8.4.2 Slope and convexity/concavity

Beyond symmetry, the slope and the convexity/concavity of the order book are two further key dimensions for describing its topological features. Slope is defined as the rate at which resting depth decays as distance from the mid price increases. It reflects the distribution of liquidity along the price axis and directly determines the market's capacity to absorb the impact of large orders.

A gentle slope means that liquidity is evenly distributed across a relatively wide price range. In this configuration, even a relatively large market order produces only relatively mild price slippage, because the subsequent depth can quickly take over and absorb the impact. Conversely, a steep slope indicates that liquidity is highly concentrated near the mid price. Once the price moves beyond this narrow high-liquidity band, the market rapidly enters a thin liquidity vacuum. In an order book with a steep slope, the market's impact-absorption capacity is significantly reduced, and even a tiny order imbalance can trigger a violent price jump.

Convexity/concavity, in turn, is the rate of change of the slope; it reveals the second-

derivative feature of order book depth. A normal order book typically exhibits a "convex" shape — that is, depth decays at a decreasing rate as the price deviates from the mid price, with the decline in depth becoming gentler the farther out one goes. When the order book exhibits a "concave" feature, however — that is, depth drops sharply at a specific price level and then recovers only slowly in a farther price range — this usually means that some "structural obstacle" exists in the market microstructure.

In the context of perpetual futures, this structural obstacle that gives the order book a concave shape is, in the vast majority of cases, the dense liquidation-price zone. Market makers' strategic retreat carves a distinct "pit" into the depth curve. Anomalous changes in convexity/concavity — especially the emergence of a concave feature — are therefore an important topological indicator for identifying the accumulation of liquidation risk and assessing the market's extreme fragility.

8.4.3 Order book morphology

By combining topological features such as symmetry, slope, and convexity/concavity, we can classify the complex shapes of the order book into five typical morphologies. These five morphologies form a continuous spectrum reflecting the market's different phases, from extreme calm to extreme panic.

At one end of the spectrum is the "thick book" morphology, the typical representation of the market in its healthiest state: order book depth is ample, the two sides are highly symmetric, the slope is gentle and exhibits normal convexity, the market has an extremely strong impact-absorption capacity, volatility is at a low level, and market makers' willingness to supply liquidity is abundant.

When market activity declines, the order book often evolves into the "thin book" morphology, in which overall depth is limited but symmetry is still fairly well preserved. It typically appears during low-activity periods such as weekends, holidays, or the late night of a particular time zone; although depth is shallow, the absence of pronounced one-directional pressure keeps the market in a relatively balanced state.

Further deterioration presents as the "skewed book" morphology: when the market develops directional pressure or when market makers engage in one-sided risk avoidance, the order book exhibits pronounced asymmetry — deep on one side, shallow on the other. This is common in trending markets or when the price gradually approaches an important support or resistance level.

More extreme still is the "cliff book" morphology unique to perpetual futures markets: below (or above) a specific price level, depth drops abruptly, forming a distinct "cliff." This is direct microstructural evidence of the existence of a dense liquidation zone, marking that leverage accumulation has reached a statistically high level and that market makers are engaged in a defensive withdrawal of liquidity.

At the other end of the spectrum is the "vacuum book" morphology, the hallmark of severe liquidity exhaustion: bid- or ask-side depth has fallen to nearly zero, and market makers have withdrawn entirely from the order book. It typically appears at the peak moment of an extreme crisis event, such as when a liquidation cascade is in full swing, at which point the market has lost any meaningful pricing capacity.

Figure 8-8 arranges the five morphologies along a continuous spectrum from maximum resilience to complete fragility. This arrangement reveals a structural regularity: the progression from "thick book" to "vacuum book" is not a random jump but an ordered degradation — depth contracts, symmetry breaks, and the slope shifts from gentle to steep until it ruptures — with each phase containing the precursor signals of deterioration into the next.

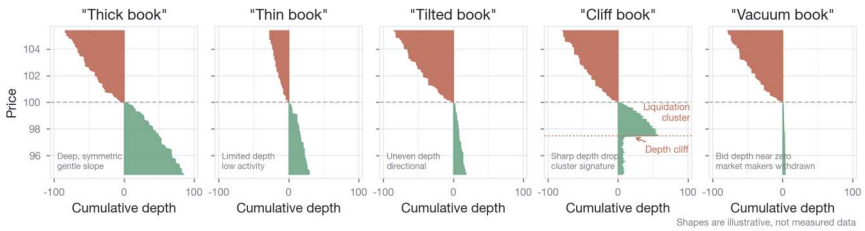


Figure 8-8. A visualization of the five order book morphologies (conceptual illustration of the five order book morphologies; synthetic depth profiles, not empirical data)

From left to right, the figure shows the ordered degradation of the order book from "thick book" to "vacuum book"; the cliff book and vacuum book are morphologies unique to perpetual futures under extreme stress and almost never appear in traditional spot markets. Table 8-3 further summarizes, in a feature matrix, the differences among the morphologies along three dimensions — topological features, market implications, and typical occurrence scenarios — providing traders and risk managers with a rapid classification and diagnostic framework based on topological features.

Morphology	Topological features	Market implications	Typical occurrence scenario
"Thick book"	Ample depth, symmetric on both sides, gentle slope	Market healthy, strong impact-absorption capacity	Calm periods, low volatility, ample market-maker confidence
"Thin book"	Limited depth, but still symmetric	Low market activity, no pronounced directional pressure	Weekends, holidays, or liquidity-trough periods
"Skewed book"	Deep on one side, shallow on the other, pronounced asymmetry	Presence of directional pressure or one-sided market-maker risk avoidance	Trending markets, or when the price approaches an important support/resistance level
"Cliff book"	Depth drops abruptly near a certain price level, exhibiting concavity	Unique to perpetual futures: dense liquidation zone, defensive market-maker retreat	Leverage-accumulation periods, with the OI-to-market-cap ratio markedly elevated
"Vacuum book"	Depth on one side (usually the bid side) is nearly zero	Liquidity collapse, complete market-maker retreat, liquidation cascade underway	The peak moment of an extreme crisis event (such as a flash crash)

Table 8-3. The feature matrix of the five order book morphologies (Data source: compiled by the author)

8.4.4 The mapping between morphology and market state

The evolution of order book morphology is not random; it maps strictly onto the deterioration of the macro market state. The transition from "healthy" to "collapsed" exhibits complex dynamics in which continuous deterioration and nonlinear jumps coexist. During the evolution from "healthy" to "tense" to "fragile," the changes in the order book's topological structure are indeed traceable and approximately continuous: the gradual widening of the spread, the progressive contraction of one-sided depth, and the steady climb of the cancellation rate all constitute traceable precursor warning signals. The transition from "fragile" to "collapsed," however, may exhibit a phase-transition-like nonlinear jump. Sornette's (2003) research on critical phenomena in financial markets shows that systemic collapse often manifests as "quasi-continuous" micro-deterioration that, upon reaching a critical threshold, suddenly triggers a violent qualitative change at the macro level [20]. The empirical observation that "70% of the damage was concentrated within 40 minutes" in the October 2025 event is precisely a textbook feature of a critical transition: the system had already accumulated sufficient fragility at the micro level, but the macro-level collapse, once triggered, erupted in a discontinuous manner, with no gradual transitional phase available for intervention in between.

The starting point of this continuous deterioration is the transition from healthy to tense, manifested as the evolution from "thick book" to "skewed book." When the market receives an initial shock or a directional expectation begins to form, the spread starts to widen, one-sided depth begins to contract, the order book evolves from a symmetric "thick book" into an asymmetric "skewed book," and the market-maker cancellation rate begins to rise, indicating that their sensitivity to risk is increasing. As leverage accumulates further and the price moves in an adverse direction, the market slides from tense into fragile: market makers foresee liquidation risk and begin to actively withdraw liquidity near the dense liquidation zone, and the "cliff book" morphology appears on the order book. This is the key marker that the market has entered a highly fragile state, indicating that even a tiny shock could trigger a chain reaction. Once the price breaks through the residual depth above the "cliff" and touches the liquidation trigger line, the market plunges from fragile into the collapse phase: a large volume of forced market orders floods in, market makers instantly cancel all remaining quotes to control their own risk exposure, bid-side depth is entirely consumed, the order book evolves into a "vacuum book," and the positive feedback loop of the liquidation cascade unfolds in full. When all the liquidation momentum has been released, the market enters the long process from collapse to recovery: the price stabilizes at the bottom, market makers re-enter the market with smaller order sizes and wider quote intervals, and the order book gradually recovers to a "thin book" morphology with limited depth. Only with the passage of time and the gradual rebuilding of confidence can depth slowly return to the "thick book" state. It should be noted that the speed of this recovery is often far slower than the speed of the collapse, reflecting the asymmetry of perpetual futures market resilience.

Figure 8-9 uses a state-transition diagram to map the characteristic microstructural signal corresponding to each phase transition (each transition signal is labeled in the figure). Particular attention should be paid to the pronounced asymmetry of the recovery path: the return from collapse to health is far longer than the deterioration path and must pass through a gradual rebuilding of "vacuum book → thin book → thick book," reflecting the time cost of market makers' capital redeployment, risk-model recalibration, and gradual restoration of confidence after an extreme event.

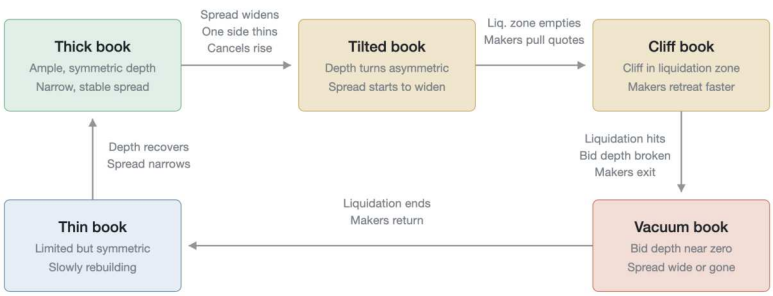


Figure 8-9. The transition path between order book morphology and market state (conceptual illustration of the morphology-to-market-state mapping path, not empirical data)

In theory, an analyst who continuously monitors the evolution of the order book's topological structure can identify the early indicators of rising market fragility before a collapse occurs. This morphology-based diagnostic capability constitutes the core foundation for the discussion of volatility clustering and the construction of a liquidity early-warning system in later chapters.

8.5 The four fingerprints of the perpetual futures order book

The underlying financial rules of perpetual futures (in particular high leverage, forced liquidation, the funding rate mechanism, and the around-the-clock, uninterrupted trading environment) significantly alter their market microstructure. These distinctive institutional features produce observable structural characteristics on the order book, which we call the "four fingerprints." These microstructural patterns simply do not exist in traditional spot markets or in futures markets with an expiry date. The four fingerprints constitute a microstructural diagnostic framework designed specifically for perpetual futures.

8.5.1 The depth cliff

In perpetual futures markets, because of the widespread use of high leverage, traders' liquidation trigger prices often cluster heavily within certain specific price ranges, forming what is called a "dense liquidation-price zone." When the price approaches these

zones, a distinct topological feature appears on the order book: the depth cliff. Specifically, bid-side (or ask-side) depth near that price level becomes abnormally thin, and the depth curve exhibits a cliff-like vertical drop.

The cause of the depth cliff is market makers' defensive order cancellation: high-frequency market makers can identify the location of dense liquidation zones with considerable precision and foresee that, once the price touches them, the liquidation engine will throw out the toxic liquidation orders described in Section 8.3.1 (involuntary, entirely price-insensitive, far larger than normal orders). Unwilling to serve as the counterparty here, they sharply reduce their resting orders in that region in advance, or push their quotes to deeper and wider levels, artificially creating a liquidity vacuum. This complete causal chain of "identify the dense liquidation zone → defensive cancellation → liquidity vacuum" was developed in Section 8.2.4.

Figure 8-10 visually shows the formation mechanism of the depth cliff and its spatial overlay with liquidation-price density.

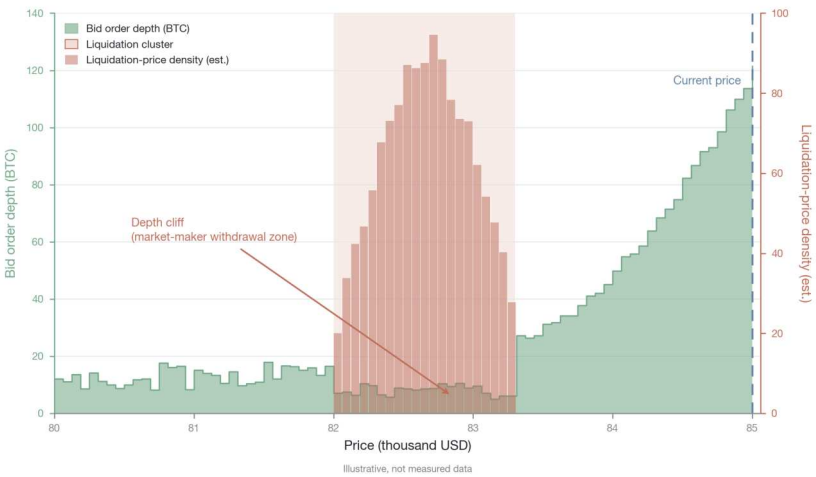


Figure 8-10. The depth cliff at the dense liquidation-price zone (conceptual illustration of the depth cliff; both bid-side depth and liquidation-price density are illustrative, not empirical data)

The existence of the depth cliff gives the dense liquidation zone a dangerous game-theoretic property: it becomes a positive-feedback region with a self-fulfilling tendency. Once the price falls into the cliff region, the lack of sufficient bid support means that even minimal ordinary selling pressure is enough to consume the remaining thin depth and push the price directly to the liquidation trigger line. The triggered liquidation orders then execute continuously in the low-liquidity range that lacks counterparty quotes, causing a violent price plunge that in turn triggers deeper liquidations. This is precisely the core transmission mechanism of the liquidation cascade at the level of the micro order book. In an on-chain transparent environment, this mechanism may even be maliciously exploited and evolve into a strategic liquidity attack.

8.5.2 The depth tide

Perpetual futures anchor the contract price to the spot index price through the funding rate mechanism. This mechanism requires longs and shorts to exchange funds at specific settlement moments (typically every 8 hours, or every hour on some high-frequency platforms). This cyclical institutional arrangement gives rise to the second fingerprint on the order book: the depth tide.

The depth tide manifests as cyclical, predictable fluctuations in the order book's depth, spread, and bid-ask asymmetry around the funding-settlement moment. According to the empirical study by Ruan and Streltsov (2022), this tidal phenomenon is especially pronounced when the expected funding rate is positive (that is, longs must pay shorts) and high [15].

In the several hours approaching settlement, the market experiences an "ebb tide" period. To avoid paying a high funding rate, cost-sensitive long traders choose to close positions briefly before settlement. This significantly increases selling pressure, and the order book develops a pronounced sell-side skew. Funding-rate arbitrageurs, meanwhile, establish short positions to collect the rate, further intensifying sell-side pressure. Faced with this one-directional, structural order-flow imbalance, market makers — to control inventory risk — actively widen the bid-ask spread and reduce resting depth.

After the settlement moment passes, the market enters a "flood tide" period. Longs that had avoided the rate re-establish their positions, arbitrageurs may close and exit, and directional pressure quickly recedes. Market makers accordingly narrow the spread and restore resting orders, and the order book's depth and symmetry return to normal levels.

The morphology and intensity of the depth tide depend heavily on the design of the funding-rate settlement frequency. The traditional 8-hour settlement cycle (such as Binance's UTC 00:00/08:00/16:00) produces the clearest tidal pattern, because the rate accumulates a large enough payable amount over the longer accumulation period that the directional order flow around settlement is sufficient to leave a discernible structural feature on the order book. As exchanges compete to introduce higher-frequency settlement mechanisms (Binance has implemented 4-hour settlement for some trading pairs, and Hyperliquid uses hourly settlement), however, the payable amount at each settlement shrinks correspondingly, and the amplitude of the tide weakens with it. If the settlement frequency were raised further to a near-real-time variable-rate system, the directional impact of a single settlement might fall below the level of market noise, at which point the diagnostic value of the depth-tide fingerprint would be significantly diminished. In addition, in an extreme funding-rate environment (an annualized rate exceeding 100%), the normal tidal pattern may fail entirely: the scale of directional position closing before settlement may exceed market makers' normal absorption capacity, causing the tide to shift from a regular ebb-and-flow into a one-directional liquidity shock.

Figure 8-11 simulates and displays this cyclical tidal process based on the characteristics of empirical data.

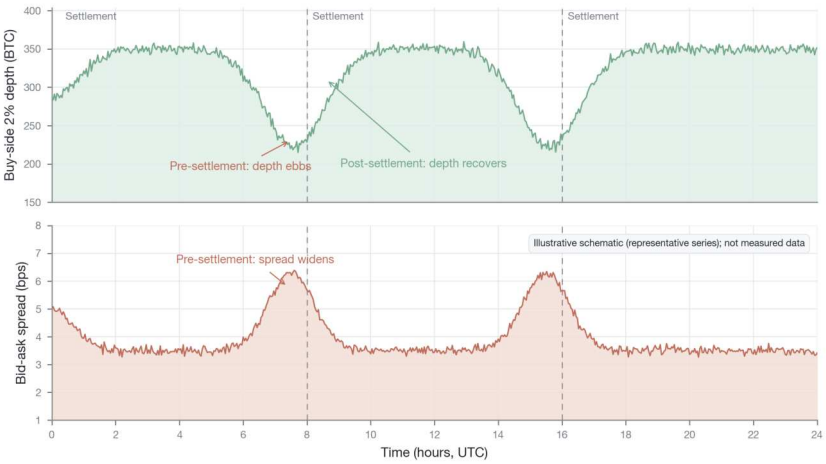


Figure 8-11. The depth tide around funding-rate settlement

Note: This figure is a theoretical illustration drawn from the characteristic parameters of the empirical data in Ruan and Streltsov (2022) [15] (a synthetic illustration of the depth tide, not empirical data). The depth tide is a microstructural feature unique to perpetual futures. It shows that even in the absence of an external macro information shock, internal institutional friction alone is enough to cause a cyclical contraction of market liquidity. For an ordinary trader, executing a large order during the ebb-tide period entails significantly higher implicit transaction costs (slippage).

8.5.3 The liquidity circadian rhythm

The cryptocurrency perpetual futures market is the world's first trillion-dollar financial market to achieve genuine around-the-clock, uninterrupted operation. The continuity of trading hours, however, does not equate to uniformity in the distribution of liquidity. On the contrary, the order book exhibits a third fingerprint highly correlated with the time zones of the world's major financial centers: the liquidity circadian rhythm.

Structurally, this rhythm exhibits a pronounced alternation of peaks and troughs. The absolute peak of liquidity typically appears during the few hours in which the European and American trading sessions overlap (approximately UTC 13:00–17:00). At this time, the world's most important institutional investors, high-frequency market makers, and active traders are online simultaneously, order book depth reaches its maximum, the spread compresses to an extremely narrow level, and the market's resilience in absorbing shocks is strongest.

By contrast, the trough of liquidity appears during the "transition period" between the close of the American market and the full activation of the Asian market (approximately UTC 00:00–05:00), as well as during traditional weekend periods. In these trough periods, the algorithms of many Western institutional market makers may enter a low-risk operating mode or reduce their quoting frequency, causing order book depth to fall significantly below peak-period levels and accompanied by a widening of the spread.

Figure 8-12, by superimposing the activity of different time zones, clearly delineates

this 24/7 liquidity rhythm.

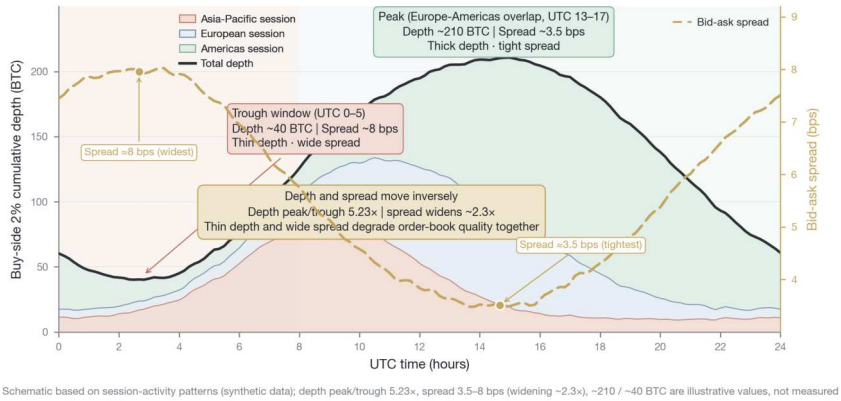


Figure 8-12. The 24/7 liquidity circadian rhythm (a stylized illustration of the 24/7 liquidity circadian rhythm: the peak of about 210 BTC, the trough of about 40 BTC, and the peak-to-trough ratio of 5.23x are synthetic representative values, not empirical data)

The existence of this rhythm reveals a serious endogenous risk of the perpetual futures market. In traditional financial markets, the market-close mechanism functionally serves to interrupt the transmission of negative feedback. The perpetual futures market lacks this interruption mechanism; in its place are cyclical liquidity-trough periods. The problem is that a trough not only fails to cushion shocks but actually becomes a window in which they amplify more easily. If a sudden piece of bad news is encountered during a liquidity-trough period, or a liquidation cascade is triggered, its destructive power is amplified exponentially, because there is simply not enough depth at that time to absorb the selling pressure.

8.5.4 The asymmetric response of the spread

In traditional low-leverage spot markets, the bid-ask spread's response to market volatility is typically approximately symmetric. That is, whether volatility is caused by a price rise or a price fall, market makers synchronously widen the spread and reduce depth on both the bid and ask sides to compensate for the increased adverse-selection risk. In the high-leverage environment of perpetual futures, however, the order book exhibits a fourth fingerprint: the asymmetric volatility response of the spread.

Specifically, during a volatility spike triggered by a sharp price decline, the widening of the order book's bid-side spread is significantly greater than the widening of the ask-side spread during volatility triggered by a sharp price rise. This asymmetry is deeply rooted in the crypto market's systematic long bias and in the asymmetry of the liquidation mechanism.

When the price falls, a large volume of high-leverage long positions faces liquidation. As noted earlier, this produces a massive volume of price-insensitive forced sell orders. Foreseeing this impending large-scale toxic selling pressure, market makers rapidly cancel their bid quotes, causing bid-side depth to plunge to near zero and the bid-side

spread to widen nonlinearly.

By comparison, although a price rise also triggers short liquidations and generates forced buy orders, in the vast majority of periods the scale of longs in the crypto market far exceeds that of shorts. The forced-buy impact during a rise is therefore typically smaller in magnitude than the forced-sell impact during a fall. The magnitude and speed of market makers' retreat on the ask side are correspondingly smaller, and the widening of the ask-side spread is relatively mild.

Figure 8-13, through a scatter plot and a fitted curve, visually shows this asymmetric response pattern.

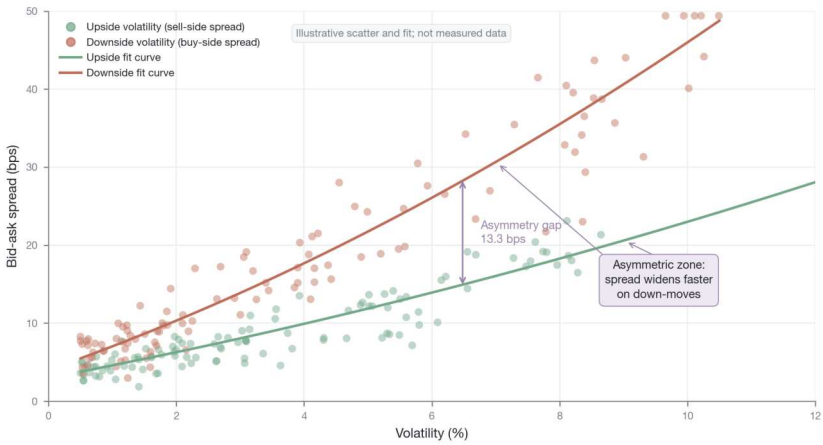


Figure 8-13. The asymmetric volatility response of the spread (a synthetic scatter illustration of the asymmetric volatility response of the spread: the 13.3 basis points is a representative asymmetry gap, not empirical data)

The scatter in the figure reveals this structural skew: the fitted slope of the bid-side spread in the negative-return (decline) range is significantly greater than that of the ask-side spread in the positive-return (rise) range, meaning that under the same magnitude of volatility shock, a decline systematically damages bid-side liquidity more than a rise damages ask-side liquidity. This is precisely the microstructural root of the perpetual futures adage that "declines tend to be faster than rises" — the forced liquidation selling pressure, the rapid retreat of market makers, and the sharp drop in depth occur simultaneously and reinforce one another, forming a high-intensity positive feedback loop.

8.5.5 The superposition effect of the four fingerprints

The four fingerprints described above (the depth cliff, the depth tide, the liquidity rhythm, and the asymmetric response) do not exist in isolation. In specific market environments, they undergo a high-risk resonance and superposition, producing a market-shock magnitude far exceeding that of any single factor. Understanding this superposition effect is key to building an advanced market-risk early-warning system.

In the most dangerous superposition scenario, the market has undergone a prolonged

one-directional rise and has accumulated a large volume of high-leverage long positions, and the funding rate remains positive and high (Fingerprint 2: longs are unwilling to close positions, and leverage is locked in). At this point, a dense liquidation-trigger zone has already formed just below the current price (Fingerprint 1: a depth cliff is emerging). If this happens to coincide with a liquidity-trough period on a weekend or in the early Asian morning (Fingerprint 3: depth is at its weakest overall), then once a tiny piece of negative news causes the price to dip slightly, it can easily break through the already-thin bid-side depth. After touching the liquidation line, the enormous selling pressure from long liquidations causes market makers to retreat sharply on one side (Fingerprint 4: the bid-side spread asymmetry widens to an extreme). In that instant, all four fingerprints point simultaneously to "liquidity collapse," and the probability of a flash crash rises sharply.

To monitor this superposition risk systematically, we can construct a "comprehensive perpetual futures order book risk dashboard" (Figure 8-14). This dashboard tracks the state indicators of the four fingerprints in real time, and when multiple indicators simultaneously break through their warning — or even danger — thresholds, it issues the highest-level systemic-risk alert.

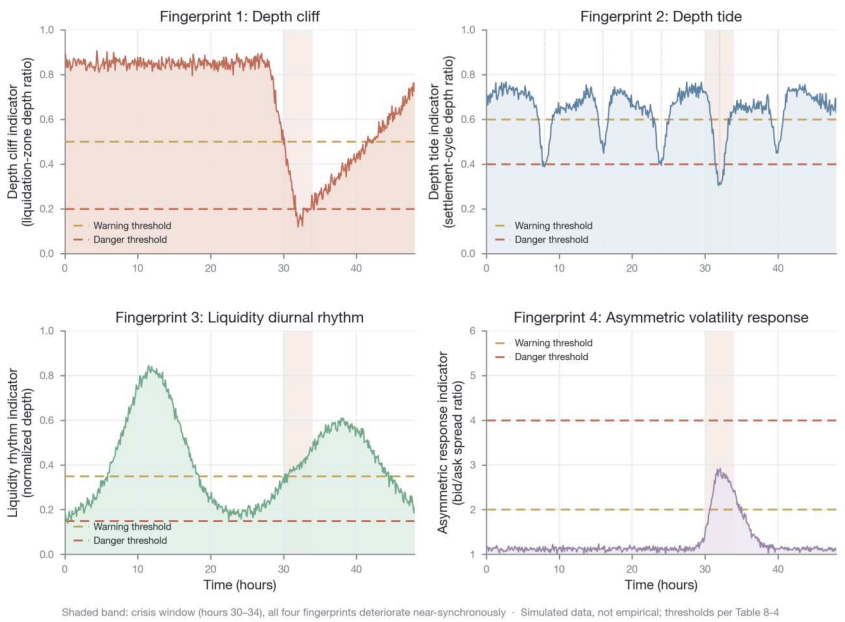


Figure 8-14. The comprehensive risk dashboard of the four fingerprints

Note: This figure is drawn from simulated data based on the characteristic parameters of historical events (the thresholds of the four fingerprints are the author's ratings, not empirical data). During the crisis period (the shaded region), the indicators of the four fingerprint dimensions deteriorate almost synchronously within an extremely short time window. This synchronized deterioration pattern is the core criterion for distinguishing a systemic-risk event from normal volatility. The diagnostic value of the dashboard lies in capturing the synchronized-deterioration signal across multiple dimensions, rather than monitoring the absolute level of any single indicator.

Table 8-4 lists the parameters of this comprehensive diagnostic matrix. The thresholds for each fingerprint dimension are calibrated on the basis of backtesting analysis of multiple historical extreme events, including the October 2025 flash crash (the thresholds are calibrated through historical backtesting; specifically, the distribution of each indicator across multiple extreme events is analyzed statistically, and the corresponding quantiles are selected as the grading boundaries).

Fingerprint dimension	Core observation indicator	Normal-state benchmark	Warning threshold (entering the tense period)	Danger threshold (approaching the collapse point)
Fingerprint 1: depth cliff	Bid-side depth in the dense liquidation zone / average depth in the normal range	$> 80\%$	$< 50\%$	$< 20\%$ (market makers have clearly retreated)
Fingerprint 2: depth tide	Depth 1 hour before settlement / average depth within the cycle	~ 90%	$< 60\%$	$< 40\%$ (accompanied by an extremely high funding rate)
Fingerprint 3: liquidity rhythm	Total depth in the current period / 24-hour peak depth	Naturally fluctuates by period (typically $> 50\%$)	$< 35\%$	$< 15\%$ (period of extreme liquidity exhaustion)
Fingerprint 4: asymmetric response	Bid-side spread widening rate during a decline / ask-side spread widening rate during a rise	~ 1.0–1.2 (slight long bias)	> 2.0	> 4.0 (one-sided liquidity is being drained)

Table 8-4. The comprehensive diagnostic matrix of the four fingerprints (Data source: compiled by the author)

The threshold settings in the diagnostic matrix above face several applicability limitations, and users should regard them as a methodological framework rather than fixed point estimates. First, the validity of the thresholds depends on a specific volatility regime: an indicator reading flagged as "dangerous" in a low-volatility environment may fall within the normal fluctuation range under a high-volatility regime, and vice versa. Fixed thresholds therefore run the risk of being invalidated by a switch in volatility regime; a more robust approach is to use conditional thresholds — that is, to replace absolute levels with the distributional quantiles of the indicator within a rolling window, so that the thresholds adapt to the market state. Second, the thresholds for Fingerprint 4 (the asymmetric spread response) — normal at about 1.0–1.2, warning at $>$ 2.0, danger at > 4.0 — are at present inferred mainly from the mechanistic narrative of the "systematic long bias" and still lack independent quantitative empirical anchoring; they should be regarded as illustrative. Moreover, all of the above thresholds are calibrated on BTC historical data as a whole, and their extension to altcoins and long-tail assets requires recalibration: the normal-state liquidity level, volatility structure, and participant composition of these assets differ systematically from those of BTC, and

applying the BTC thresholds directly may produce a large number of false positives or false negatives. The core value of this matrix lies in providing a multidimensional joint-monitoring diagnostic methodology, not a set of static numbers that can be applied mechanically.

In sum, the four fingerprints of the perpetual futures order book are both the microscopic projection of its distinctive institutional design and microstructural diagnostic tools in their own right. By integrating the signals of these four dimensions, we can move beyond surface-level price fluctuations and systematically identify how liquidity evolves from ample to fragile to collapsed.

8.6 Order book resilience

Operationalizing "resilience" from a vague qualitative concept into a measurable micro indicator is a key step in understanding the dynamic properties of the order book. The preceding sections analyzed the static structure of the order book (depth, slope, morphology) and the four fingerprints of perpetual futures, but all of these analyses are cross-sectional descriptions of the order book state "at a single moment." Introducing the time dimension, we will argue that the resilience of the perpetual futures order book is systematically weaker than that of the spot market — not a technical defect but an intrinsic cost that necessarily accompanies the institutional combination of "high leverage + liquidation mechanism + no expiry date + 24/7 trading." The positive-feedback mechanism of the liquidation cascade makes shocks not only larger in scale but also self-amplifying, thereby significantly lengthening the recovery time. The measurement of resilience provides a key micro foundation for the market-quality assessment in Chapter 25. In the October 2025 flash crash, the market experienced the extreme asymmetry of "collapse in about 2 hours, recovery over more than 24 hours," which is a real-world portrayal of the resilience characteristics of perpetual futures [2].

8.6.1 The operational definition and measurement of resilience

In the market microstructure literature, liquidity is typically decomposed into three dimensions: tightness (the size of the spread), depth (the order size the market can absorb without causing a significant price move), and resilience (the speed of recovery to a normal state after a shock). The first two dimensions are static descriptions of the order book at a single moment, whereas resilience characterizes the order book's dynamic adjustment capacity. In his seminal paper, Large (2007) explicitly proposed that an electronic limit order book possesses resilience when it can rapidly restore its normal shape after a large trade [21]. This definition transforms resilience from a vague intuition into an operational concept that can be empirically tested with time-series data.

We operationally define order book resilience as follows:

> the time required for the order book's spread and depth to recover to their pre-shock levels after it is subjected to a standardized shock.

The specific measurement method comprises the following steps. First, select a "standardized shock" as the starting point of the measurement. A standardized shock refers to a market order that consumes a certain proportion (for example, 50% or more) of the depth near the current best bid and ask, or an observable price jump driven by an exogenous event. Record the instant the shock occurs, t_0, at which the spread widens from its pre-shock steady-state level S_0 to S_1, while cumulative depth falls from D_0 to D_1. Subsequently, track the recovery process of these two core state variables. Record the time required for the spread to recover to the pre-and-post-shock mean $0.5 \times (S_0 + S_1)$, denoted τ_{spread}; record the time required for depth to recover to $0.5 \times (D_0 + D_1)$, denoted τ_{depth} [22]. The resilience indicator is a function of these two time variables. The smaller the value of τ, the faster the market recovers from the liquidity vacuum, and the stronger the resilience of the order book.

A more precise operational definition of "recovery" itself is needed. The framework above uses the midpoint of the pre-and-post-shock mean (that is, a 50% recovery rate relative to the pre-shock level) as the recovery threshold, but different analytical purposes may require different recovery criteria: the τ values corresponding to a 75% recovery rate and a 90% recovery rate may differ by several fold. In addition, the recovery speed differs significantly across price tiers: depth near the best bid and ask typically recovers fastest, because market makers prioritize re-posting at the most profitable best-price levels, whereas the deeper regions of the order book far from the mid price take longer to recover, because the return to providing liquidity at these levels is lower and the adverse-selection risk is higher [9]. A single τ value may therefore mask the significant heterogeneity in recovery speed across price tiers.

Empirical research shows that in traditional markets the different dimensions of the order book deteriorate simultaneously after a shock but recover at significantly different speeds. Wuyts (2008) documented this pattern of synchronous deterioration and uneven recovery [22], and Lo and Hall (2015) further confirmed that the spread typically recovers faster than depth [23]. This finding has an intuitive economic interpretation: after a shock, market makers tend to first provide a small amount of liquidity at the best price to narrow the spread (a "tentative" return that carries little risk), and then gradually replenish the depth at each price tier in response to market feedback (a "committed" return that requires greater capital investment and stronger confidence).

This asymmetric recovery pattern of "spread recovers first, depth recovers later" is further amplified in perpetual futures markets. As shown in Figure 8-15, in a typical shock event the spread and depth of the spot market can both recover to close to pre-shock levels within a relatively short time (on the order of minutes). The perpetual futures market, by contrast, not only experiences a larger initial shock (because of the amplification effect of the liquidation cascade) but also follows a more tortuous recovery path, in which a liquidation-triggered "secondary shock" may appear after the initial shock, repeatedly interrupting the recovery process. More critically, perpetual futures depth often cannot fully recover to its pre-shock level but instead stabilizes at a lower "new baseline," reflecting market makers' permanent reassessment of risk exposure after an extreme event.

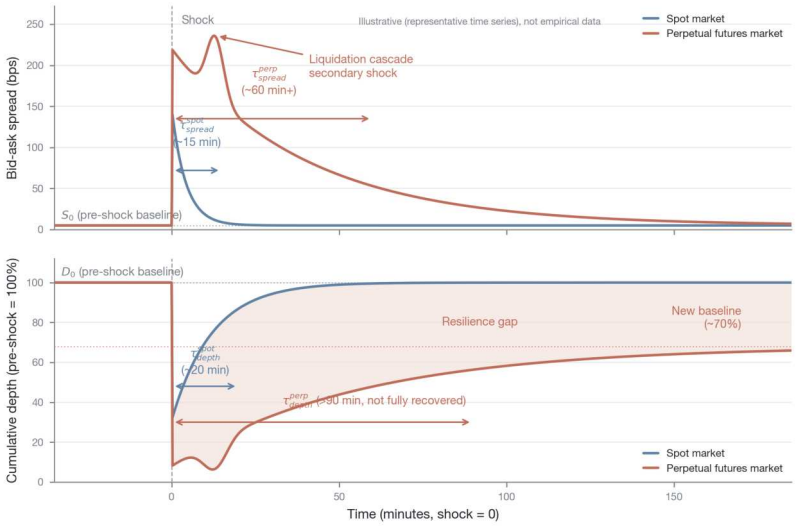


Figure 8-15. The resilience measurement of the shock-recovery time series (a conceptual time-series illustration of order book shock-recovery resilience; the τ, the ~70% new baseline, and similar values are conceptual illustrative values, not empirical data)

The upper and lower panels of the figure track, respectively, the post-shock recovery paths of the spread and of depth. In the spot market (dashed line), both the spread and depth return to close to pre-shock levels within minutes; in perpetual futures (solid line), the initial shock is not only larger ($S_1 - S_0$ and $D_0 - D_1$ are both significantly greater than in spot) but also, in the initial recovery phase, a liquidation cascade produces a "secondary shock" (marked in red in the upper panel as "liquidation-cascade secondary shock"), so that the recovery curve is sawtooth-shaped rather than a smooth monotonic path. More critical is the recovery endpoint: spot depth returns to close to D_0, whereas perpetual depth stabilizes at a significantly lower "new baseline" D_0', with the gap $\Delta D = D_0 - D_0'$ quantifying market makers' permanent reassessment of risk exposure.

8.6.2 The structural disadvantage in resilience

By comparing the recovery paths of perpetual futures and the spot market, one finds that the resilience of perpetual futures is systematically weaker than that of spot. This disadvantage does not stem from a deficiency in technical performance but is caused by three structural reasons derived from its core institutional features.

The foremost structural reason is the positive-feedback mechanism of liquidation. In a spot market, the shock produced by a large sell order is typically a one-time, exogenous event. Once the seller has completed the sale, the selling pressure ends. After absorbing the shock, if market makers assess that no new fundamental information is driving events, they quickly adjust their quotes, and price and depth tend to recover naturally. In perpetual futures markets, however, the situation is entirely different. A large sell order not only produces the initial price shock but is also highly likely to trigger the forced liquidation of leveraged long positions. The liquidation engine throws these positions into the market as market orders, producing a new round of selling pressure. This new selling pressure pushes the price down further, thereby triggering deeper liquida-

tions. The shock is no longer an isolated one-time event but a self-amplifying positive feedback loop. Market makers are keenly aware of the destructive power of this positive feedback loop. After a shock, therefore, a rational market maker will not immediately provide liquidity but must wait until "all liquidations that can be triggered have been completed" and confirm that the selling pressure has been thoroughly exhausted before beginning to slowly restore quotes and depth [2]. This rational behavior of "waiting for the liquidations to end" is precisely what lengthens the duration of the market's low-resilience state.

The flash crash of October 10, 2025, provides extreme empirical evidence of this mechanism. FTI Consulting's post-mortem report noted that on that day the depth near the best bid of BTC perpetual futures contracted by more than 90%, and the bid-ask spread widened from the normal single-digit basis points to double-digit percentages [2]. Two different depth measures must be distinguished: the ">90%" contraction described under "stress depth" in Section 8.2 characterizes the instantaneous stress-depth collapse at the 21:15 UTC cascade peak, whereas the following persistent decline characterizes the residual depth level that had not recovered weeks after the crisis; the two differ in measure and in moment. As for the persistent residual, CoinDesk Research's follow-up tracking data show that the order book depth of BTC within the 1% price range (two-sided) fell from about $20 million before the shock to about $14 million, a decline of nearly one-third, and had still not recovered to pre-shock levels five weeks after the event [3]. Amberdata's analysis in February 2026 went further, noting that BTC's order book depth within the 10-basis-point range plunged about 63%, from a peak of $38 million in September 2025 to $14 million (this 10-basis-point window differs in measure from the 1% window above, and its baseline moment is also the September 2025 peak rather than the eve of the crash, so the two sets of figures cannot be compared directly), and that "the market never fully recovered" [24]. Together, these data indicate that the collapse took only about 2 hours, whereas the recovery (if it can be called a "recovery") remained incomplete months later.

Attributing the persistent depression of depth entirely to "resilience failure," however, may be an oversimplification. Market makers' reduction of quote size and depth after the crisis was partly a rational repricing based on higher volatility expectations: after an extreme event, market makers systematically revise their estimates of future volatility upward, and a wider spread and thinner depth are their optimal response in a higher-risk environment, rather than pure loss of confidence or inadequacy of capacity [12]. In addition, the actual losses suffered during the crisis also consumed market makers' available capital, limiting their capacity to restore their market-making scale. The post-crisis "new baseline" of depth therefore contains both a structural resilience defect (that is, the system lacks a mechanism at the institutional level to promote rapid recovery) and market makers' reasonable behavioral adjustment based on a new information set. Distinguishing these two components is of significant analytical value for assessing whether the market is truly "damaged" or merely "repriced."

At the same time, the lowering of market makers' retreat threshold further weakens resilience. As discussed in Section 8.2.2, perpetual futures abolish the expiry date, causing market makers to bear open-ended inventory risk and depriving them of the definite risk-termination point and natural exit of the traditional futures "hold to delivery."

This open-ended risk exposure significantly lowers market makers' retreat threshold when market uncertainty rises: at the earliest sign of an anomaly, they are more inclined to cancel orders immediately to protect capital rather than "wait a little longer," which weakens their willingness to "stay in the market and provide resilience" and causes the order book to enter a "vacuum" state faster after a shock. CoinDesk Research described this phenomenon as "a prudential contraction of market-making commitment" and regarded it as a "structural shift" rather than a temporary panic reaction [3].

Building on this, the absence of a "reset window" caused by the 24/7 operating model constitutes another important factor. Traditional financial markets have fixed trading hours, and the closed period provides the market with a natural "resilience-recovery window": market makers can reassess their risk exposure, adjust pricing models, replenish margin, and re-enter the next day in a fresher state and with more ample capital. This window also provides three often-overlooked functions — time to digest information (rationally assessing new information during non-trading hours and avoiding overreaction under immediate emotion), cross-time-zone coordination (participants in different time zones fully pricing in the same event during their respective active periods), and a definite point for clearing and settlement (providing an institutional deadline for margin calls and risk-position adjustments). The around-the-clock operation of perpetual futures means all three functions are simultaneously absent, and market makers must make "stay or leave" decisions in real time under a continuous flow of information and pressure, so that the system's continuous fatigue further weakens its recovery capacity after an extreme event. Combined with the "liquidity circadian rhythm" analyzed in Section 8.5.3, if a shock happens to coincide with a liquidity-trough period (such as between the American close and the Asian open), the absence of market-making force worsens resilience further. The October 2025 event occurred at 20:50 UTC, at the tail end of the American trading session, when the activity of European and American market makers was declining and Asia-Pacific had not yet fully started — not the absolute trough of the 24-hour cycle, but already on the downward slope of the transition from peak to trough [2]. This coincidence of the liquidity level with the moment of crisis may have had a strategic component, and it also amplified the system's vulnerability to shocks during a relatively weak period.

Figure 8-16, in the form of a causal diagram, provides a systematic visualization of the three structural reasons above and the specific paths by which they transmit to order book resilience.

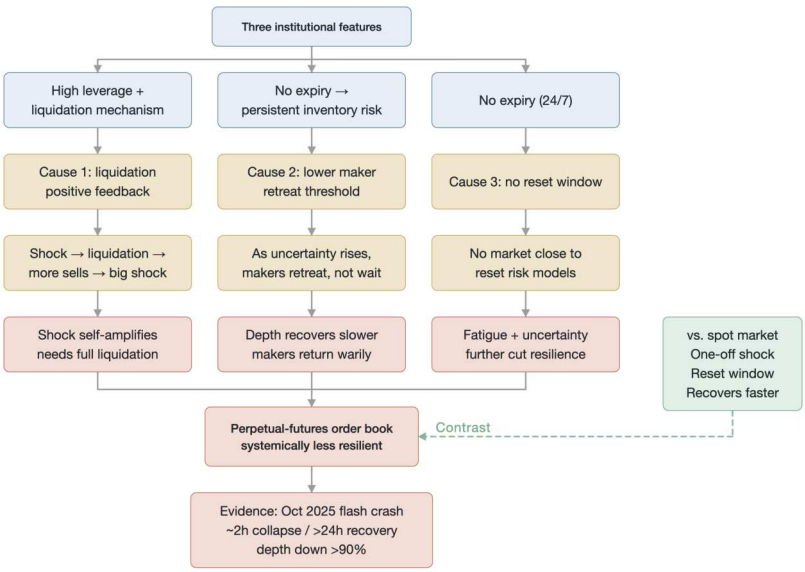


Figure 8-16. The causal mechanism of the resilience disadvantage of perpetual futures (an illustration of the three-cause causal mechanism; at the bottom, "October 2025 flash crash depth contraction >90%" is hard data, and "collapse in about 2 hours / recovery over more than 24 hours" is a representative time scale; following CoinDesk, November 15, 2025 [3])

The causal diagram reveals the synergy of three paths: the liquidation positive feedback produces immediate, self-amplifying destructive power during the shock; the low retreat threshold weakens the order book's defensive depth even before the shock; and the absence of a 24/7 reset window eliminates the institutional opportunity for recovery after the shock. The simultaneous presence of all three means that the perpetual market is already in a resilience-weakened state before the shock, suffers self-amplifying destruction during the shock, and lacks a recovery mechanism after the shock. This triple superposition, which spot markets do not possess, is precisely the structural root of why perpetual resilience is systematically inferior to that of spot (this tendency is inferred mainly from the institutional causal chain above and is corroborated by a small number of events such as October 2025; its cross-event generality still awaits testing against more extreme events).

8.6.3 A comparison of resilience across architectures

Beyond the difference in asset type (perpetual futures vs. spot), the market structure in which the matching engine resides also strongly affects the order book's resilience. Under different technical architectures, the measurement of resilience exhibits a trade-off between "verifiability" and "recovery speed."

On centralized exchanges, the order book and matching engine run on private servers, and market makers, using extremely low network latency and zero-gas-cost quote updates, can complete quote cancellations and resets at the millisecond level. On the surface, CEX resilience therefore "looks" strongest: the spread after a shock can often narrow within an extremely short time, and depth too seems to recover quickly.

This resilience, however, is unverifiable. An external observer cannot know for certain whether this rapidly recovered depth contains genuine trading intent or is merely "ghost quotes" posted to obtain market-maker rebates or maintain market rankings. The problems of "flickering liquidity" and "conditional depth" analyzed in Section 8.2 are especially prominent in the assessment of CEX resilience. A CEX order book that rapidly "recovers" its nominal depth after a shock may still have extremely weak stress depth.

By comparison, a central limit order book on an application chain is subject to the physical constraint of block time. For example, on an application chain with a block time of about 200 milliseconds to 1 second, there is a hard upper limit on the frequency of market-maker quote updates, which makes its surface recovery speed inferior to that of a CEX. But every order submission, modification, and cancellation on an on-chain CLOB is recorded in a distributed ledger, and its recovery process is fully transparent and auditable: every unit of depth that is rebuilt locks in real margin, rather than a "ghost" that can be canceled instantly at zero cost (this "capital authenticity" mechanism and its contrast with CEX ghost liquidity are detailed in Section 8.9.2). The resilience of an on-chain CLOB, therefore, though inferior to that of a CEX in speed, may be more reliable in "quality."

As for DEXs deployed on general-purpose public chains (such as the Ethereum mainnet), because the block-confirmation time is long (about 12 seconds) and they face intense gas competition, their order book resilience is the weakest and their recovery speed the slowest. When updating quotes, market makers face high on-chain costs and enormous adverse-selection risk (in the several seconds of waiting for block confirmation, the market price may already have moved substantially), making depth recovery after a shock exceedingly difficult. This architecture, however, offers the highest degree of transparency and censorship resistance.

This difference reveals a core insight: in assessing market resilience, a trader must choose between "surface-level rapid recovery" and "verifiable genuine recovery." This is precisely the concrete manifestation, at the level of order book microstructure, of the "impossible triangle" in Chapter 5 and the "risk visibility-manageability proposition" in Chapter 6. A high-frequency trader who pursues execution speed may place greater value on the surface resilience of a CEX, whereas an institutional investor concerned with counterparty risk and market fairness may place greater trust in the verifiable resilience of an on-chain CLOB.

8.6.4 Resilience and market quality

Resilience is not merely a technical micro indicator but a core dimension of the five-dimensional market-quality model to be established in Chapter 25. It directly determines traders' behavioral patterns and cost expectations in a given market and has a significant effect on the overall efficiency and stability of the market.

In a market with strong resilience, traders can more boldly use market orders to execute their trading strategies, because they know that even if a large order punches through the current depth and produces a price shock, market makers will quickly fill the liquidity gap and the market price will soon recover to fair value. This expectation

makes transaction costs highly certain, and traders can estimate fairly accurately the total cost of executing a trade (including the spread cost and the market-impact cost), enabling more effective portfolio management and risk control.

Conversely, in a market with weak resilience, traders must adopt a more conservative execution strategy. They need to split a large order into multiple small orders, or use algorithms such as time-weighted average price and volume-weighted average price to disperse the execution impact. Because in such a market "the shock may not recover and may even self-amplify," the positive-feedback mechanism of the liquidation cascade means that a large trade may trigger a price move far exceeding expectations. Transaction costs in this environment are full of uncertainty, and the trader faces not only the known spread cost but also an unpredictable tail-impact risk.

At the macro level, the strength of resilience also affects the market's price discovery efficiency. In a market with strong resilience, the price can quickly return to the fundamentals-driven fair value after a shock, and a temporary supply-demand imbalance does not produce a lasting distortion of the price. In a market with weak resilience, however, a shock may cause the price to deviate from fair value for a long time, or even trigger, through a liquidation cascade, a price collapse unrelated to fundamentals, severely damaging the accuracy of price discovery. The effective operation of the price discovery mechanism to be analyzed in Chapter 13 depends largely on the order book having sufficient resilience to absorb and digest information shocks.

The perpetual futures market's weaker resilience means it faces greater tail risk under extreme events. This is not a system "defect" that can be solved by patching code or adjusting parameters but an intrinsic cost that necessarily accompanies the institutional combination of "high leverage + no expiry date + 24/7 trading." These features are interrelated: high leverage provides traders with capital efficiency but also plants the seeds of a liquidation cascade; the absence of an expiry date eliminates rollover costs but also deprives market makers of a risk-termination point; and 24/7 trading provides around-the-clock market access but also eliminates the buffer window for resilience recovery. The theoretical discussion of "the cost of abolishing the expiry date" in Chapter 2 finds here, through the micro indicator of resilience, a complete and quantifiable expression. Understanding this point has important practical significance for traders in formulating risk-management strategies, for regulators in assessing market stability, and for protocol designers in optimizing liquidation mechanisms.

8.7 The evolution of matching-engine architecture

The preceding analysis shows that the resilience of the order book is structurally constrained by institutional features such as the liquidation mechanism, market-maker behavior, and market continuity. These microstructural manifestations (such as the width of the spread, the thickness of depth, and the strength of resilience), however, are also constrained to a considerable degree by the underlying technical architecture that carries the order book. A matching engine running on a sub-millisecond centralized server and one running on a public chain with second-level block confirmation will produce entirely different microstructural characteristics even when facing exactly the

same order flow. The deployment location and implementation of the matching engine have undergone a clear evolutionary path from fully centralized to application-specific. This path reflects not only the iteration of technology but also decentralized exchanges' ongoing trade-offs across three dimensions: performance (efficiency), decentralization (fairness and security), and sovereignty (protocol autonomy). Understanding this evolutionary thread is a prerequisite for grasping the institutional significance of matching-mechanism design in Section 8.8, and it also lays the groundwork for understanding how Hyperliquid, in Chapter 9, stakes out its distinctive position on this spectrum.

8.7.1 Centralized exchanges

In the early stages of the development of both traditional finance and cryptocurrency, the centralized exchange was the only viable solution for providing high-performance trading. Under this architecture, the maintenance of the order book and the operation of the matching engine are placed entirely within the private server cluster controlled by the exchange. Users deposit their assets into the exchange's custodial wallet, and all order submission, modification, matching, and cancellation are completed within the exchange's internal system.

The core advantage of this design lies in its high performance. Because all operations are conducted in memory and are not subject to the constraints of a blockchain consensus mechanism, a CEX can achieve sub-millisecond (<1 ms) matching latency, with throughput reaching hundreds of thousands of orders per second or even higher. For example, Binance has stated that the regular processing capacity of its futures matching engine can reach 100,000 orders per second, with average latency of about 5 milliseconds [25]. Note that these figures are the official values published for Binance's production environment and reflect performance under normal load. This low-latency, high-throughput environment provides a suitable trading environment for high-frequency traders and market makers, enabling them to update quotes at high frequency and thereby maintain, on the surface, narrower spreads and seemingly ample depth. For the perpetual futures market, which pursues execution speed, the CEX has long been regarded as the only venue capable of meeting the needs of professional traders.

This performance, however, comes with several structural risks. The most prominent is the counterparty risk brought by extreme centralization: users' assets are held in custody by the exchange, and once the exchange faces internal misappropriation, a hacking attack, or bankruptcy (such as the FTX incident of 2022), users' funds face the risk of total loss. This risk is especially acute in the perpetual futures market, because high-leverage trading requires users to deposit large amounts of margin at the exchange. At the same time, the unverifiability of the matching process constitutes a major hidden danger: external participants cannot verify the authenticity of order book depth, nor can they know for certain whether the exchange is using its informational advantage to manipulate internally, such as front-running user orders, wash trading to inflate volume, or improper operations during liquidation. In addition, the centralized architecture introduces censorship risk: a centralized entity can unilaterally freeze user accounts, reject specific transactions, or shut down specific markets under regulatory pressure, and users lack an effective remedy against this.

8.7.2 General-purpose-chain DEXs

To eliminate the centralization risk of the CEX, the blockchain community initially attempted to deploy the order book and matching engine directly on a general-purpose public chain (such as the Ethereum mainnet). The design goal of this approach has a clear decentralization orientation: all orders and the matching process are executed transparently on-chain, users' assets are self-custodied by smart contracts, and anyone can verify the fairness of the matching. This fully on-chain decentralized approach, however, immediately ran into a severe performance bottleneck.

A general-purpose public chain was designed to handle various types of smart-contract transactions, not optimized specifically for high-frequency trading. Ethereum's limited throughput (about 15 TPS before the Merge, improved somewhat afterward but still far from sufficient) and long block-confirmation time (about 13 seconds during the PoW period) cannot meet the performance requirements of real-time order book matching. In this environment, market makers must pay high gas fees each time they update a quote, and in the several seconds of waiting for block confirmation they face significant adverse-selection risk: the market price may have moved substantially before the quote is confirmed, making the market maker's resting order a stale quote easily exploited by arbitrageurs. In addition, because all pending transactions queue in the public mempool waiting to be packed, this architecture is vulnerable to maximal extractable value (MEV) attacks such as front-running and sandwich attacks, further worsening conditions for market makers.

Facing the performance dilemma, many DEXs adopted the compromise of a hybrid architecture. A typical case is dYdX v3, which used StarkEx as its Layer 2 settlement layer. In this model, the maintenance of the order book and the matching engine still run on off-chain centralized servers controlled by the project team; orders submitted by users are first sent to these servers for matching, while the trade results are periodically batched and submitted to the Ethereum mainnet through zero-knowledge proofs for on-chain settlement and verification. This architecture achieves smart-contract self-custody of user assets (users' funds are locked in a smart contract on Ethereum, and even if the off-chain servers go down, users can withdraw funds through an on-chain mechanism), thereby eliminating part of the counterparty risk. But its core matching component is still centralized and opaque, and the project team's servers remain a single point of failure and a potential manipulation node. The hybrid architecture makes a pragmatic compromise between performance and decentralization, but it still cannot fully resolve the contradiction of the "impossible triangle."

8.7.3 The application-chain architecture

As DeFi infrastructure matured, and especially as modular blockchain development frameworks such as the Cosmos SDK became widespread, a new paradigm began to dominate the development of high-performance DEXs: the application-chain architecture. An application chain is an independent blockchain tailor-made for a single application, whose consensus mechanism, network parameters, and state-machine logic can all be deeply optimized for that application's needs. dYdX's migration from Ethereum L2 to the independent dYdX Chain (v4) built on the Cosmos SDK marked the full

establishment of this trend.

The core motivation of the application-chain architecture is to ease the trade-off between performance and decentralization. In the dYdX v4 architecture, the order book is no longer hosted by a single centralized server but is jointly maintained and synchronized in the memory of the network's decentralized validators (about 50 to 60 active validators). When a user submits an order, that order is rapidly propagated to all validator nodes through a peer-to-peer network. Matching is performed by the proposer of each block based on the order book in its memory, and the trade results are then packed into a block and verified and confirmed by all validators through the CometBFT consensus mechanism. The dYdX v4 architecture therefore exhibits the hybrid characteristics of an off-chain order book and on-chain settlement: order matching is completed in the validators' off-chain memory, and only the final trade results and state changes need to pass through consensus confirmation. This design substantially improves performance while preserving decentralization. It does, however, introduce a specific trade-off: validators can observe pending orders before a block is produced, leaving potential room for validator-level MEV extraction (such as front-running based on order information).

Beyond dYdX v4, emerging platforms such as Hyperliquid have further expanded the performance frontier of the application-chain architecture. Hyperliquid uses its self-developed HyperBFT consensus mechanism (inspired by the academic HotStuff BFT protocol), which breaks the sequential constraint of traditional consensus through pipelining to achieve (for geographically co-located clients) a median end-to-end order latency of about 0.2 seconds (200 milliseconds) and a 99th percentile below 0.9 seconds; benchmark tests of its HyperCore component show it can process up to 200,000 orders per second, but this figure comes from a benchmark-test environment rather than production load and is not directly comparable to Binance's production-environment data. The actual throughput and latency data under extreme stress scenarios (such as the liquidation peak during the October 2025 flash crash) have not been publicly disclosed by either platform. Lighter.xyz takes another technical route — a zkRollup architecture based on Ethereum — achieving near-instant transaction confirmation (millisecond-level) through a single sequencer, while using zero-knowledge proofs to submit transaction batches to the Ethereum mainnet to inherit its security. This approach offers extremely high performance at the sequencer level, but its degree of decentralization depends on the design of the sequencer (currently mostly a centralized sequencer, with plans to gradually decentralize in the future).

The degree of decentralization of Hyperliquid, however, needs to be assessed with caution. As of this writing, Hyperliquid's validator set is largely dominated by nodes operated by Hyperliquid Labs, validator admission is not fully permissionless, and sequencing power is substantively concentrated in the core team. The JELLY incident of March 2025 highlighted this centralized operational reality: when the perpetual futures of the JELLY token exhibited suspected manipulation that threatened the solvency of HLP, the Hyperliquid team intervened directly in the market, voting to delist the JELLY contract and forcibly settling all positions at a price favorable to HLP. This intervention technically protected the platform's financial stability but also exposed the tension between its "decentralized exchange" positioning and its actual capacity for cen-

tralized intervention. From a regulatory perspective, this centralized control capacity may expose Hyperliquid to legal-characterization risks similar to those of a centralized exchange with respect to the Howey test and exchange-registration obligations.

The rise of the application-chain architecture has demonstrated that providing a high-performance order book meeting the needs of professional traders, without sacrificing the principle of decentralization, is now technically feasible. This technological advance has had multiple effects on the microstructure of the perpetual futures market: the on-chain-verifiable matching process eliminates the "black box" problem of the CEX, the self-custody asset model eliminates counterparty risk, and performance metrics approaching those of a CEX make professional market makers willing to deploy strategies on these platforms, thereby improving the depth and resilience of the order book.

Figure 8-17 uses an architecture-comparison diagram to visualize the three-stage evolution of the matching engine from the centralized exchange, through the general-purpose-chain hybrid DEX, to the application-chain DEX, highlighting the core differences of each stage in the hosting location of the order book, the settlement mechanism, and the asset-custody model.

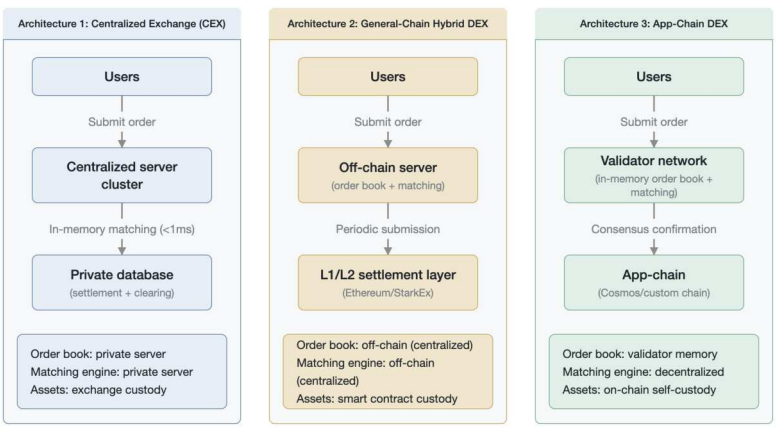


Figure 8-17. A comparison of the three matching-engine architectures (conceptual comparative illustration of the three matching architectures, not empirical data)

From left to right, the figure presents the core decentralization increments of the three architectures: from the decentralization of asset custody (the hybrid DEX migrates user assets to on-chain smart-contract self-custody, but the order book and matching are still off-chain, with only settlement on-chain) to the decentralization of the matching process itself (the application-chain DEX has validators jointly maintain the order book, matching is confirmed through consensus, and assets are fully self-custodied on-chain). The trust basis of the order book market gradually shifts from reliance on the operator's reputation to a verifiability guarantee grounded in cryptographic proof and distributed consensus.

8.7.4 A multidimensional comparison of the three architectures

To present the differences among these three architectures more clearly, we compare them across several key dimensions — matching location, settlement layer, performance metrics, custody model, transparency, MEV risk, and protocol sovereignty (see Table 8-5).

Dimension	CEX (e.g., Binance)	General-purpose-chain hybrid DEX (e.g., dYdX v3)	Application-chain DEX (e.g., dYdX v4, Hyperliquid)
Matching location	Centralized server (private)	Centralized server (private)	Decentralized (validator network)
Order book location	Centralized server (private)	Centralized server (private)	Validator memory (off-chain synchronization) / on-chain state
Settlement layer	Private database	Ethereum L2 (e.g., StarkEx)	Independent application chain (e.g., Cosmos, Hyperliquid L1)
Throughput	Extremely high (~100k+ TPS)	High (limited by L2 performance)	Very high (>10k TPS, Hyperliquid reaching 200k)
Latency	Extremely low (<1 ms)	Low (~10–100 ms)	Relatively low (~100 ms – 1 s)
Custody model	Exchange custody (centralized)	Smart-contract self-custody	On-chain self-custody
Transparency	Opaque (black box)	Partially transparent (settlement results on-chain)	Highly transparent (fully on-chain verifiable)
MEV/manipulation risk	High internal-manipulation risk	Moderate (depends on the fairness of the L2 sequencer)	Relatively low (jointly supervised by decentralized validators)
Protocol sovereignty	Full sovereignty (controlled by a corporate entity)	No sovereignty (depends on the governance of the underlying L1/L2)	Full sovereignty (protocol/DAO autonomy)

Table 8-5. A comparison of the core dimensions of the three matching-engine architectures (Data source: compiled by the author from the public documentation of each platform)

The "application-chain DEX" category in the table covers multiple implementation paths with significant internal architectural differences: dYdX v4 is built on the Cosmos SDK and relies on CometBFT consensus, Hyperliquid runs on its own Layer 1 and uses its self-developed HyperBFT consensus, and Lighter.xyz uses an Ethereum zkRollup architecture and relies on a single sequencer. These three paths differ fundamentally in security assumptions (independent consensus vs. inheriting Ethereum's security), degree of decentralization (dozens of validators vs. a single sequencer), and MEV-protection mechanisms. Grouping them into the same category facilitates macro comparison, but the reader should attend to the specific architectural details of each platform.

Table 8-5 shows clearly that the three architectures form a progressive spectrum between performance and decentralization. The CEX occupies the performance extreme but faces the most limitations in decentralization, transparency, and the security of user assets. The general-purpose-chain hybrid DEX achieves a breakthrough in asset self-custody, but its core matching component remains a centralized core weakness. The application-chain DEX achieves significant progress in both performance and decentralization; although it still lags the CEX on the pure latency metric (about 200 milliseconds vs. <1 millisecond), this gap is narrowing rapidly and is already entirely acceptable for the vast majority of non-high-frequency trading strategies.

As shown in Figure 8-18, in the logarithmic two-dimensional space of throughput and latency, the application-chain DEX (green) is migrating rapidly from the upper-left corner (low throughput, high latency) toward the lower-right corner where the CEX resides (high throughput, low latency), gradually approaching its performance frontier. This convergence will reshape competition: as the performance gap narrows, the relative advantage of the CEX will lie more in compliance licenses, fiat on-ramps and off-ramps, liquidity network effects, and brand recognition, while the unverifiability of its matching and its custody risk will face greater scrutiny as transparency becomes a focus of competition.

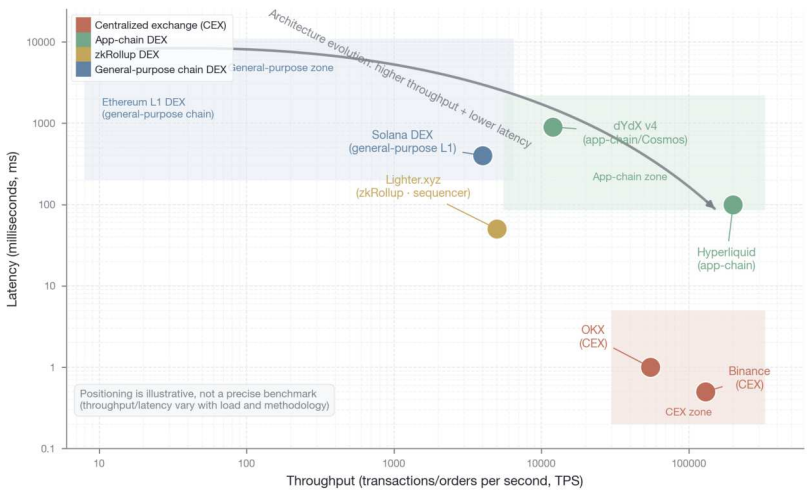


Figure 8-18. The positioning of matching engines in the throughput-latency space (conceptual illustration of two-dimensional throughput-latency positioning; the coordinates of each platform are representative orders of magnitude based on official documentation, not measured against a unified benchmark; Hyperliquid's about 200k is a benchmark-test value, not production load)

This architectural evolution is not merely technical progress; it is the infrastructural precondition for the evolution of the market microstructure. The architecture of the matching engine determines how fast market makers can update quotes (affecting resilience), whether order book depth can be externally verified (affecting the degree of the depth illusion), and whether MEV attackers can use information asymmetry to front-run (affecting the proportion of toxic order flow). Building on this, Section 8.8 further analyzes how the design of the matching mechanism itself (from continuous

auction to batch auction) optimizes market fairness and efficiency through institutional innovation, within the constraints of a given technical architecture.

8.8 Continuous auction and batch auction

Having explored the on-chain evolution of the automated market maker and the central limit order book, we need to examine closely a core assumption of the order book model itself: continuous-time matching. Traditional financial markets and early decentralized exchanges mostly adopt a continuous-auction mechanism by default — that is, orders are matched immediately upon arrival against existing orders in the order book according to the principles of price priority and time priority. As high-frequency trading rose and the characteristics of blockchain infrastructure became apparent, however, the structural defects of this continuous-time matching gradually became evident. This section analyzes how the frequent batch auction and its variants optimize the market microstructure through the discretization of time.

8.8.1 The speed race and latency arbitrage

In the continuous limit order book model, the matching engine follows a strict serial processing logic. When new public information emerges in the market (for example, the price of an asset changes on another exchange), market makers and high-frequency arbitrageurs receive this signal simultaneously. A rational market maker needs to immediately cancel its old quote and submit an updated one to avoid being filled at a stale price, whereas a high-frequency arbitrageur (or scalper) attempts to trade with the market maker at the stale price before the market maker cancels, thereby capturing a risk-free profit [26].

This competition gradually turns into a speed race centered on latency advantage. In traditional financial markets, this race takes the form of enormous hardware investment in microwave communication networks, direct fiber connections, and exchange colocation, with participants competing on infrastructure for an advantage of a few microseconds or even nanoseconds. In their widely cited paper "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," Budish, Cramton, and Shim (2015) noted that the continuous limit order book drives price correlations toward zero at micro time scales, thereby generating obvious mechanical arbitrage opportunities [27]. They emphasized that market competition has not eliminated the scale or frequency of these arbitrage opportunities but has merely raised the speed threshold required to capture them.

In the blockchain environment, because of network latency, the discreteness of block time, and miners'/validators' control over transaction ordering, this speed race evolves into a contest over MEV. When the on-chain price deviates from an external oracle or centralized-exchange price, searchers ensure that their arbitrage transactions are ordered ahead of the market maker's cancellation transaction by paying a high priority fee or directly bribing the block builder. This behavior, known on-chain as latency arbitrage or front-running, is essentially identical to high-frequency arbitrage in traditional markets: both exploit the loophole of the time-priority principle in the CLOB.

The time-priority principle imposes a significant social cost. Under the continuous-auction mechanism, a market maker must bear the risk of having its stale quotes selectively attacked by a faster arbitrageur. This risk does not stem from any inaccuracy in the market maker's pricing model but is caused purely by the physical limitation of the system's serial processing of orders in continuous time. The profit the arbitrageur extracts constitutes a direct loss to the market maker, and this loss is ultimately passed on to all ordinary traders through the market mechanism.

8.8.2 The formation of toxic order flow and adverse selection

When arbitrageurs use their speed advantage to trade against market makers' stale quotes, this order flow is called "toxic order flow." The essence of toxic order flow is the adverse selection caused by information asymmetry. In their classic microstructure theory, Glosten and Milgrom (1985) noted that a market maker facing informed traders who may possess private information or a speed advantage will necessarily suffer adverse-selection losses [26].

In perpetual futures markets, the transmission mechanism of toxic order flow is especially pronounced. When the market is highly volatile, there is not only front-running by high-frequency arbitrageurs but also a large volume of triggered liquidation orders. These liquidation orders often sell or buy at market regardless of cost, forming an extreme toxic order flow. To measure the toxicity of this order flow in real time, Easley, López de Prado, and O'Hara (2011) proposed the VPIN indicator [28]. When the VPIN indicator rises, it means that the toxicity of the order flow is increasing and that the adverse-selection risk faced by market makers rises significantly.

Facing highly toxic order flow, a rational market maker has only two choices: either widen the bid-ask spread to compensate for the expected adverse-selection loss, or reduce resting depth at the best bid and ask to lower risk exposure.

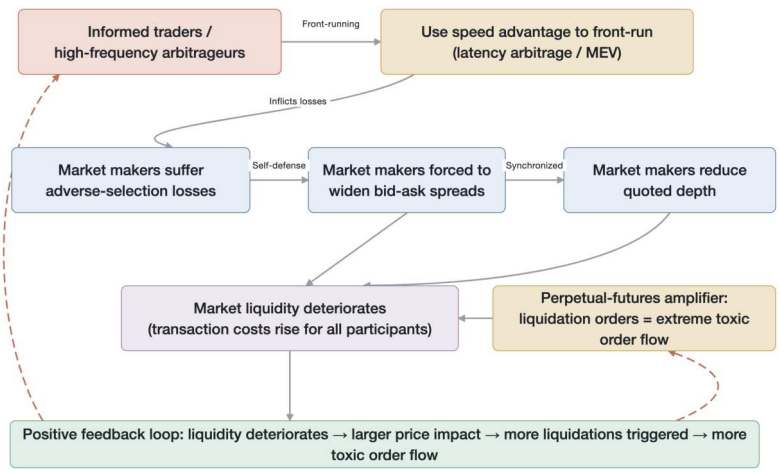


Figure 8-19. The transmission mechanism of toxic order flow (illustration of the transmission mechanism of toxic order flow, not empirical data; theoretical basis: Glosten & Milgrom 1985 [26], Easley,

López de Prado & O'Hara 2011 [28])

As shown in Figure 8-19, the adverse selection triggered by toxic order flow forces market makers to widen spreads and reduce depth, leading to a deterioration of overall liquidity, so that ordinary traders (natural order flow) bear higher spreads and slippage; in the extreme, the drying up of liquidity amplifies price shocks and triggers more liquidations, forming a positive feedback loop that intensifies market instability. Solving the problem of toxic order flow is therefore not only a matter of market-maker profitability but is key to maintaining market health and protecting the interests of ordinary traders.

8.8.3 The frequent batch auction

To fundamentally solve the problems of latency arbitrage and adverse selection brought by the continuous-auction mechanism, academics and industry proposed an alternative market microstructure: the frequent batch auction (FBA). The core idea of the FBA is to abandon the continuous-time assumption, discretize time into extremely short intervals (for example, 100 milliseconds or one block time), and clear orders through a uniform-price two-sided auction at the end of each time window [27].

Under the FBA mechanism, all orders arriving within a time window are treated as arriving simultaneously. The system aggregates all buy orders (constructing the demand curve) and sell orders (constructing the supply curve) and searches for a single clearing price that maximizes the trading volume. All buy orders with a price higher than (or equal to) the clearing price, and all sell orders with a price lower than (or equal to) the clearing price, are executed at this uniform clearing price.

The FBA eliminates the micro-scale speed advantage through time discretization. Within the same auction cycle, an order submitted 1 millisecond — or even tens of milliseconds — later enjoys exactly the same execution priority as an order submitted earlier, as long as it arrives before the window closes. This mechanism pulls the competition among participants back from the "speed dimension" to the "price dimension." Market makers no longer need to invest heavily for a few milliseconds of cancellation speed, and arbitrageurs can no longer profit risk-free by relying on a speed advantage alone to exploit market makers' stale quotes.

Gong, Liu, and Kate (2025) conducted an empirical analysis of BTC-USD and ETH-USD trading data on the decentralized exchange dYdX [29]. Their study uses the realized spread as the measure of transaction cost. The results show that, compared with the CLOB model, the FBA can reduce overall transaction costs by about 21% to 37% (BTC-USD by about 24%–37% and ETH-USD by about 21%–34%, depending on the auction frequency of 5, 10, or 15 seconds). This improvement mainly benefits market makers: because the elimination of latency arbitrage lowers the adverse-selection risk they face, they are able to offer narrower quoted spreads. Note that the above data are based on a sample under normal market conditions, and the batch interval introduced by the FBA itself produces an opportunity cost of execution delay: in a high-volatility environment, a trader may miss favorable price movements while waiting for the next auction window, and whether this delay cost partly offsets the spread improvement was not included in the study's analysis. In addition, whether this conclusion applies to less liquid long-tail assets awaits further verification. Their study notes that when public

information in the market (such as price-oracle updates) is more frequent than private information, the FBA has a clear welfare advantage over the CLOB.



Figure 8-20. The transaction-cost-reduction effect of the FBA (the only empirical figure in this chapter: relative transaction cost is baselined at CLOB = 100%; the FBA lowers BTC-USD by 37% to 63% and ETH-USD by 21% to 79%; the 37% and 21% are the two endpoints of the 21%–37% range, depending on the auction frequency of 5, 10, or 15 seconds; in the main text the overall reduction is 21%–37%, BTC about 24%–37%, and ETH about 21%–34%; data source: Gong, Liu & Kate 2025 [29], based on BTC-USD and ETH-USD trading data from the dYdX exchange, January 2023 – October 2024)

As shown in Figure 8-20, by eliminating latency arbitrage the FBA substantially lowers market makers' adverse-selection risk, making them willing to offer narrower spreads and deeper liquidity; its uniform-price clearing also provides price improvement for market orders — buyers may be filled below their expected ceiling, and sellers above their expected floor.

8.8.4 The dual-flow batch auction

Although the FBA succeeds in eliminating the speed advantage, it remains an "all-to-all" market structure. In a traditional FBA, market makers may still be matched against one another. To optimize the market structure further, institutions such as Jump Crypto proposed the dual-flow batch auction (DFBA) mechanism [30].

The core innovation of the DFBA is to explicitly distinguish order flow into two categories — makers (liquidity providers) and takers (liquidity consumers / natural order flow) — and to strictly prohibit makers from being matched against one another. In the DFBA, each auction cycle actually contains two independent auction processes: the buy-side auction matches makers' sell orders against takers' buy orders, and the sell-side auction matches makers' buy orders against takers' sell orders.

This design brings several key advantages. It eliminates adverse-selection competition among market makers. In a continuous auction or an ordinary FBA, when an external price changes, a market maker must guard not only against arbitrageurs but also

against other, faster-reacting market makers. By prohibiting maker-to-maker matching, the DFBA allows market makers to quote their best prices with confidence, without worrying about being "sniped" by peers.

At the same time, the DFBA strongly encourages market makers to display genuine liquidity depth. Because the clearing price is a uniform price determined by the supply and demand curves, and fills at the clearing price are typically allocated pro rata, market makers have a strong incentive to provide as large a size as possible at the best price in order to capture more taker order flow.

A key implementation challenge of the DFBA mechanism lies in the classification standard for makers and takers. In a continuous auction, this distinction is clear: whoever rests passively on the order book waiting to be filled is a maker, and whoever actively issues a taking instruction is a taker. In a batch-auction environment, however, all orders participate in clearing "simultaneously" when the window closes, and the boundary between maker and taker becomes blurred. The classification method used in Jump Crypto's proposal is based on the ex ante declaration of order type and submission intent [30], but this may give rise to circumvention behavior, in which participants disguise what is essentially active trading as passive resting in order to enjoy maker treatment. As of this writing, no mainstream exchange has fully deployed the DFBA mechanism in a production environment, and the robustness of its classification standard remains at the stage of theoretical discussion.

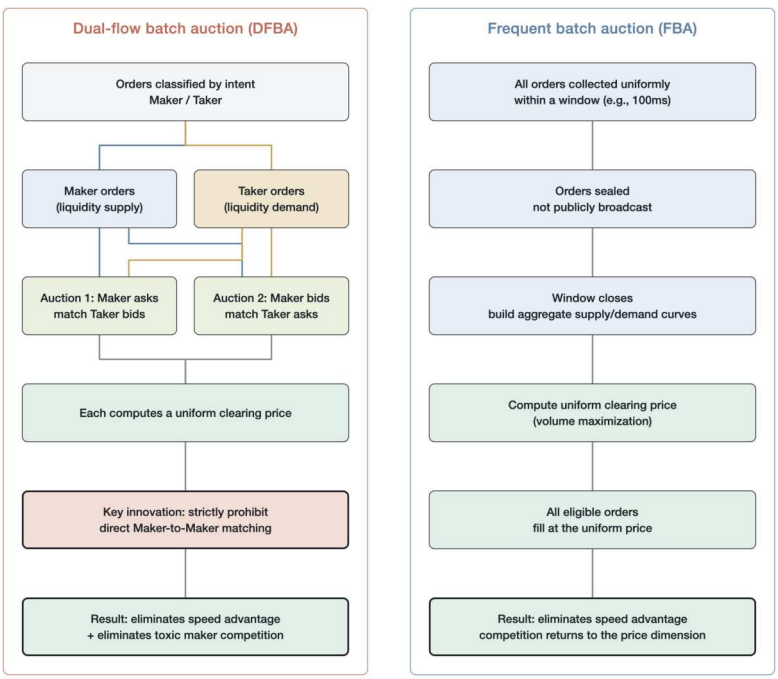


Figure 8-21. A process comparison of the FBA and the DFBA (illustration of the process comparison between the FBA and the DFBA, not empirical data; following Jump Crypto 2025 [30])

As shown in Figure 8-21, while inheriting the FBA's advantage of eliminating the speed advantage, the DFBA lowers the adverse-selection risk market makers face through structured order isolation. This improved environment ultimately translates into narrower spreads and deeper market depth, thereby benefiting the natural order flow the market is meant to serve.

8.8.5 Injective's FBA practice

A representative case of putting the FBA mechanism into practice on a blockchain is the Injective protocol. Injective is an application chain optimized specifically for decentralized finance, and it is the first blockchain network to embed the frequent batch auction mechanism directly into the Layer 1 consensus layer [31].

Implementing the FBA on a traditional smart-contract platform faces many challenges, especially the problem of order privacy protection. If orders are publicly broadcast during the auction window, they can easily trigger front-running or sandwich attacks by MEV bots. Injective effectively solves this problem by combining the Tendermint consensus mechanism with a sealed-bid design.

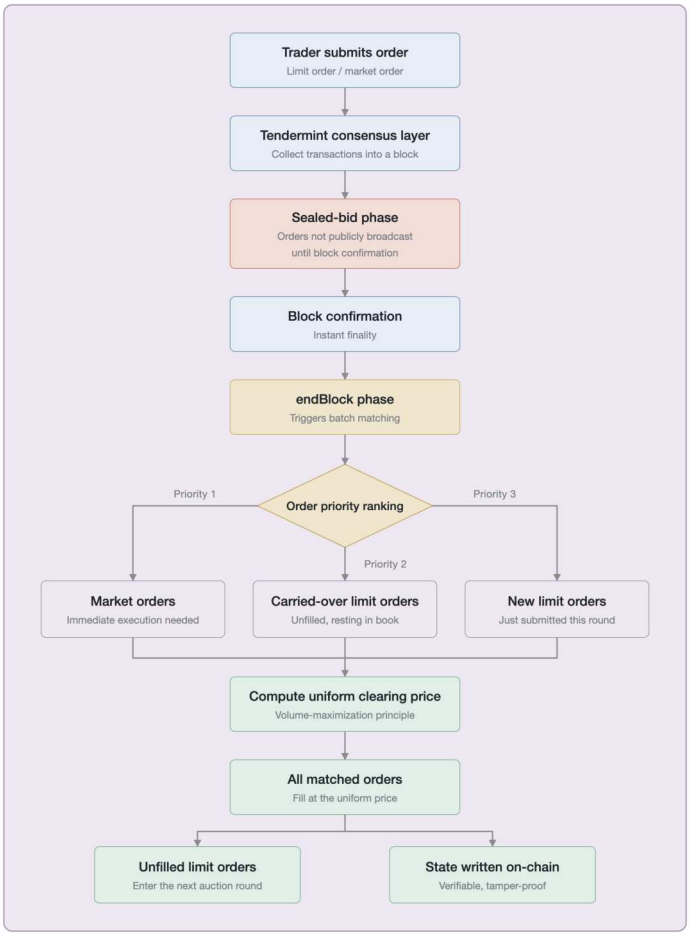


Figure 8-22. Injective's on-chain FBA execution flow (illustration of Injective's on-chain FBA execution flow, not empirical data; following Injective Protocol 2024 [31])

Injective's FBA execution flow (as shown in Figure 8-22) is deeply bound to the block life cycle. In the phase in which a block collects transactions, all submitted orders are in a sealed state and do not expose their specific price and quantity information in the public mempool. Only after the block is confirmed (which, thanks to Tendermint's instant finality, typically takes only an extremely short time) are the orders decrypted.

In the final phase of the block life cycle, the system triggers the batch matching engine. The matching logic follows a strict priority order: market orders have the highest priority to ensure the need for immediate execution, followed by unfilled limit orders from the previous round, with the lowest priority assigned to newly submitted limit orders in the current round. After determining the pool of orders participating in the matching, the engine computes the uniform clearing price that maximizes the trading volume and settles all eligible orders at this price.

Injective's practice demonstrates the feasibility and superiority of the FBA in an on-chain environment. By synchronizing time discretization with block generation and supplementing it with cryptographic means to protect order privacy, Injective has successfully built an order book market on-chain that can effectively resist latency arbitrage and harmful MEV. As of this writing, however, the liquidity and volume data for Injective's perpetual futures market remain relatively limited, making it difficult to empirically compare the FBA's real effect on this platform against theoretical expectations. Whether the theoretical advantages of the FBA hold equally in a low-liquidity environment (that is, whether uniform-price clearing can still achieve meaningful price improvement when there are few participants and possibly only a handful of orders within a batch-auction window) still requires more practical data for verification. This provides not only a predictable trading environment for market makers but also lays a microstructural foundation for the decentralized perpetual futures market to advance toward institutional-grade liquidity.

8.9 The information ecosystem of on-chain and off-chain order books

Having explored the matching mechanism of the order book, we must turn our attention to the data the order book itself carries. The order book is not only where trading happens but also the market's most important source of microstructural information. From centralized exchanges to decentralized exchanges, the migration of the order book involves both a transformation of infrastructure and a structural change in the information ecosystem. This section compares the fundamental differences between on-chain and off-chain order books in transparency, verifiability, and the barrier to data acquisition, and explores how these differences give rise to an entirely new on-chain data-analysis ecosystem.

8.9.1 The characteristics of the CEX order book

On centralized exchanges, the order book is highly opaque and unverifiable. Traders receive order book snapshots and incremental updates through the APIs (such as WebSocket) the exchange provides, but this data is generated entirely by the exchange's centralized servers. Users cannot independently verify whether a resting order at a specific price level truly exists, nor can they confirm whether the exchange is conducting wash trading or spoofing internally to manufacture false liquidity depth.

What this unverifiability buys is high performance. Because there is no need to reach consensus in a distributed network, a CEX's matching engine can achieve submillisecond latency. More importantly, canceling or updating a quote on a CEX is typically free (or extremely low-cost). This zero-cost quote-update mechanism lets high-frequency market makers revise their quotes dozens of times per second in response to tiny market fluctuations.

This information ecosystem, based on centralized trust and zero-cost updates, however, causes the CEX order book to be flooded with a large volume of "ghost liquidity." When the market fluctuates violently, these seemingly substantial resting orders are often canceled in an instant, causing liquidity to vanish suddenly precisely when it is most needed.

8.9.2 The characteristics of the on-chain CLOB

Entirely unlike the CEX, a central limit order book built fully on a blockchain (such as Hyperliquid or Injective) offers a fully transparent and verifiable information ecosystem. In an on-chain CLOB, every resting order, cancellation, and trade record is packed into a block as a transaction and broadcast to the entire network through the consensus mechanism. Anyone can run a full node, replay all transaction states from the genesis block, and thereby reconstruct locally an order book precisely consistent with the on-chain state.

This verifiability changes the trust basis of order book depth, but its meaning needs to be defined precisely. An on-chain resting order, at the legal and operational level, is still a revocable offer, just like an order on a traditional order book: a market maker can submit a cancellation transaction in the next block to cancel a prior quote. The POPCAT incident analyzed in Section 8.2.5 has already clearly demonstrated this: the $20 million order wall the manipulator posted on-chain was ultimately withdrawn in full. An on-chain resting order therefore does not possess "execution irrevocability." What the on-chain order book genuinely improves is the degree to which "capital authenticity" is verifiable: every resting order must lock in real on-chain margin, and both submitting and canceling orders require paying gas fees, which significantly raises the capital cost of spoofing at the economic level. On a centralized exchange, posting and canceling a false order of $100 million is nearly costless, but in an on-chain environment the manipulator must at least actually lock in the corresponding margin assets, substantially raising the economic barrier to spoofing.

At the same time, the full transparency of the on-chain order book is a double-edged sword. The distribution of liquidation prices, the location of large positions, and the

price ranges with weak liquidity — this information is visible equally to all market participants, including potential attackers. As the POPCAT incident demonstrated, an attacker can use on-chain-visible position-distribution information to locate a manipulation target precisely and design a targeted liquidity-attack strategy. This "information symmetry" aids price discovery efficiency in a normal market, but in an adversarial environment it also provides malicious actors with a complete "map of the attack surface."

This verifiability, however, comes at the cost of speed and flexibility. The update frequency of an on-chain order book is limited by the block time of the underlying blockchain. Even on a high-performance application chain or Layer 2 network, this time is typically between a few hundred milliseconds and a few seconds. In addition, every on-chain state update (including cancellations) requires paying a gas fee. This time constraint and cost friction force on-chain market makers to abandon high-frequency quote-update strategies in favor of wider spreads and more stable resting-order strategies. As a result, although the depth of an on-chain order book may not be as "thick" as that of a CEX, the liquidity it provides is more genuine and more durable.

8.9.3 The democratization of data access

In traditional finance and CEX markets, obtaining high-quality order book data is an expensive privilege. Exchanges typically divide data into different tiers (such as Level 1, Level 2, and Level 3) and charge high subscription fees for Level 3 data, which provides full depth and tick-by-tick trade records. In addition, to obtain the lowest-latency data feed, institutions must also pay expensive colocation fees. This data-commercialization model creates a stratification of information access, placing ordinary retail traders at a structural disadvantage.

On the surface, the public and transparent nature of the blockchain seems to "democratize" this reading capability. On-chain, order book data becomes a public good. No centralized entity can restrict access to this data, and there is no need to pay expensive API subscription fees. Anyone willing can obtain the lowest-level smart-contract state through an RPC node and parse out the most complete Level 3 order book data.

This gives rise to an interesting "transparency paradox," however: although the data is public, the barrier to extracting valuable information from it is extremely high. To reconstruct and analyze an active on-chain order book in real time requires building a high-availability node cluster, writing complex smart-contract event-parsing logic, and handling a massive data stream. For an ordinary DeFi user facing a vast volume of hexadecimal raw data, the barrier to interpreting the data is actually higher than when using a CEX's standardized API. The "reading" capability has therefore been democratized in theory, but in engineering practice it remains in the hands of a small number of professional teams with powerful infrastructure capabilities.

8.9.4 The on-chain data-analysis ecosystem

It is precisely on the basis of this "transparency paradox" that the decentralized perpetual futures market has given rise to an entirely new on-chain data-analysis ecosystem.

Since the underlying order book data is public and tamper-proof, third-party developers can build a rich variety of visualization tools and real-time monitoring platforms on top of it, filling the information gap between ordinary users and professional market makers.

In this new ecosystem, we see many innovative data products. For example, through joint analysis of on-chain open interest and order book depth, data platforms can plot high-precision "liquidation heatmaps." These heatmaps can intuitively show users the price ranges in which large volumes of high-leverage positions are clustered, and the chain reaction that may be triggered when these positions are liquidated.

In addition, targeting the "whale behavior" unique to the on-chain order book, analysis tools can track in real time the resting and cancellation actions of specific addresses, identifying potential "order-propping" or "order-suppressing" behavior. Because the transaction history of an on-chain address is fully public, users can even assess whether the address behind these large orders has sufficient financial strength, thereby judging the true intent of these orders.

Feature dimension	Centralized exchange order book	Fully on-chain order book
Data-generating entity	Exchange's centralized servers	Decentralized consensus network
Transparency and verifiability	Black box, unverifiable, risk of false trades	Fully transparent, cryptographically verifiable, every trade recorded
Data-acquisition cost	High (Level 3 data and colocation require payment)	Free (data is a public good, obtained via RPC)
Engineering-parsing barrier	Relatively low (standardized API directly outputs structured data)	Extremely high (requires parsing underlying contract events and the state tree)
Quote-update friction	Extremely low (sub-millisecond latency, typically no cancellation fee)	Relatively high (limited by block time, requires paying gas fees)
Liquidity characteristics	Large depth, but containing a large volume of easily cancelable "ghost liquidity"	Relatively smaller depth, but strongly committed and more genuine liquidity

Table 8-6. A comparison of the core differences between the CEX and on-chain order books at the level of the information ecosystem (Data source: compiled by the author)

Table 8-6 summarizes the core differences between the CEX and on-chain order books at the level of the information ecosystem. Overall, the on-chain order book shifts trust from "reliance on a centralized entity" to "verification of cryptography and code." Although this shift brings engineering challenges, it compresses the space for a centralized operator to conduct unverifiable black-box operations (such as internal wash trading, zero-cost spoofing, and improper operations during liquidation), and it lays the data foundation for building a fairer, more transparent decentralized derivatives market. As the POPCAT incident in Section 8.2.5 shows, however, the on-chain transparent environment does not eliminate market manipulation itself but merely changes the cost structure and traceability of manipulation. As infrastructure continues to improve and third-party analysis tools mature, the information ecosystem of the on-chain order book

is expected to continue developing toward greater openness and efficiency.

8.10 Chapter summary

Taking the order book as its central object, this chapter has constructed an analytical framework for understanding the operating logic of the perpetual futures market from the standpoint of microstructure. The starting point of this framework is the redefinition of the order book as an information machine: the five signal channels of spread, depth, slope, flow, and dynamics continuously gather the private information and risk preferences of market participants and transform them into observable state variables. By cross-validating the joint changes in these five channels (rather than observing any single indicator in isolation), one can diagnose the market state systematically, mapping the continuous deterioration from "healthy" to "collapsed" into identifiable structural features that can be flagged early.

The depth illusion is the core risk concept this chapter reveals. A systematic gap exists between the visible nominal depth on the order book and the trading size the market can genuinely bear under stress, and this gap is amplified in the perpetual futures market by three mechanisms: high-frequency market makers' flickering quotes cause a large volume of visible depth to exist only within a millisecond-scale time window; market makers' liquidity commitment is implicitly premised on "normal market conditions" and triggers a collective retreat once volatility rises; and the strategically constructed false depth demonstrated by the Hyperliquid POPCAT incident shows that visible depth can even be actively constructed as a tool of manipulation. Empirical data from the October 2025 flash crash show that the reduction between nominal depth and stress depth exceeded 90%, providing quantitative validation of this concept.

Building on this, the chapter has proposed the four microstructural fingerprints of the perpetual futures order book: the depth cliff, the depth tide, the liquidity circadian rhythm, and the asymmetric volatility response of the spread. None of these four patterns exists in traditional spot markets or in futures markets with an expiry date; they arise, respectively, from perpetual futures' unique institutional features of the liquidation mechanism, the funding rate, 24/7 trading, and the systematic long bias. When the four fingerprints resonate and superimpose in a specific market environment (for example, when a high-funding-rate period coincides with a liquidity-trough period and a dense liquidation zone approaches the current price), the systemic risk the market faces far exceeds the level any single fingerprint could foreshadow.

The resilience analysis further reveals the structural fragility of the perpetual futures market along the time dimension. The positive-feedback mechanism of the liquidation cascade gives shocks a self-amplifying character, market makers face open-ended inventory risk because of the absence of an expiry date and therefore have a lower retreat threshold, and 24/7 uninterrupted trading eliminates the resilience-recovery window that the closed period provides in traditional markets. These three factors together produce the asymmetric resilience characteristic of "collapse measured in minutes, recovery measured in weeks," with the fact that depth had still not recovered to pre-shock levels months after the October 2025 event serving as direct evidence.

Finally, the chapter has traced the evolutionary path of the technical architecture that carries the order book from the centralized exchange to the application chain, as well as the institutional innovation of the matching mechanism from continuous auction to batch auction. The application-chain architecture achieves the verifiability of the matching process through a decentralized validator network, eliminating the CEX's counterparty risk and the possibility of manipulation; the frequent batch auction and the dual-flow batch auction, through time discretization and order-flow separation, reduce at the mechanism level the harm that latency arbitrage and toxic order flow inflict on market makers, thereby lowering traders' implicit costs. These advances at the technical and institutional levels are reshaping the microstructural foundation of the perpetual futures market, and their implementation and effects on specific platforms will be analyzed in detail in Chapter 9, using Hyperliquid as a case study.

References

[1] Amberdata. (2025, October). *How $3.21B vanished in 60 seconds: October 2025 crypto crash explained through 7 charts*. https://blog.amberdata.io/how-3.21b-vanished-in-60-seconds-october-2025-crypto-crash-explained-through-7-charts

[2] FTI Consulting. (2025). *Crypto crash Oct 2025: Leverage meets liquidity*. https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity

[3] CoinDesk. (2025, November 15). Crypto liquidity still hollow after October crash, risking sharp price swings. *CoinDesk*. https://www.coindesk.com/markets/2025/11/15/crypto-liquidity-still-hollow-after-october-crash-risking-sharp-price-swings

[4] Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press.

[5] Harris, L. (2003). *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press.

[6] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

[7] Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, *3*(2), 5–39. https://doi.org/10.21314/jor.2001.041

[8] Bouchaud, J.-P., Farmer, J. D., & Lillo, F. (2009). How markets slowly digest changes in supply and demand. In T. Hens & K. R. Schenk-Hoppé (Eds.), *Handbook of Financial Markets: Dynamics and Evolution* (pp. 57–160). North-Holland. https://doi.org/10.1016/b978-012374258-2.50006-3

[9] Cont, R., Kukanov, A., & Stoikov, S. (2014). The price impact of order book events. *Journal of Financial Econometrics*, *12*(1), 47–88. https://doi.org/10.1093/jjfinec/nbt003

[10] Tiniç, M., Sensoy, A., Akyildirim, E., & Corbet, S. (2023). Adverse selection in cryptocurrency markets. *The Journal of Financial Research*, *46*(2), 497–546. https://doi.org/10.1111/jfir.12317

[11] Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56. https://doi.org/10.1016/s1386-4181(01)00024-6

[12] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[13] CoinDesk. (2025, November 13). Peak degen warfare: Alleged POPCAT manipulation hits Hyperliquid with $4.9M loss. *CoinDesk*. https://www.coindesk.com/markets/2025/11/13/peak-degen-warfare-alleged-popcat-manipulation-hits-hyperliquid-with-usd4-9m-loss

[14] Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The Flash Crash: High-frequency trading in an electronic market. *The Journal of Finance*, *72*(3), 967–998. https://doi.org/10.1111/jofi.12498

[15] Ruan, Q., & Streltsov, A. (2022). *Perpetual futures contracts and cryptocurrency market quality* (SSRN Working Paper No. 4218907, posted 2022, rev. 2024). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4218907 https://doi.org/10.2139/ssrn.4218907

[16] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

[17] Commodity Futures Trading Commission. (2023, January 9). CFTC charges Avraham Eisenberg with manipulative and deceptive scheme to misappropriate over $110 million from the Mango Markets exchange [Press Release No. 8647-23]. https://www.cftc.gov/PressRoom/PressReleases/8647-23

[18] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, *135*(2), 293–319. https://doi.org/10.1016/j.jfineco.2019.07.001

[19] Cont, R., Stoikov, S., & Talreja, R. (2010). A stochastic model for order book dynamics. *Operations Research*, *58*(3), 549–563.

[20] Sornette, D. (2003). *Why Stock Markets Crash: Critical Events in Complex Financial Systems*. Princeton University Press.

[21] Large, J. (2007). Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, *10*(1), 1–25. https://doi.org/10.1016/j.finmar.2006.09.001

[22] Wuyts, G. (2008). *The impact of liquidity shocks through the limit order book* (CFS Working Paper No. 2008/53). Center for Financial Studies.

[23] Lo, D., & Hall, A. (2015). Resiliency of the limit order book. *Journal of Economic Dynamics and Control*, *61*, 222–244. https://doi.org/10.1016/j.jedc.2015.09.012

[24] Amberdata. (2026, February). Bitcoin below $70K: The crash, the data, and what comes next. *Amberdata Blog*. https://blog.amberdata.io/bitcoin-below-70k-the-crash-the-data-and-what-comes-next

[25] Binance. (n.d.). Binance Futures for institutions: Pioneering performance & technology. *Binance Blog*. https://www.binance.com/en/blog/futures/binance-

futures-for-institutions--pioneering-performance--technology-421499824684900642

[26] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100. https://doi.org/10.1016/0304-405x(85)90044-3

[27] Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, *130*(4), 1547–1621. https://doi.org/10.1093/qje/qjv027

[28] Easley, D., López de Prado, M. M., & O'Hara, M. (2011). The microstructure of the "Flash Crash": Flow toxicity, liquidity crashes, and the probability of informed trading. *The Journal of Portfolio Management*, *37*(2), 118–128. https://doi.org/10.3905/jpm.2011.37.2.118

[29] Gong, T., Liu, Z., & Kate, A. (2025). The case of FBA as a DEX processing model. In *Proceedings of Financial Cryptography and Data Security (FC'25)*. https://arxiv.org/abs/2302.01177 https://doi.org/10.1007/978-3-032-07024-1_7

[30] Jump Crypto. (2025). *Dual flow batch auction*. https://jumpcrypto.com/writing/dual-flow-batch-auction/

[31] Injective Protocol. (2024). A new era of Injective: Limitless scale. *Injective Blog*. https://blog.injective.com/en/a-new-era-of-injective-limitless-scale/

Hyperliquid's Microstructure Innovations

Chapter 7 defined the order as a complex vehicle that carries intent, risk, and information, and it systematically analyzed the signal strength and maximal extractable value (MEV) exposure of different order types. Building on this, Chapter 8 dissected the structural dynamics of the order book, revealing microstructural fragilities such as depth illusion and liquidation cascades, and traced the evolution of matching mechanisms from continuous auction to batch auction. Together, the analytical framework established by these two chapters points to a core question that remains unanswered: when a large number of buy and sell orders arrive within the same time window, whose order should be executed first? This ordering problem is decisive for the profits and losses of market participants, shapes the structure of the market ecosystem, and even defines market fairness. The central tension of market microstructure is that a market must simultaneously provide traders with ample liquidity and efficiently aggregate information into prices, and the ordering mechanism is precisely what mediates this tension. In a traditional centralized exchange (CEX), this core mechanism resides within an opaque, closed architecture. The matching engine processes orders at microsecond or even nanosecond speeds, and its internal ordering logic, latency compensation, and potential differential treatment are entirely opaque to the outside world. This opacity gives rise to the costly speed race analyzed in Section 8.8.1. Aquilina, Budish, and O'Neill (2022) found that the social cost of this speed race across the world's major exchanges amounts to billions of dollars annually [1]. Participants lay dedicated fiber-optic lines, build microwave towers, and colocate their servers in exchange data centers, all to gain a microsecond-level physical head start.

The advent of blockchain, and in particular the rise of general-purpose smart contract platforms such as Ethereum, transferred the power to order transactions from private exchange operators to networks of validators. A distinction is warranted here: *decentralization* in this context operates on two different levels — the protocol-level decentralization of rule definition (the ordering rules are encoded in open-source software rather than privately controlled by a single entity) and the operational decentralization of rule enforcement (the size of the validator set, the distribution of stake, and the governance structure). This *open but semantically agnostic* ordering paradigm created transparency, but it also gave rise to the problem of MEV. In a public mempool, all pending transactions are fully visible, allowing MEV searchers to bribe block producers to arbitrarily reorder, insert, or censor transactions, and thereby to execute predatory

strategies such as sandwich attacks and front-running. In this environment, market makers face not only the adverse selection risk of traditional markets but also continuous value extraction from the MEV supply chain, leaving them exposed to greater risk than in the CEX environment.

It is against this backdrop that the emergence of Hyperliquid offers an important case study. It is not merely a decentralized exchange (DEX) but an open and observable microstructure experiment: all of its rules, trades, and order book states are recorded on-chain and can be audited and verified by anyone, in marked contrast to the *black box* of a CEX. Addressing the limitations of the first two paradigms, Hyperliquid's designers proposed an entirely new solution — semantic ordering: identifying the economic meaning of a transaction at the consensus-protocol level and ordering transactions differentially on that basis, placing liquidity-protecting actions ahead of liquidity-consuming ones (its definition and implementation are detailed in Section 9.1).

This chapter analyzes Hyperliquid's microstructure system layer by layer: the two mechanism innovations of semantic ordering and cancellation priority, and how the discrete-time property — an inevitable consequence of on-chain matching — provides their structural foundation. It examines how these elements operate in concert and how they trade off among liquidity, efficiency, fairness, transparency, and robustness, while also probing their boundary conditions and failure modes. Among these, the JELLY incident provides a critical stress test that exposes systemic fragilities lying beyond microstructural protection. Understanding these underlying ordering and matching mechanisms is also the key foundation for analyzing the higher-level liquidity incentive structures built on top of them, such as the HLP Vault.

9.1 The semantic ordering mechanism

How transaction ordering fundamentally determines the ultimate profits and losses of market participants is a core question in understanding market microstructure. The closed time-priority of the CEX, the open bribe-priority of the general-purpose DEX, and the semantic priority of Hyperliquid — these three paradigms lead to sharply different outcomes in the distribution of value. As the first layer and guiding design principle of Hyperliquid's mechanism innovations, semantic ordering interprets the economic meaning of transactions directly at the consensus-protocol level and, on that basis, makes explicit priority choices that embody a value judgment.

Semantic ordering refers to a mechanism in which the consensus protocol determines a transaction's execution priority within a block according to its economic intent, rather than by a bare timestamp or the fee paid. This marks a paradigm shift: from semantically agnostic processing that treats all transactions *identically* to semantically aware processing that treats different transactions *differently*.

9.1.1 The limitations of traditional paradigms

In a traditional CEX, the matching engine is a highly optimized piece of software whose primary design goals are throughput and low latency. To achieve high performance, the engine must be kept as simple as possible, treating every incoming instruction —

whether a new order, a cancellation, or an amendment — as a data packet devoid of economic meaning. Minimizing per-order semantic processing has long been regarded as advantageous for matching speed. The engine's ordering logic follows a strict price-time priority principle: at the same price, the timestamp at which an order reaches the server determines its priority. This ordering process takes place in a fully closed, opaque *black box*; outside participants can neither observe nor audit its internal workings, which in principle leaves room for the exchange to treat clients differentially. This *semantically agnostic plus closed ordering* paradigm made the *speed race* the primary dimension of market competition — the source of the billions of dollars in social cost noted at the start of this chapter [1]. In a continuous-time black-box game, speed directly determines profitability.

The emergence of general-purpose blockchains such as Ethereum transformed the closed ordering regime of the CEX. On an on-chain DEX, transactions are broadcast to a public mempool, and the validators responsible for assembling blocks determine the final ordering. This *semantically agnostic plus open ordering* paradigm achieved procedural transparency but also introduced the problem of MEV. Anyone can observe the pending transactions in the mempool, and this transparency creates significant opportunities for MEV extraction. Typical MEV strategies include sandwich attacks, front-running, and back-running. Daian et al. (2020) first systematically described this phenomenon and warned that it could threaten the consensus stability of blockchains [2]. Under this paradigm, block producers effectively act as allocators of ordering rights, and the highest-bidding MEV searcher obtains a priority position in the transaction ordering. Value is transferred from ordinary users and liquidity providers to a complex supply chain composed of searchers, builders, and proposers. For market makers, this environment is harsher than the CEX: they face not only traditional adverse selection risk but also additional value extraction from the MEV supply chain.

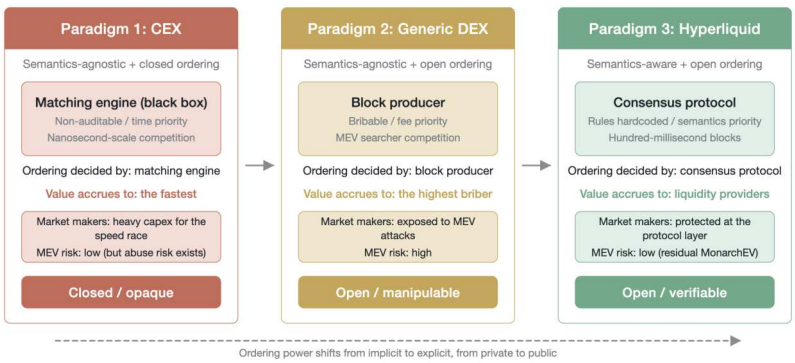


Figure 9-1. Comparison of the three ordering paradigms (conceptual architecture, not empirical data; the MEV and *bribable* mechanisms follow Daian et al. 2020 [2], and the social cost of the CEX speed race follows Aquilina et al. 2022 [1]; the multi-level ordinal ranking of *semantic priority* is an analytical characterization, whereas what the documentation records is *cancellations and post-only orders ahead of GTC/IOC*)

As shown in Figure 9-1, the three paradigms evolve progressively — ordering rights move from an unauditable matching engine, to a bribable block producer, and finally

to public information hardcoded in the consensus protocol. The core trend is that ordering power moves from implicit to explicit and from private to public.

9.1.2 Definition and implementation

Hyperliquid's designers recognized that the core problem of the first two paradigms is rooted in semantic agnosticism. Whether the closed black box of the CEX or the open auction of the general-purpose DEX, both treat all transactions as homogeneous data packets, entirely disregarding the fundamental economic differences among transaction types. A market maker's cancellation instruction and an arbitrageur's aggressive order are economically very different in meaning, yet under both paradigms they are treated as entirely equivalent.

Hyperliquid's solution embeds this capacity for distinction directly into the blockchain's consensus protocol. When its layer-1 (L1) blockchain protocol processes the transactions within a block, it does not order them by time or fee but first performs a semantic classification. Each transaction is parsed for its type tag — the order types defined in Chapter 7, such as Cancel, post-only, GTC, and IOC — and is then explicitly sorted into two priority categories [3], as shown in Table 9-1.

Priority	Transaction type	Economic meaning
System (highest)	Mark price updates, funding rate settlement	System maintenance: ensuring the correctness of price anchoring and contract settlement.
High	`Cancel`, `Amend/Replace`, `Post-only`	Liquidity protection or provision: a market maker withdraws a potentially loss-making quote, modifies the parameters of an existing order, or passively submits a limit order that only rests on the book and does not take.
Standard	Liquidation engine forced-closure orders	Solvency maintenance: forced liquidations triggered by the liquidation engine when margin is insufficient.
Low	`GTC`, `IOC`, triggered conditional orders	Liquidity consumption: a taker actively accepts the current quote to execute, consuming liquidity on the order book.

Table 9-1. Hyperliquid transaction priority classification (Data source: compiled by the author)

The priority tiering in Table 9-1 reflects a fine-grained design trade-off. Placing liquidation engine orders at *standard* priority rather than *high* priority carries important implications: if liquidation orders were given too low a priority, solvency protection would be delayed and systemic risk would rise; if too high, an attacker might deliberately trigger liquidations to obtain high-priority execution, creating a new attack vector. Standard priority allows market makers to withdraw their quotes before liquidation orders execute; this protects market makers but may also degrade the quality of

liquidation execution, a trade-off examined concretely in the JELLY incident in Section 9.4.3. An Amend/Replace order is semantically equivalent to the atomic operation of a Cancel plus a new order; it inherits high priority because its core intent is still to adjust an existing supply of liquidity rather than to consume liquidity. Hyperliquid's official documentation and its January 2024 technical blog post [4] explicitly record a protocol-level rule that orders transactions by semantic type: cancellations and post-only orders execute ahead of GTC/IOC taker orders. (The documentation describes within-block ordering as three classes: (1) actions that do not send a GTC/IOC order to the book, (2) cancellations, and (3) actions that send a GTC/IOC order.) The listing above, which places system-maintenance transactions and the liquidation engine in separate *System* and *Standard* tiers, is an interpretive reconstruction made in this chapter for analytical convenience; the documentation does not explicitly frame mark price and funding rate updates and liquidation as ordered priority tiers. The claims below that *liquidation sits at standard priority and that market makers can cancel before a liquidation fills* actually rest on the documented mechanism that *cancellations precede all fillable orders, and a liquidation fill is a fillable order*, rather than on the four-tier ordinal ranking itself, which the documentation does not confirm.

When a block is processed, transactions execute in strict sequence according to the priority tiers above. These rules are encoded in the open-source software run by all validator nodes, and changing them requires the majority consensus of the validator set rather than the unilateral decision of any single entity. However, as Section 9.4.7 will analyze, the size of the validator set and the distribution of stake determine the actual strength of this safeguard. In economic terms, this *semantically aware plus open ordering* paradigm amounts to an explicit value judgment made at the consensus level: actions that protect and provide liquidity take priority over actions that consume it. Table 9-2 provides a systematic comparison of the three ordering paradigms.

Dimension	CEX (semantically agnostic + closed)	General-purpose DEX (semantically agnostic + open)	Hyperliquid (semantically aware + open)
Who decides ordering	Matching engine (unauditable)	Block producer (bribable)	Consensus protocol (rules hardcoded)
Value allocation logic	Accrues to the fastest	Accrues to the highest briber	Favors liquidity providers
Market maker's position	Must incur costs to join the speed race	Disadvantaged (exposed to MEV attacks)	Protected at the protocol level
MEV risk	Low (but risk of exchange abuse of power)	High	Low (only residual MonarchEV)

Table 9-2. Systematic comparison of the three ordering paradigms (Data source: compiled by the author)

A key pattern emerges from the comparison in Table 9-2: in the CEX value accrues to the fastest, in the general-purpose DEX to the highest briber, and in Hyperliquid it is systematically directed toward liquidity providers. The three paradigms also handle MEV risk in sharply different ways: the CEX internalizes it as the exchange's private

informational advantage, the general-purpose DEX externalizes it as open rent-seeking competition, and Hyperliquid compresses it to a structural residual through semantic rules.

To make the differences among these three paradigms concrete, the following analysis traces how ordering determines trading outcomes by comparing the results of an identical set of trading instructions under the three ordering regimes. In a hypothetical scenario, the price of ETH on Binance drops suddenly from \$2,000 to \$1,980 at t = 0. At this moment, on a given trading platform, three participants issue instructions almost simultaneously (Table 9-3). The three instructions reach the platform within an extremely short time window (< 200 ms), and the different ordering rules determine the final distribution of profits and losses.

Participant	Role	Action	Intent
Alice	Market maker	Cancel — withdraw the \$1,995 bid (10 ETH)	Avoid being filled at a stale price and protect herself against adverse selection
Bob	Arbitrageur	IOC sell order — sell 10 ETH at \$1,995	Exploit the price lag by selling high on this platform and buying back on Binance at \$1,980 for a profit
Carol	Ordinary trader	GTC limit buy order — post a bid for 5 ETH at \$1,970	Believing ETH will keep falling, hopes to acquire it at a lower price

Table 9-3. The three core participants and their actions (across the CEX, general-purpose DEX, and Hyperliquid paradigms) (Data source: compiled by the author)

Under the traditional CEX's speed-priority paradigm, Bob holds the latency advantage: his IOC sell order reaches the matching engine before Alice's Cancel instruction and executes 10 ETH at \$1,995. Bob immediately hedges by buying back on Binance at \$1,980, earning a profit of \$150. Alice's cancellation arrives too late; her \$1,995 bid is passively filled at a stale price, buying 10 ETH worth only \$1,980, for a loss of \$150. Bob's profit comes entirely from a physical latency advantage. Under the general-purpose DEX's bribe-priority paradigm, both Bob's IOC sell order and Alice's Cancel are broadcast to the public mempool. An MEV searcher, Dave, spots the opportunity and pays a bribe to the block builder to place his own sell order ahead of Alice's Cancel. Dave sells 10 ETH at \$1,995 into Alice's stale bid, and Alice's cancellation fails because her order has already been filled. Dave profits (net of the bribe), while Alice again loses \$150. Under Hyperliquid's rule-priority paradigm, the three instructions arrive within the same block window. The consensus protocol automatically groups them by category: Alice's Cancel is high priority, while Bob's IOC and Carol's GTC are low priority. The execution order is as follows: Alice's Cancel is executed first, her \$1,995 bid is successfully withdrawn, and she suffers no loss. Bob's IOC sell order cannot execute because the counterparty no longer exists. An MEV searcher's aggressive order would likewise be ranked behind Alice's Cancel, so a sandwich attack is structurally infeasible.

Table 9-4 summarizes the profit-and-loss outcomes for each participant under the three paradigms.

Participant	CEX	General-purpose DEX	Hyperliquid
Alice (market maker)	−$150	−$150	$0
Bob (arbitrageur)	+$150	$0	$0
Carol (ordinary trader)	$0 (not filled)	$0 (not filled)	$0 (not filled)
Dave (MEV searcher)	Not present	+$100 (assumed)	Cannot act
Builder/exchange	Implicit gain	+$50 (assumed)	No additional gain

Table 9-4. Summary of profit-and-loss outcomes for each participant under the three paradigms (Dave exists only in the general-purpose DEX scenario) (Data source: compiled by the author)

This case reveals not merely a technical difference but a fundamental question about who value belongs to: when a price dislocation occurs, to whom should the economic value it generates accrue? The first two paradigms assign it to the fastest participant and to the one best at manipulating ordering, respectively (as in Table 9-2 and the discussion above), whereas the answer of the third paradigm is that this value is captured by no one — it is not transferred away from the liquidity provider (Alice) but remains in the market (in the Hyperliquid column, every party's profit and loss is $0). In other words, Hyperliquid protects liquidity providers from predation rather than redistributing value to market makers. Its design embeds a core assumption: the adverse selection risk borne by market makers constitutes the supply cost of liquidity as a public good, and the protocol should compensate for this structurally.

Crucially, the protection above holds only when Alice's Cancel instruction and Bob's IOC order fall within the same block. If Bob's aggressive order is included in block N while Alice's cancellation falls into block N+1 owing to network latency, then Alice's stale quote is still executed in block N, and the market maker's adverse selection loss is no different from the CEX scenario. The probability of this cross-block failure scenario depends on the network latency between the market maker and the validator nodes and on the timing relationship at the block boundary, which Section 9.3.7 analyzes in detail.

9.1.3 Effects on market makers

The deeper effect of semantic ordering on market makers is that it systematically alters the cost structure of the market-making business. A market maker's equilibrium spread must cover the various costs it faces. According to the classic model framework established by Glosten and Milgrom (1985), these costs consist mainly of adverse selection costs, inventory risk costs, and order processing costs [5]. In the CEX environment, one must add the high infrastructure cost of maintaining a speed advantage. In the general-purpose DEX environment, adverse selection costs are amplified by the presence of MEV.

By granting market makers' defensive instructions (such as cancellations) high priority, semantic ordering directly reduces the adverse selection risk they face. Under normal market conditions, a market maker can, with high probability, successfully cancel before

informed traders exploit its stale quote, thereby protecting itself from loss. However, this protection probability drops significantly during network congestion or near a block boundary, where a cancellation instruction may fall into the next block owing to transmission latency, leaving the market maker exposed to adverse selection in the current block. In its pricing model, a market maker should reserve a buffer for this tail risk rather than treat cancellation priority as unconditional protection. This protocol-level protection means market makers no longer need to invest heavily to compete in the speed race, which significantly lowers infrastructure costs. The decline in these two core costs allows market makers to remain profitable even while quoting narrower spreads. Drawing on the model framework of Glosten and Milgrom [5], Figure 9-2 conceptually presents the decomposition of a market maker's equilibrium spread across the three market environments. In the semantic ordering environment, the substantial reduction in adverse selection and infrastructure costs significantly compresses the total spread, ultimately benefiting all traders.

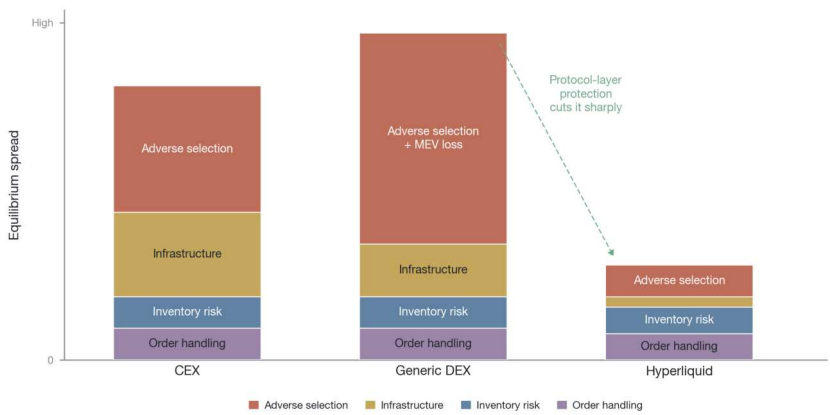


Figure 9-2. Decomposition of a market maker's equilibrium spread across three market environments (conceptual illustration; relative heights are assigned by the author and are not empirical data; the spread-decomposition framework follows Glosten and Milgrom 1985 [5], where [5] supports the adverse selection component, while the infrastructure cost is an extension added in this chapter and is not part of [5]; for the general-purpose DEX, the adverse selection segment is folded into the MEV loss)

As Figure 9-2 shows, the compression of the equilibrium spread in the semantic ordering environment comes mainly from two components: adverse selection costs fall sharply owing to protocol-level protection, and infrastructure costs approach zero as the speed race loses its purpose. Inventory risk and order processing costs differ little across the three environments and are only weakly linked to the ordering mechanism.

9.1.4 Effects on arbitrageurs

Semantic ordering does not eliminate MEV; rather, it structurally redefines and redistributes it. Traditional sandwich-attack and front-running MEV is largely suppressed because aggressive orders have lower priority than market makers' defensive orders. MEV does not vanish entirely, however, but persists in a more constrained form. Because the ordering of transactions within a single priority category is still determined

by the block proposer, the proposer retains a degree of ordering power. This ordering power held by the block proposer within the framework of the protocol's rules is termed *residual MEV* in this chapter; it also corresponds to the *Monarch* category in the {Monarch, Moloch, Mafia} extractable-value framework proposed by Sun (2022) [25] in the MEV literature — the value extracted by a single sequencer by virtue of its ordering monopoly (the term is not official Hyperliquid terminology, nor is it equivalent to *residual MEV* in the generic sense).

MonarchEV is far smaller in scale than traditional MEV and arises mainly in the following scenario: when two arbitrageurs submit identical arbitrage transactions within the same block, the block proposer can choose which executes first. This power can be *sold*, with an arbitrageur paying part of its profit to the block proposer as an *ordering fee*. But because such arbitrage opportunities are themselves greatly reduced by the protection afforded to market makers, the absolute scale of MonarchEV is very limited. It is closer in nature to a theoretical residual risk than to a serious problem in practice. This mechanism pushes high-frequency traders away from latency arbitrage that relies on micro-timing advantages and toward the discovery of cross-market price discrepancies based on information analysis, a behavior pattern closer to traditional cross-market arbitrage (Section 9.2.3 further analyzes the positive effect of this strategic shift on price discovery).

9.1.5 Theoretical foundations

The design of semantic ordering is supported by well-established theory in financial economics. As noted earlier, the model of Glosten and Milgrom (1985) shows that a market maker's spread is directly related to the adverse selection risk it faces [5]. By granting priority to cancellation operations, semantic ordering systematically lowers the probability of adverse selection faced by market makers, thereby allowing them to quote tighter spreads. Viewed through Harris's (2003) framework for analyzing market microstructure, semantic ordering is in essence a major innovation at the level of *trading rules* [6]. It elevates the protection of liquidity providers from a reliance on market makers' own investment (speed) or on ex post compensation in market design to an ex ante guarantee at the protocol level.

The deeper logic of this value judgment is that liquidity is a public good of the market from which all traders benefit, yet its providers (market makers) bear a disproportionate share of the risk. Without protection, market makers would widen their spreads to compensate for this risk, and wider spreads mean that all traders pay higher transaction costs. Protecting market makers at the protocol level is therefore not an unfair tilt but a mechanism design intended to improve the overall efficiency of the market. By lowering market makers' risk, it lowers the spread compensation they demand and ultimately reduces transaction costs for all traders.

However, the systematic suppression of informed traders by semantic ordering may also produce unintended efficiency consequences. The impossibility theorem of Grossman and Stiglitz (1980) holds that if informed trading is unprofitable, the incentive to acquire and process information weakens, and the efficiency with which market prices aggregate information declines accordingly [7]. When semantic ordering makes it difficult to

exploit an informational advantage against a market maker's stale quote, the incentive for arbitrageurs to acquire and transmit new information is also weakened. This is not a design flaw of semantic ordering but an inherent design trade-off: between improving liquidity supply and preserving price discovery incentives, semantic ordering clearly chooses the former. Each ordering paradigm is a design choice made under particular constraints, with its own trade-offs, and should not be placed within a simple linear narrative of *progress*.

Semantic ordering offers a new dimension for market design: it makes explicit the priority of different trading behaviors and hardcodes that priority into the consensus rules. Yet semantic ordering is not the only path to solving the MEV problem, and situating it within the broader range of mitigation approaches helps clarify its distinctive position.

9.1.6 A comparison of MEV mitigation approaches

Current mainstream MEV mitigation approaches can be grouped into several categories according to their technical path. Private-mempool approaches, exemplified by Flashbots Protect and MEV-Share, work by encrypting or privatizing the transaction-submission process to prevent searchers from observing the contents of unconfirmed transactions, thereby eliminating front-running and sandwich attacks that rely on information visibility [8]. Such approaches have been widely adopted in the Ethereum ecosystem, but their limitation is that they do not change the ordering rules themselves: once a transaction is decrypted, its ordering is still decided by the block builder, so MEV is redistributed rather than eliminated. Cryptography-based approaches, exemplified by the threshold-encrypted mempool, use cryptographic means to ensure that transaction contents remain invisible until ordering is fixed, technically eliminating content-based MEV extraction. However, such approaches introduce additional computational and communication latency and cannot treat different transaction types differentially; in their encrypted state, all transactions are treated as equivalent. Intent-based approaches, exemplified by CoW Protocol and UniswapX, abstract a user's trading need into an *intent*, and specialized solvers compete to find the optimal execution path [9]. This approach performs well in swap scenarios, but its applicability is limited to transaction types that can be expressed as outcome constraints and cannot cover the full order-book trading environment.

Hyperliquid's semantic ordering occupies a distinctive position among these mechanisms. Unlike the approaches above, semantic ordering explicitly classifies the economic meaning of transactions at the consensus-protocol level and, on that basis, makes priority decisions that embody a value judgment. This makes it a rare MEV mitigation mechanism among the mainstream approaches surveyed in this chapter that explicitly expresses, at the protocol level, a value orientation about *which market behaviors should be given priority protection*. However, this design requires an application-specific chain as its implementation vehicle and cannot be deployed on a general-purpose L1, because Hyperliquid's L1 blockchain is itself infrastructure custom-built to support semantic ordering. Table 9-5 provides a systematic comparison of these approaches.

Approach category	Representative	Protection mechanism	Scope of MEV reduction	Applicable setting	Limitation
Private mempool	Flashbots Protect	Hides transaction contents	Front-running, sandwich attacks	General-purpose L1	MEV redistributed, not eliminated
Threshold encryption	Shutter Network	Encryption until ordering is fixed	Content-based MEV	General-purpose L1	Added latency, no differential treatment
Intent solving	CoW Protocol	Off-chain competitive solving	Swap slippage	Swap scenarios	Not applicable to the order book
Semantic ordering	Hyperliquid	Type classification at the consensus layer	Front-running, adverse selection	Dedicated order-book chain	Requires an application-specific L1
Distributed ordering rights	dYdX v4	Ordering by validator consensus	Single-point ordering manipulation	Dedicated order-book chain	Semantic priority not explicitly encoded

Table 9-5. Comparison of MEV mitigation approaches (Data source: compiled by the author)

The comparison with dYdX v4 is especially instructive, because, like Hyperliquid, it is a decentralized order-book exchange on an application-specific chain. dYdX v4 mitigates single-point ordering manipulation by distributing ordering rights to validator consensus, but it does not classify different transaction types semantically at the protocol level, and all orders are treated equally when ordered. As a result, dYdX v4's market makers do not enjoy Hyperliquid-style protocol-level cancellation-priority protection, but they also avoid the suppression of informed-trader incentives that semantic ordering entails. The two paths strike different balances between liquidity protection and price discovery efficiency.

The value of surveying these solutions is that it reveals Hyperliquid's design choice to be not without alternatives but a particular philosophical and engineering trade-off made among several viable paths.

9.1.7 The interaction of semantic ordering with the transaction lifecycle

Semantic ordering does not act in isolation on the transaction-ordering stage; it interacts deeply with the other core subsystems of perpetual futures, and these interactions both amplify the protective effect of semantic ordering and, under certain conditions, introduce new dimensions of risk.

At the level of the margin system, the interaction between semantic ordering and liquidation cascades differs between isolated margin and cross margin. Under isolated margin, the liquidation of a single position does not affect other positions, and cascade risk is confined to a single trading pair; under cross margin, the liquidation of one position may render overall margin insufficient and trigger a chain of liquidations across trading pairs. Semantic ordering provides the same ordering protection in both modes, but under cross margin the transmission path of a liquidation cascade is more complex,

and its systemic risk lies beyond what the ordering layer can cover.

At the level of funding rate settlement, the periodic settlement of the funding rate is placed at the system's highest priority (Table 9-1), which means that settlement transactions cannot be front-run by market participants at the ordering level (although, as Section 9.4.3 shows, the external price inputs on which the funding rate and mark price depend can still be manipulated; the protection here is confined to on-chain execution ordering). This design safeguards the integrity of the funding rate mechanism as a price-anchoring tool for perpetual futures, and its detailed workings are analyzed in Chapter 10.

At the level of the liquidation engine, the design choice to place liquidation orders at standard priority produces subtle but important consequences. When a market maker's Cancel instruction and the liquidation engine's forced-closure order coexist within the same block, the Cancel, as high priority, executes before the liquidation order — that is, the market maker can withdraw its quote before the liquidation order seeks a fill. This protects market makers from being forced to absorb liquidated positions at extreme prices, but it may also leave liquidation orders facing thinner order-book depth at execution, degrading the quality of liquidation execution. The JELLY incident is precisely an extreme manifestation of this interaction: when many market makers exercised their cancellation priority before liquidation was triggered, an oversized liquidated position, unable to find sufficient depth on the order book, was transferred to the HLP Vault, as analyzed in depth in Section 9.4.3. These interactions show that the effects of semantic ordering cannot be assessed in isolation from the broader contract-settlement system in which it is embedded. The design details of the liquidation mechanism and the margin system are discussed in Chapter 11 and Chapter 10, respectively.

The effectiveness of this macro-level framework depends heavily on its most contested concrete implementation in practice: cancellation priority.

9.2 The cancellation priority mechanism

Cancellation priority is the most directly consequential concrete rule within the semantic ordering framework. The rule stipulates that, within a block's processing cycle, every instruction that seeks to cancel an existing order has an execution priority absolutely higher than that of every new instruction that seeks to fill those orders. This design gives market makers a right of last look enforced by the consensus protocol.

In traditional financial markets, last look is a market maker's final opportunity to accept or reject a client order in light of the latest market information. This power has been widely questioned for its opacity and potential conflicts of interest. Hyperliquid, however, fundamentally reworks this concept. Here, last look is not a market maker's private privilege but a public, universal rule hardcoded into the consensus protocol. All market participants are fully aware of the rule's existence and can adjust their strategies accordingly. This complete transparency means cancellation priority is no longer an unfair advantage but a known initial condition that reshapes the market game.

9.2.1 Hedging adverse selection

To understand the effect of cancellation priority, one must return to the core market-microstructure concept analyzed in Section 7.1.3: adverse selection. A market maker's core business is to quote continuously under uncertainty (its strategy framework is given in Section 7.7.1); its profit comes from the bid-ask spread, while its greatest risk comes from trading against better-informed *informed traders*. When external information causes the true value of an asset to change, informed traders can arbitrage against a market maker's stale, not-yet-updated quote, and the loss this imposes on the market maker is the adverse selection cost. The classic model of Glosten and Milgrom (1985) systematically shows that a market maker's quoted spread must be wide enough to cover the expected loss it incurs when trading against informed traders [5].

In the CEX and general-purpose DEX environments, market makers have limited means to combat adverse selection: they mainly rely on investing in faster technical systems to cancel first, or on widening their quoted spread to pass the risk cost on to all traders up front. The cancellation priority mechanism offers a third solution: at the protocol level, it compresses the probability of within-block adverse selection under normal market conditions to a residual level near zero (with risk remaining only in a narrow window near the block boundary). In essence, it establishes a layer of structural separation, enforced by the consensus protocol, between the aggressive orders of informed traders (such as arbitrageurs) and the defensive cancellation instructions of market makers.

Figure 9-3 compares the execution outcomes of the same set of trading instructions in the CEX and Hyperliquid scenarios.

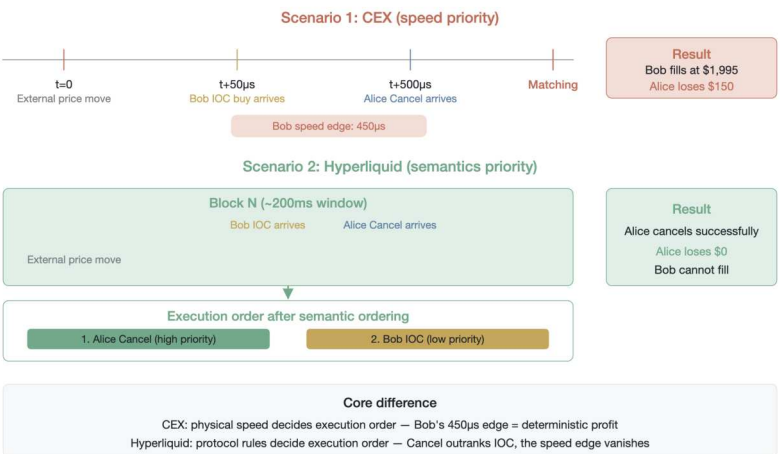


Figure 9-3. Comparison of execution timelines for the CEX and Hyperliquid (hypothetical worked example, not empirical data; $1,995, $150, 450 μs, and approximately 200 ms are all illustrative values; the mechanism by which the Cancel precedes the IOC is grounded in Section 9.1.2 and [5])

When the external market price changes, market maker Alice and arbitrageur Bob send instructions to the system at the same time. Even if Bob's instruction physically arrives first, as long as it falls within the same block as Alice's cancellation instruction,

the protocol's consensus rules force Alice's cancellation to be processed first. This means the protocol gives Alice a protection window of about 200–400 ms (one block time), allowing her to withdraw a stale quote before it is executed. This mechanism significantly lowers a market maker's adverse selection cost, enabling it to provide liquidity at narrower spreads over longer periods.

9.2.2 A comparison with the IEX speed bump

The design philosophy of cancellation priority is often compared with the speed bump of the IEX exchange. IEX's design, popularized by Michael Lewis in his book *Flash Boys* [10], imposes a physical delay of about 350 microseconds on all orders entering its matching engine by routing them through 38 miles of coiled fiber-optic cable, with the aim of eliminating the speed differential between high-frequency trading (HFT) firms and other market participants. Although the two share a similar goal — creating a fairer market environment — they differ fundamentally in philosophy and mechanism (Table 9-6).

Dimension	IEX speed bump	Hyperliquid cancellation priority
Mode of action	Physical layer: delays all orders indiscriminately.	Protocol layer: orders differentially according to the economic meaning of transactions.
Philosophy	Procedural fairness: attempts to have all participants reach the finish line *simultaneously*.	Outcome-oriented: prioritizes protecting the interests of liquidity providers.
Protective effect	Limited: can bridge only microsecond-level speed differences.	Provides strong protection within one block time (about 200–400 ms) (with a narrow gap at the block boundary).
Market impact	Fairly controversial; some studies argue it harms price discovery efficiency.	Explicitly redistributes value from predatory arbitrageurs to market makers.

Table 9-6. Cancellation priority versus the IEX speed bump (Data source: compiled by the author)

The speed bump is a form of *physical slowdown* that attempts to patch a flawed rule system through technical means. Cancellation priority, by contrast, is a *change of rules* that fundamentally alters the rules of the game. The behavioral difference between the two under liquidity-stress scenarios is especially critical: IEX's delay is symmetric, applying equally to all orders (including cancellations), so during a liquidity contraction the rate at which market makers withdraw slows in step with the rate at which takers consume, and the net change in liquidity is relatively gradual. Hyperliquid's mechanism, by contrast, is asymmetric: only cancellations receive priority acceleration, and in a liquidity-contraction scenario, where cancellations far outnumber new resting orders, net liquidity may drop sharply within a single block. This asymmetry strengthens liquidity supply in normal markets but may accelerate liquidity evaporation during periods of stress. The D-Limit order type that IEX later introduced also began to incorporate a semantically aware design philosophy, indirectly corroborating the need

to solve the ordering problem at the level of protocol rules.

9.2.3 Structural effects

The cancellation priority mechanism has profound structural effects on the market ecosystem, the most direct of which is a fundamental shift in the dimension along which market makers compete. In the CEX, a two-dimensional competition of *speed plus model* dominates, whereas in Hyperliquid the importance of speed is greatly diminished and the competitive focus shifts to *model quality* itself. This significantly lowers the barrier to entry for market making, enabling small quantitative teams that lack the capacity to invest in top-tier infrastructure to compete, thereby increasing the diversity of liquidity supply in the market. By comparing market makers' cost structures across the two environments, Figure 9-4 illustrates this structural shift from infrastructure investment to competition on model quality.

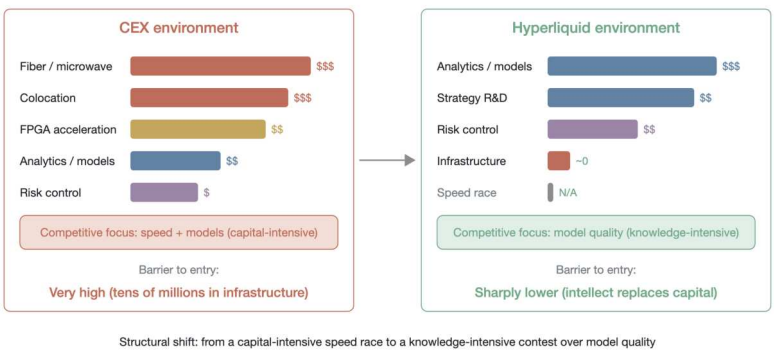


Figure 9-4. Comparison of market makers' cost structures across trading environments (conceptual illustration, not empirical data; $/$$/$$$, ~0, and N/A are qualitative cost tiers, not measured monetary values; maps to Sections 9.2.3 and 9.5.1)

Figure 9-4 reveals a structural shift: infrastructure investment falls from a dominant position in the CEX environment to a marginal level in the Hyperliquid environment, while model development and data analysis take its place as the core cost items. The direct consequence is that the capital barrier to entry for market making falls while its intellectual barrier to entry rises.

The mechanism also forces high-frequency takers to shift their strategies from latency arbitrage toward information-driven approaches. The traditional strategy of sniping stale quotes becomes unprofitable, forcing these traders toward strategies that rely more on information analysis and predictive capability — for example, predicting market makers' behavior patterns, or acting as a price-transmission intermediary across exchanges to promote price convergence, rather than arbitraging price lags.

At the level of spread structure, the effect of cancellation priority is equally significant. By lowering adverse selection and infrastructure costs, market makers can remain profitable even while quoting narrower spreads. Empirical data show that, during calm market periods, the spreads on Hyperliquid's major trading pairs are typically compa-

rable to, or even better than, those of top-tier CEXs (the measurement basis for this comparison and the static-versus-effective-depth qualifications are detailed in Section 9.5.4). However, the cost of these narrow spreads reveals a *liquidity paradox*. The research framework of Foucault, Kadan, and Kandel (2005) on the resilience of limit order markets helps in understanding this phenomenon [11]: cancellation priority lowers the option cost of a market maker's resting order, and under normal market conditions this incentive leads market makers to provide tighter spreads and deeper order books; but the same mechanism also means that, during periods of market stress, market makers can withdraw all of their liquidity with frictionless speed. The prospect theory of Kahneman and Tversky (1979) offers a behavioral explanation: when market volatility spikes, loss aversion drives market makers to exercise their cancellation right immediately rather than wait and assess [12]. Because the execution cost of cancellation (including its psychological cost) is reduced to near zero at the protocol level, panic withdrawal may occur faster and more completely on Hyperliquid than in the CEX environment. This constitutes a procyclical effect: the very mechanism that improves liquidity in normal times accelerates its evaporation in a crisis. Potential mitigations include introducing minimum market-making obligations or imposing a cancellation cooldown during extreme volatility, but these options are in tension with the core design philosophy of cancellation priority.

For ordinary traders, the net effect is positive. They enjoy lower transaction costs, and the risk of their market orders being sandwiched is significantly reduced. But they must also adapt to a new market characteristic: some of the liquidity displayed on the order book may be apparent liquidity — the *depth illusion* defined in Section 8.2 appears in a new form under the cancellation priority mechanism, in that when they attempt to execute a large order, this liquidity may vanish as market makers exercise their cancellation priority. For institutional traders who need to execute large orders, the complete transparency of the on-chain environment poses a structural challenge to execution quality. Traditional execution algorithms such as the time-weighted average price (TWAP), the volume-weighted average price (VWAP), and iceberg orders face a fundamental dilemma in a fully transparent on-chain environment: the size, timing interval, and direction of every child order are visible in real time, allowing competitors to infer the total size and execution strategy of the parent order. The optimal-execution framework of Almgren and Chriss (2001) shows that information leakage increases the permanent price-impact cost [13]: when market participants can anticipate the later portions of a large order, they adjust their quotes in advance, causing execution prices to deteriorate systematically. This transparency cost partly offsets the spread-level benefit that cancellation priority brings to institutional traders.

The effectiveness of the cancellation priority mechanism depends on a key structural foundation: the discrete-time window inherent to on-chain matching (about 200–400 ms). It is this window that forcibly reduces the resolution of the speed race from the nanosecond to the hundred-millisecond scale, making rules rather than speed the dominant factor in ordering. Its structural origin as an inevitable consequence of on-chain matching, and the mechanism-design space it opens up, are detailed in Section 9.3.

9.3 Discrete-time matching: an inherent property of on-chain matching

Semantic ordering and cancellation priority define the rules for *who should be processed first*, but these rules can take effect only if the matching engine can *collect* multiple transactions within a time window before ordering them together. Discrete-time matching provides this structural foundation. Discrete-time matching is not a mechanism invented by Hyperliquid but an inherent structural feature of any on-chain matching system: a blockchain reaches consensus and processes transactions in units of blocks, so once a matching engine is placed on-chain, it must process transactions in batches by block, switching from continuous time to a discrete-time window (about 200–400 ms). This structural feature is not itself complex, but Hyperliquid astutely recognized the mechanism-design space it creates and built semantic ordering and cancellation priority on top of it, fundamentally transforming the market's speed game.

9.3.1 Continuous- versus discrete-time games

To understand the economic consequences of discrete-time matching, one must first examine its opposite: the inherent problems of continuous-time matching. Budish, Cramton, and Shim (2015) argue that the continuous-time limit order book inherently gives rise to a speed race, because under this mechanism even a microsecond-level speed advantage can be converted into a deterministic profit [14]. This channels vast social resources into a *zero-sum* or *negative-sum* race to invest in infrastructure — building dedicated microwave towers, laying undersea fiber-optic cables — and ultimately concentrates the market ever more tightly among a few top-tier HFT firms.

> In traditional financial markets, the communication latency between the Chicago Mercantile Exchange (CME) and the New York Stock Exchange is a classic arena of the speed race. HFT firms invested hundreds of millions of dollars to build microwave-tower networks solely to shorten data-transmission latency along the Chicago–New York corridor from about 6.0 ms to about 3.8 ms (data from roughly 2014–2018; the latency of the most advanced microwave networks today may be lower). In a continuous-time *winner-take-all* game, this roughly 2.2 ms advantage can generate tens of millions or even hundreds of millions of dollars in profit each year, because the participant that reacts first to a price change captures almost all of the arbitrage opportunities. Hyperliquid's block time of about 200–400 ms is precisely where the economic essence of discrete-time matching lies: it transforms the winner-take-all game of continuous time into equal competition within a discrete-time window — as long as instructions fall within the same block, speed differences within the block no longer generate economic value at the protocol level (a quantitative worked example is given in Section 9.3.5).

Hyperliquid's L1 uses the HyperBFT consensus algorithm, a derivative implementation of the HotStuff family. Under normal network conditions, HyperBFT's block latency is about 200 ms; under high load (as during the JELLY incident), block time may extend to 400 ms or longer owing to mempool congestion and increased communication latency among validators. The 200–400 ms range of block times reflects precisely this load-dependent dynamic. Hyperliquid's matching mechanism operates in discrete time, with the block as its basic unit. The processing flow for each block (as shown in Figure

9-5 [3]) is uniform: collect transactions from the mempool, order them according to the semantic rules, match them one by one within the ordered sequence under a price-first, time-second principle, and finally update the order book state all at once and commit the block. This batch-processing model has significant microstructural consequences.



Figure 9-5. Hyperliquid's block-processing flow (mechanism illustration; the flow follows Hyperliquid documentation [3]; the *priority tiers: System/High/Standard* shown in the figure are an interpretive reconstruction in this chapter — what the documentation [3][4] explicitly records is only *cancellations and post-only orders ahead of GTC/IOC taker orders*, and placing mark price and funding rate updates and liquidation in separate ordered tiers is an interpretive reconstruction made in this chapter for analytical convenience rather than the documentation's own text; see the note to Table 9-1)

Figure 9-5 reveals an important design feature: semantic ordering takes place after transaction collection and before sequential matching — a three-stage flow of *classify, then order, then match* — which is the concrete way in which semantic ordering and discrete-time matching operate in concert.

9.3.2 The displacement of speed competition

The discrete-time matching mechanism displaces the dimension of the speed race. Within a block, the execution order of all transactions of the same type is determined by the order in which the block proposer receives them; but because the mempool typically buffers several blocks' worth of transactions, most transactions wait briefly in the pool before being packaged. This greatly dilutes the *who sends first* speed advantage within a block. Speed competition is not eliminated entirely, however, but is displaced to the level between blocks. If a participant's transaction can enter the mempool earlier than others', it is more likely to be packaged into an earlier block. The speed race therefore falls from the *microsecond-level* competition of the CEX to the hundred-millisecond-level competition of Hyperliquid.

This between-block speed competition is especially pronounced near the *block boundary*. As a block's packaging deadline approaches, a residual speed competition persists within a fuzzy interval of about 10–20 ms. Even so, compared with the 1–10 microsecond competition window of the CEX, Hyperliquid has reduced the intensity of speed competition by about three orders of magnitude.

9.3.3 A comparison with frequent batch auctions

Although Hyperliquid nominally still uses a price-time-priority central limit order book (CLOB) mechanism (its structure and rules are covered in Section 8.1.1), its discrete-time batch-processing property gives it some of the characteristics of the frequent batch auction (FBA) introduced in Section 8.8.3. The core idea of the FBA is to collect all orders within a time window and then execute all matchable orders at a single uniform clearing price, thereby eliminating the speed advantage entirely in theory [14]. Hyperliquid does not adopt a uniform clearing price but continues to use price-time-priority sequential matching within the block. Within a single priority category, transactions are arranged in the order in which the block proposer receives them from the mempool, effectively forming a first-in-first-out (FIFO) ordering. However, because the block proposer retains final control over the received ordering, this constitutes a concrete dimension of MonarchEV. This within-priority ordering rule gives rise to a key difference from the FBA.

> Suppose that within a block window the order book has sell orders at $2,002, $2,005, and $2,010, and that market buy orders from three buyers — Alice, Bob, and Carol — arrive simultaneously, with total demand less than or equal to the total supply at or below the $2,005 price level. Under the FBA mechanism, all buyers execute at the uniform clearing price of $2,005, and their arrival times are entirely irrelevant. Under the Hyperliquid mechanism, by contrast, the earliest-arriving buyer, Alice, first executes a portion at $2,002 and then the remainder at $2,005, obtaining a better average execution price. The later-arriving buyers, Bob and Carol, can execute only at $2,005. In this scenario, Alice's speed advantage — even if only at the millisecond level — brings her an economic benefit. Moreover, the block proposer can choose which of Bob and Carol executes first, which constitutes one source of residual MEV.

The differences in pricing mechanism, residual speed advantage, and MEV risk distribution revealed by this case can be compared systematically in Table 9-7.

Dimension	Hyperliquid block matching	Theoretical FBA
Pricing mechanism	Sequential matching within the block (price-time priority)	Uniform clearing price
Speed advantage	Greatly weakened but not fully eliminated	Fully eliminated in theory
MEV risk	Residual MonarchEV present (ordering power)	None in theory

Table 9-7. Key differences between the Hyperliquid and FBA mechanisms (Data source: compiled by the author)

Hyperliquid's design can be seen as a balance between the theoretical goal of the FBA and the practical constraints of the CLOB. It largely achieves the FBA's main goal (greatly weakening the speed race) while retaining the user experience of a traditional order book and a relatively simple implementation logic.

Behind this mechanistic difference lies a deeper divergence in design philosophy. Hyperliquid and the FBA share a common idea — both counter the speed race by discretizing time [14] — but they differ fundamentally in implementation path and philosophical

orientation. The core of the FBA is a uniform clearing price, so that all trades within the auction window execute at the same price, eliminating the incentive to front-run. Hyperliquid, by contrast, remains a price-time-priority order book: within a block, transactions execute sequentially by semantic priority and price level, and the price can move. Hyperliquid absorbs the core idea of the FBA but combines it with the traditional order-book mechanism and the innovative rule of semantic ordering. Hyperliquid differs structurally from the FBA along the dimensions of matching mechanism, price formation, ordering rules, and HFT-suppression strategy, and Table 9-8 provides a systematic comparison.

Dimension	FBA	Hyperliquid discrete-time matching
Matching mechanism	Batch auction, uniform clearing price	Sequential matching, batch processing
Price	Only one clearing price per batch	Price can vary continuously within a batch
Ordering rule	No ordering needed (or by price only)	Semantic ordering takes priority
Suppression of HFT	Eliminates front-running	Eliminates front-running but preserves room for other strategies

Table 9-8. Discrete-time matching versus the FBA (Data source: compiled by the author)

From a theoretical perspective, the core logic by which Budish et al. [14] argue for the FBA is that it breaks the *prisoner's dilemma* of the speed race. The discrete-time property inherent to on-chain matching achieves the same effect, and Hyperliquid adds to it a dimension the FBA lacks: semantic ordering. In the FBA, all participants are entirely equal within each auction window; in Hyperliquid, market makers' cancellation instructions are explicitly given higher priority. This is an important design difference: the FBA pursues *procedural equality*, whereas Hyperliquid pursues *outcome optimization*, explicitly placing the interests of liquidity providers above those of liquidity consumers. Hyperliquid can therefore be seen as a more explicit, deliberately weighted, and pragmatic implementation of the FBA idea in a blockchain-native environment.

9.3.4 Synergy with semantic ordering

The discrete-time property of on-chain matching and semantic ordering are mutually reinforcing complements, and their combination forms the foundation of Hyperliquid's microstructure. Discrete time creates a *batch window* that provides the basis for implementing semantic ordering. In continuous-time matching, orders are processed one by one in order of arrival, making it impossible to establish priority relationships among different order types. Conversely, semantic ordering solves an inherent problem of discrete time: within a batch window, absent explicit ordering rules, the block proposer would have complete freedom to order all transactions, which would lead to unconstrained MEV extraction. Through predefined priority rules, semantic ordering strictly limits the block proposer's discretion, compressing the MEV space to the single dimension of *ordering within the same category.*

Neither element can stand without the other. Without discrete time, semantic ordering cannot be implemented in continuous time; without semantic ordering, discrete time would degenerate into a pure block-space auction, leading to uncontrolled block-proposer MEV. Other decentralized exchanges share the structural feature of discrete time but have not converted it into a microstructural advantage. Hyperliquid's innovation lies in recognizing the mechanism-design space embedded in the discrete-time window inherent to on-chain matching and, on top of it, building semantic ordering and cancellation priority, thereby creating a self-consistent microstructure with few direct precedents in the known practice of either traditional finance or crypto finance.

Figure 9-6 presents the synergy between discrete-time matching and semantic ordering and its effect on the market ecosystem.

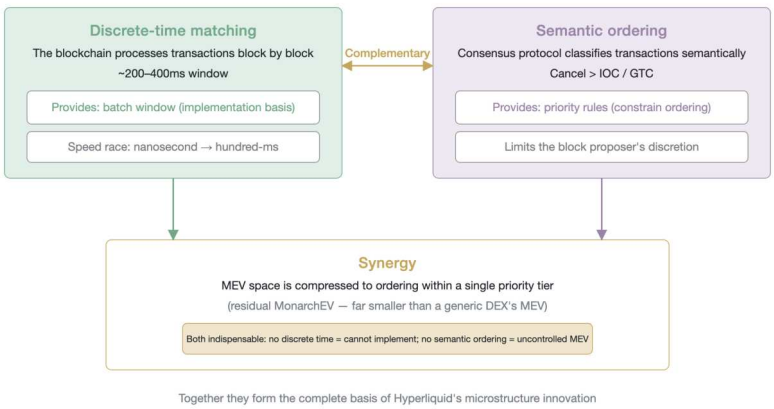


Figure 9-6. The synergy between discrete time and semantic ordering (conceptual illustration, not empirical data; MEV is compressed to ordering within a single priority category — that is, residual MonarchEV, a term drawn from the {Monarch, Moloch, Mafia} extractable-value framework of Sun 2022 [25] and not official Hyperliquid terminology)

Figure 9-6 shows that discrete time creates the batch window that makes semantic ordering implementable, while semantic ordering constrains the block proposer's discretion; the interaction of the two compresses the MEV space to the single dimension of ordering within a priority category, whose scale and harm are far smaller than the unconstrained MEV extraction seen on a general-purpose DEX.

9.3.5 The weakening of the speed race

In the CEX's continuous-time matching environment, a speed lead is measured in nanoseconds (billionths of a second), and HFT firms invest enormous sums to achieve it. In Hyperliquid's discrete-time matching environment, by contrast, the smallest unit of time is one block time (a few hundred milliseconds). This means that whether a trading instruction arrives with a latency of 1 microsecond or 100 ms, as long as it falls within the same block window it is treated as *equal* along the time dimension. What determines execution priority is no longer physical speed but the protocol's semantic rules.

This compression of time granularity from nanoseconds to milliseconds renders the traditional speed advantage of HFT almost entirely ineffective. Drawing on the theoretical framework of Budish et al. [14] and Hyperliquid's technical documentation [3], Figure 9-7 compares the effective competition time windows of different trading platforms.

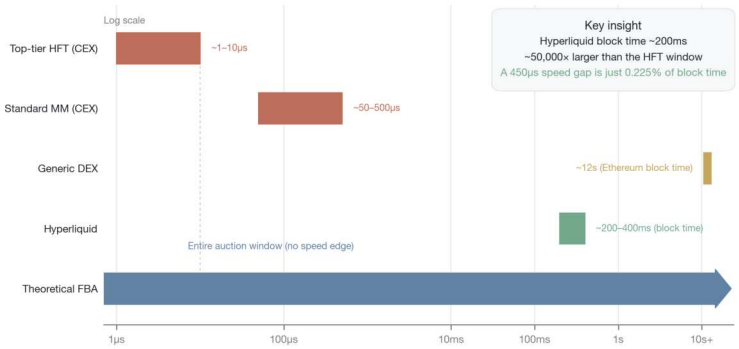


Figure 9-7. Comparison of effective competition time windows across trading platforms (representative parameters, not an external dataset; logarithmic time axis; Ethereum's block time of about 12 s and Hyperliquid's of about 200–400 ms follow Hyperliquid documentation [3] and refer to block production and confirmation rather than the roughly 70 ms single-block finality; the 1–10 μs for top-tier HFT and 50–500 μs for standard market making are industry orders of magnitude; the theoretical FBA spanning the entire window follows Budish et al. 2015 [14]; *about 50,000 times the roughly 4 μs window of top-tier HFT* is an arithmetic inference)

A simple model can quantify this effect. Suppose that on a CEX, HFT firm A has a latency of 50 microseconds and market maker B has a latency of 500 microseconds; A is 450 microseconds faster than B, a gap sufficient for A to capture B's stale quote first on every price change. But on Hyperliquid, where the block time is 200 ms (200,000 microseconds), the 450-microsecond latency difference between the two is only 0.225% of the block time. As long as the instructions fall within the same block, this difference is irrelevant. The *resolution* of the speed race is forcibly lowered from the nanosecond to the hundred-millisecond scale, substantially equalizing the competitive conditions among participants and shifting the competitive focus from infrastructure speed to model quality and information-analysis capability.

> On January 11, 2024, the U.S. Bureau of Labor Statistics released higher-than-expected consumer price index (CPI) data, setting off a chain reaction in the markets. At t = 0 the data are released; at t + 5 ms, algorithmic trading systems in the Chicago futures market react first and BTC futures prices fall; at t + 15 ms, top-tier HFT market makers on Binance begin adjusting their quotes; at t + 50 ms, spot prices on Binance have finished adjusting. Between t + 50 ms and t + 200 ms, market makers on Hyperliquid detect the price change on Binance and send Cancel instructions. At the same time, arbitrageurs also detect the spread and send IOC buy orders in an attempt to attack the stale quotes. The instructions of both fall within the same block. At t + 200 ms, the block is processed. Under the cancellation priority rule, the market makers' cancellation instructions execute before the arbitrageurs' buy orders. The market makers successfully withdraw their stale quotes and avoid loss, while the arbitrageurs' buy orders fail for lack

> of a counterparty. This case reveals a core characteristic of Hyperliquid's microstructure: its price discovery carries a structural lag of about 150–200 ms, giving it a price-following relationship to external markets. It is precisely this lag window that gives market makers ample time to send Cancel instructions, and the cancellation priority rule ensures that these instructions execute ahead of an attack, so that discrete time and cancellation priority form an effective synergy here. At the block boundary (for example, around t + 200 ms), a brief, temporary absence of liquidity may occur, because the market makers' old quotes have been withdrawn while their new quotes have not yet been confirmed on-chain.

9.3.6 Block time as a hyperparameter

The preceding analysis has treated the 200–400 ms block time as a given system parameter, but from a mechanism-design perspective, block time is an adjustable hyperparameter whose value directly determines the trade-off surface between the strength of market-maker protection and price discovery efficiency. When block time is short (for example, close to 50 ms), fewer transactions are collected within each block window, the reaction time available to market makers is compressed, the probability that a Cancel instruction lands in the same block as an aggressive order declines, adverse selection risk partly returns, market makers are forced to quote wider spreads to compensate, and the system's speed-competition characteristics begin to approach those of the CEX. When block time is long (for example, more than 2 seconds), market makers gain an ample protection window and adverse selection risk falls further, but the price-following lag widens in step. Within each block cycle, order-book quotes drift steadily away from external market prices, so market makers must widen their spreads in advance to cover this quote-staleness risk. The risk arises from within-block price drift, not from any need to hold positions longer, since market makers can still cancel and re-quote each block. The equilibrium spread therefore traces a U-shaped curve: when block time is too short, adverse selection costs push the spread up; when block time is too long, inventory risk and price staleness push the spread up; and in the middle region there is an optimal block time that minimizes the total spread.

Hyperliquid's current choice of a 200–400 ms block time can be located within this framework. Compared with other L1s, Ethereum's 12-second block time offers ample room to implement semantic ordering but too large a price-following lag; Solana's block time of about 400 ms is close to Hyperliquid's but does not introduce semantic ordering; and the subsecond block times of layer-2 (L2) chains such as Arbitrum likewise lack differential treatment of transaction types. Hyperliquid's distinctiveness lies in combining a relatively short block time with semantic ordering, seeking to operate near the low point of the U-shaped curve. The U-shaped curve above is a qualitative description, and its quantitative calibration faces significant challenges. The parameters of the FBA model of Budish et al. [14] are estimated from the volatility characteristics and participant structure of U.S. equity markets, and extending them directly to the crypto perpetual futures market would require recalibration to markedly different volatility distributions, leverage structures, and participant compositions. The optimal block time may differ significantly across trading pairs — for example, the difference in volatility characteristics between BTC perpetual futures and long-tail tokens implies

that their optimal points on the U-shaped curve may lie far apart — yet Hyperliquid currently applies a uniform block time to all trading pairs. As the underlying consensus technology advances, block time may shorten further, which would improve the price discovery lag but would also compress the protection window provided by cancellation priority; the dynamic balance between the two requires ongoing empirical calibration.

9.3.7 Boundary effects and MonarchEV

Discrete-time matching has inherent limitations, one distinctive risk being the block-boundary effect. Because transactions are processed in discrete blocks, transactions that happen to fall near the boundary between two blocks may have sharply different outcomes. As shown in Figure 9-8, market maker Alice sends a Cancel instruction 1 ms before block N's deadline tN; but if network latency delays that instruction so that it reaches the validator node 1 ms after tN, it is assigned to block N+1. During the processing of block N, Alice's stale quote is still active and may be filled by an arbitrageur within that block.

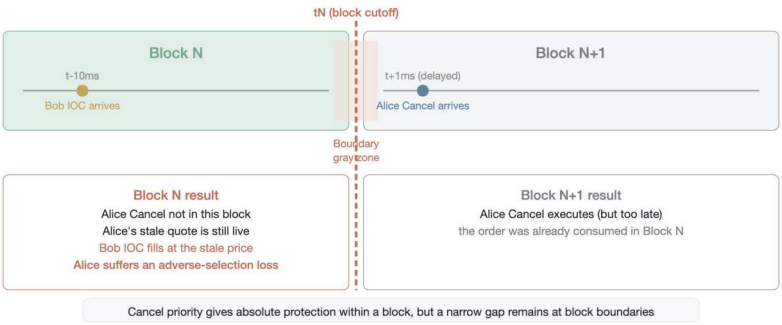


Figure 9-8. The block-boundary effect (hypothetical worked example, not empirical data; t − 10 ms and t + 1 ms are illustrative timestamps showing that cancellation priority provides near-absolute protection within a block but leaves a narrow gap at the block boundary; maps to Section 9.3.7)

This means that although cancellation priority provides near-absolute protection within a block (conditional on the cancellation being submitted and received by the proposer within the same block), a narrow protection gap remains at the block boundary, its width depending on fluctuations in network latency and typically ranging from a few milliseconds to a few tens of milliseconds. Market makers must therefore still attend to the network latency between themselves and the validator nodes, but the aim is to ensure that instructions reliably land in the intended block rather than to compete within a block — the competitive focus shifts from a *within-block speed race* to a *cross-block reliability race*, whose investment scale is far smaller than that of the traditional speed-infrastructure race.

Moreover, discrete-time matching does not eliminate MEV entirely. HyperBFT selects block proposers through a deterministic leader-rotation mechanism, which means that the identity of the next block's proposer is technically predictable. MEV searchers can exploit this predictability by submitting priority-ordering requests directly to the

validator about to produce a block. Within a block, the power to order transactions of the same priority (for example, two IOC orders) still rests with that block's proposer — the MonarchEV defined in Section 9.1.4. Although this MEV is far smaller in scale and harm than the MEV on a general-purpose DEX, it remains a concentration of power worth attending to within the Hyperliquid ecosystem.

The discrete-time property inherent to on-chain matching is the structural foundation of Hyperliquid's microstructure innovations. By switching time from continuous to discrete, the blockchain fundamentally changes the market's competitive paradigm: it not only greatly weakens the advantage of traditional HFT but also provides the necessary conditions for implementing the two mechanism innovations of semantic ordering and cancellation priority. The combination of the discrete-time property with these two mechanism innovations forms a complex synergistic system whose inherent trade-offs and risk boundaries equally merit examination.

9.4 Synergies and trade-offs

The power of the three-layer mechanism — semantic ordering, cancellation priority, and discrete-time matching — arises not from any layer operating in isolation but from the complex system they form in concert. The discrete-time property inherent to on-chain matching divides the continuous flow of time into discrete block windows, creating the necessary conditions for semantic ordering within a batch; semantic ordering, in turn, defines a non-economic, rule-based priority for the flow of transactions within that window, at the core of which is granting cancellation operations priority. Acting together, this layered structure significantly alters the market's microdynamics but also introduces new trade-offs and risk boundaries. This synergy ultimately drives a positive feedback loop aimed at strengthening the market's core function — liquidity: a self-reinforcing transmission chain running from market-maker protection to narrower spreads, greater trading volume, and the entry of more market makers, whose complete chain is laid out in Section 9.4.8 (Figure 9-9). In practice, however, this self-reinforcing loop faces multiple constraints, and its inherent trade-offs and risk boundaries merit examination one by one.

9.4.1 Transparency effects

One of Hyperliquid's core value propositions is its high degree of transparency, and the tension between *verifiability and strategy privacy* analyzed in Section 7.5 is concentrated here. All trade histories, order book snapshots, and even the protocol's ordering rules themselves are recorded on-chain and can be audited by anyone. This design eliminates the *black box* problem of centralized exchanges and establishes trust grounded in verifiable facts. However, this complete transparency also carries an asymmetric side effect: in eliminating one form of information asymmetry, it creates a new dimension of information competition, giving rise to a race in analytical capability.

Because all data are publicly available, quantitative analysis teams can use sophisticated algorithms to *reverse-engineer* the trading strategies of other market participants, market makers in particular. By analyzing a particular market maker's quoting behav-

ior, inventory management, and cancellation patterns under different market conditions, one can infer the key parameters of its model. These reverse-engineered strategy characteristics are called a *strategy fingerprint*. Having captured a rival's strategy fingerprint, a competitor can develop targeted countermeasures — for example, slightly improving on the rival's quote after it posts — and thereby systematically capture its order flow.

> A small quantitative market-making team, AlphaQuant, deployed an innovative market-making model on Hyperliquid and performed excellently at first. Months later, however, its profits fell sharply. Investigation revealed that a competitor had fully replicated its core algorithmic model by analyzing all of its public on-chain trade data. The competitor had captured AlphaQuant's *behavioral fingerprint* — for example, its cancellation thresholds and spread-adjustment patterns at particular volatility levels — and developed dedicated countermeasures on that basis. Because of Hyperliquid's transparency, AlphaQuant's strategic edge was rapidly eroded. In response, the team had to develop *strategy obfuscation* techniques, such as adding random noise to its quotes and randomizing cancellation delays, but this sacrificed a certain proportion of its market-making efficiency. Strategy obfuscation necessarily reduces signal precision in theory, because introducing noise also disturbs the market maker's own optimal quoting decisions; the exact magnitude of the efficiency loss, however, depends on the obfuscation parameters and market conditions, and public empirical measurements are still lacking. Although this analytical competition does not consume physical resources the way the CEX speed race does, it likewise raises the barrier to market making and poses a new challenge to small participants. This reveals an inherent contradiction: cancellation priority lowers the speed barrier to market making, but complete transparency raises the barrier of strategic complexity. The deeper risk is that on-chain transparency exposes not only individual strategies but also makes coordinated group behavior possible. The complete observability of market makers' positions and the HLP Vault's holdings, amplified by social media, can catalyze coordinated behavior against a particular participant; the JELLY attacker analyzed in Section 9.4.3 exploited precisely the public visibility of the HLP Vault's risk exposure to design its attack path.

9.4.2 The price discovery lag

Hyperliquid's mechanism design makes an explicit trade-off between *protecting liquidity* and *promoting price discovery*, choosing the former. Cancellation priority and discrete-time matching enable market makers to provide liquidity safely, but they also reduce market makers' incentive to quote proactively and push prices toward a new equilibrium level. Market makers increasingly favor a passive quoting strategy: quoting off the price of a mature external market such as Binance plus a fixed spread, and using cancellation priority to protect themselves when the external price moves, rather than actively predicting the direction of prices.

This pattern gives Hyperliquid's price discovery process a structural dependence: its price updates follow the price changes of external markets passively and with some lag (one block time). This inefficiency in price discovery is not a technical flaw but a cost paid to protect market makers in exchange for greater liquidity. Quantitatively, this lag is about one block time (200–400 ms) under normal market conditions [4]. During sharp market swings, because rising uncertainty may prompt market makers to withdraw from the market collectively, order-book depth can drop abruptly, and the

lag in price adjustment may widen to several block times. This structural dependence on external markets is the inherent fragility of its price-following posture. In systemic deleveraging events (such as the global liquidity crisis of March 2020 or the FTX collapse of 2022), Hyperliquid faces not merely a one-block-time price-following lag but a second-order risk: when several CEXs experience liquidity droughts simultaneously, the price reference source itself becomes unreliable. Hyperliquid's oracle price (the core input to the mark price) is a weighted median of external CEX spot prices, on top of which the mark price adds components such as an order-book exponential moving average (EMA); if these CEXs exhibit significant price deviations or trading halts under extreme stress, the anchoring function of the oracle price and mark price fails simultaneously, potentially triggering improper liquidations or erroneous funding rate settlement.

9.4.3 The JELLY incident: attack chain and mechanism analysis

Any rule system, however finely designed, can be exploited by participants in unexpected ways. Kirilenko et al.'s (2017) study of the 2010 *flash crash* shows that even the rules of a mature market can be exploited in unintended ways under extreme conditions [15]. The JELLY incident of March 26, 2025, stress-tested Hyperliquid's rule system; its attack path is closely related to the liquidation-cascade transmission mechanism analyzed in Section 8.3.1 and exposed systemic fragilities lying beyond microstructural protection.

JELLY was, at the time, a newly listed low-liquidity token on Hyperliquid, with an average daily trading volume of only a few million dollars on external spot exchanges. The timeline of the attack can be reconstructed as follows. First, the attacker created three independent on-chain accounts (the absence of any linked-account detection meant this step triggered no risk-control alerts) and deposited a total of about $7.17 million in USDC [16]. Second, the attacker established hedged positions across the three accounts: two accounts opened JELLY long positions totaling about $4.05 million, while the third opened a JELLY short position of about $4.1 million, whose effective leverage was such that a 400% rise in the mark price would produce a loss far exceeding the principal. The attacker then bought JELLY heavily on external spot exchanges, using its extremely thin external liquidity depth to drive the price up several-fold (reported increases range from about 400% to several tens of times depending on the source; the settlement price of $0.0095 was the attacker's opening price). The sharp move in the external price was transmitted through the oracle to Hyperliquid's mark price. As described in Section 9.4.2, the mark price takes the weighted median of external CEX spot prices as its core input and adds components such as an order-book EMA on top; when the attacker's manipulated external price carried enough weight, the oracle price and mark price shifted substantially as a result. The shift in the mark price caused the short account's maintenance margin ratio to fall below its threshold, triggering forced liquidation.

Hyperliquid's liquidation waterfall follows a specific priority sequence: the liquidation engine first attempts to liquidate the position on the order book via limit orders; if order-book depth is insufficient or slippage exceeds a preset threshold, the position is transferred to the insurance fund; and if the insurance fund's resources are insufficient to

cover it, the position is ultimately transferred to the HLP Vault as the backstop. In the JELLY incident, because the short position far exceeded the token's order-book depth on Hyperliquid (and, as analyzed in Section 9.1.7, market makers could withdraw their quotes before liquidation orders executed by exercising cancellation priority, further weakening available depth), the position rapidly broke through the first two lines of defense in the liquidation waterfall and was transferred to the HLP Vault. As the price of JELLY kept climbing, the unrealized loss on the short position absorbed by the HLP Vault reached about $13.5 million at one point (roughly 6% of the Vault's total assets); this figure exceeds the attacker's initial principal because, under high leverage, a short position's loss is theoretically unbounded. The real threat lay not in the 6% snapshot loss at that moment but in the trend: a short position's loss is theoretically unbounded and expands linearly as the mark price continues to rise. Absent timely intervention, the loss would rapidly erode the HLP Vault's total assets of about $230 million, constituting a systemic threat [17].

Hyperliquid's validator set reached consensus within about 2 minutes, voting to delist the JELLY perpetual futures and to forcibly settle all JELLY positions at the attacker's original opening price of $0.0095, rather than at the prevailing market price of about $0.50 [17]. Within 24 hours of the incident, about $200 million in funds flowed out of Hyperliquid.

9.4.4 Risk transmission and counterfactual analysis

The risk-transmission path of the JELLY incident follows a clear causal chain: external price manipulation → mark-price shift → insufficient margin → forced liquidation → absorption by the HLP Vault → widening unrealized loss → a shock to market confidence → outflow of funds → declining order-book depth → potential cascade liquidations. The roughly $200 million outflow within 24 hours accounted for a significant share of Hyperliquid's total value locked (TVL) at the time, and had this outflow continued to accelerate, it could have triggered the bank-run-style negative feedback loop described by Diamond and Dybvig (1983) [18], in which depositors' rational individual behavior (withdrawing funds to protect themselves) produces a self-fulfilling liquidity crisis at the collective level.

Counterfactual analysis helps assess the reasonableness of the intervention decision. Had the validators not intervened and had the JELLY price continued to rise, the HLP Vault's short-side loss would have grown linearly with the mark price (a short position's loss is theoretically unbounded); given the constraint of the Vault's total assets of about $230 million, a further several-fold rise in the mark price could have rendered the Vault insolvent. The Vault's insolvency would mean not only direct losses for depositors but also, through the confidence channel, a large-scale withdrawal of funds that would accelerate the reversal of the flywheel. Viewed through the lens of cross-market contagion, JELLY's vulnerability characteristics — extremely thin external liquidity depth, no position limits, and high leverage — are not an isolated phenomenon. Several long-tail token perpetual futures with a similar risk profile exist on Hyperliquid, and the correlation risk of a simultaneous attack on multiple instruments constitutes a systemic exposure that has not yet been adequately assessed.

9.4.5 Legal and governance implications of the emergency intervention

The validators' decision to force settlement at the opening price of $0.0095 rather than the market price of about $0.50 has economic implications worth analyzing in depth. This choice of settlement price protected the HLP Vault from further losses but simultaneously harmed the interests of legitimate JELLY long holders: participants who held JELLY long positions during the attack for ordinary trading reasons were forced to close out at a price far below the market price, and their profit opportunity was unilaterally taken away.

From the standpoint of legal characterization, this intervention could face varying degrees of scrutiny under different regulatory frameworks. Under the anti-manipulation provisions of the U.S. Commodity Futures Trading Commission (CFTC), settling at a non-market price could be seen as an exchange's artificial intervention in market prices. Under the European Union's MiCA framework, emergency measures must satisfy a proportionality requirement — the scope of intervention should be commensurate with the risk being guarded against. A comparable precedent in traditional finance is exchanges' power to make ex post adjustments to erroneous trades and abnormal prices (for example, Rule 588 on *trade cancellation and price adjustment* in the CME rulebook, which handles erroneous trades outside a no-review range, along with the related authority exchanges hold under market emergency conditions); such mechanisms are typically accompanied by pre-disclosed procedural rules and an ex post review process. Hyperliquid's validator intervention was completed in about 2 minutes, and this timeline reveals the double-edged nature of a small validator set in crisis response: the capacity for rapid coordination allows it to act decisively before losses widen, but the same feature also means that a handful of participants can make major decisions affecting all users within an extremely short time and without any preset governance-deliberation procedure. This tension points directly back to the qualified statement about decentralization in Section 9.1.2: validator consensus is technically *majority rule*, but in a small validator set the practical meaning of *majority* differs from the public's perception of *decentralized governance*.

9.4.6 The boundary of protection

The JELLY incident reveals the boundary of protection of Hyperliquid's microstructure system. Placing these mechanisms within a complete risk-management framework helps pinpoint their scope: pre-trade controls include margin requirements, position limits, and linked-account detection as admission controls; at-trade controls include transaction ordering, matching rules, and price protection as execution-layer mechanisms; and post-trade controls include liquidation execution, absorption of risky positions, and insurance-fund management as settlement-layer mechanisms. The protection afforded by semantic ordering, cancellation priority, and discrete-time matching is concentrated in the at-trade stage: semantic ordering eliminates bribe-based ordering manipulation, cancellation priority eliminates the within-block adverse selection risk that market makers face, and discrete time eliminates the nanosecond-level speed-race advantage. The combined effect of these three layers of protection is to compress the space for MEV extraction to the single residual dimension of ordering within the same priority cate-

gory.

The JELLY attack path, however, bypassed precisely the at-trade layer: the attacker exploited the absence of linked-account detection and inadequate position limits at the pre-trade layer, together with the liquidation-absorption rules and cross-market price manipulation at the post-trade layer. Even if semantic ordering worked perfectly, and even if every Cancel instruction executed successfully within its intended block, the JELLY attack could still be carried out along exactly the same path. This shows that microstructural protection is a necessary but not sufficient condition: a complete market-safety system also requires independent defense mechanisms at the pre-trade layer (dynamic margin, position limits, cross-market monitoring) and the post-trade layer (liquidation rules, Vault governance, insurance-fund adequacy).

9.4.7 Validator economics and centralization risk

The effectiveness of the semantic ordering rules ultimately depends on those who enforce them: the set of validator nodes. Hyperliquid uses the HyperBFT consensus algorithm (a derivative of the HotStuff family), and its current validator set is relatively small (about 16–24 active validators, roughly 21 as of early 2026) [3]. This small size offers engineering advantages in delivering high throughput and low latency, but it also raises questions about the credibility of rule enforcement and about centralization risk.

On the MonarchEV dimension, validators (block proposers in particular) have ordering discretion within a single priority category. The preceding analysis characterized MonarchEV as a *theoretical residual risk*, but the accuracy of this judgment depends on the size and incentive structure of the validator set. In a validator set composed of a few nodes, the identity of the block proposer is highly predictable owing to deterministic leader rotation, and searchers can submit priority-ordering requests directly to the validator about to produce a block. Whether validators accept such requests directly determines the extent to which MonarchEV turns from a *theoretical residual* into an *actual revenue source*.

On the operational-risk dimension, a set of about 21 validators faces multiple infrastructure risks. Key management is the foremost concern: the Hyper Foundation controls a majority of the staked share, and a leak of its keys could in theory control the consensus process. Geographic and infrastructure concentration is the second layer of risk: if a majority of validators run on the same cloud provider, a single infrastructure failure could halt the chain or degrade it severely. Client-software homogeneity is the third layer of risk: all validators run the same client implementation, so a single software bug could paralyze the entire network; the Ethereum community's experience in promoting client diversity shows that this risk cannot be ignored in high-value systems. With open interest of $6.5 billion, the consequences of any of these single points of failure could be catastrophic.

On the moral-hazard dimension, the precedent set in the JELLY incident, where the validator set overrode the settlement price, creates an implicit expectation: when a position is large enough to threaten the solvency of the HLP Vault, validators may step in to protect it. This *too big to fail* implicit guarantee may be exploited strategically by large traders, who may game the validators' intervention behavior and take on risk

exposure beyond the normal risk-control range, because they expect the system to backstop them in extreme situations. This moral hazard is structurally similar to the problem posed by systemically important institutions in traditional finance, but in a decentralized protocol there is no mature regulatory framework to constrain it.

9.4.8 The core trade-offs

Taken together, the synergy of this microstructure system can be summarized by the positive feedback loop shown in Figure 9-9.

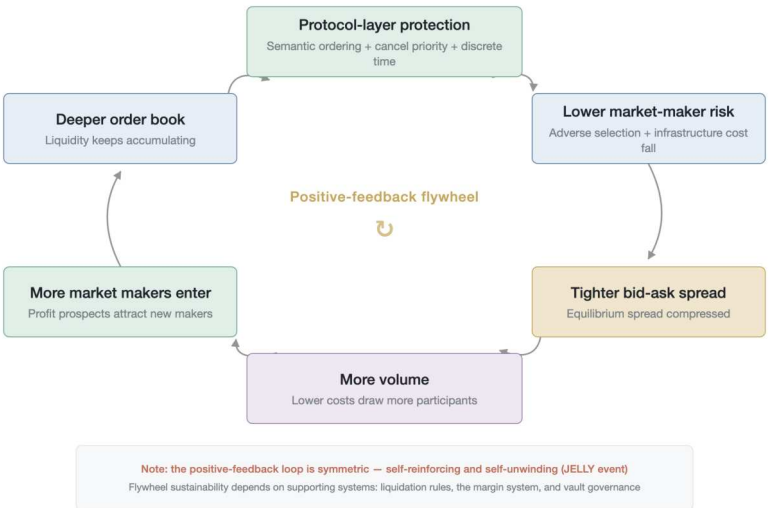


Figure 9-9. The synergy and positive feedback loop of the microstructure system (conceptual illustration, not empirical data; a six-node self-reinforcing flywheel; the footnote *symmetry of self-reinforcement and self-unraveling* is validated by the JELLY incident of March 26, 2025, per [16][17]; its sustainability depends on supporting systems such as liquidation rules, the margin system, and Vault governance)

The positive feedback loop shown in Figure 9-9 follows the transmission chain of *protocol-level protection* → *lower market-maker risk* → *narrower spreads* → *greater trading volume* → *the entry of more market makers*, and its self-reinforcing nature means that once the initial conditions are met (that is, once the mechanism's protection is sufficiently effective), the accumulation of liquidity accelerates. This positive feedback loop is not without cost, however, and Hyperliquid's microstructure system makes a series of explicit, systematic trade-offs among the goals it pursues. These trade-offs are not judgments of *right* and *wrong* but value choices among different dimensions of market quality, and together they define the platform's market characteristics. These trade-offs span multiple dimensions — liquidity versus price discovery, market-maker protection versus taker rights, transparency versus strategy privacy, security versus performance, and rule certainty versus flexibility — and on each dimension Hyperliquid makes an explicitly weighted choice. Table 9-9 systematically summarizes the choice, benefit, and cost on each dimension.

Dimension	Hyperliquid's choice	Gains	Sacrifices
Liquidity vs. price discovery	Leans toward liquidity	Tighter spreads, deeper order books, lower transaction costs	Slower price discovery, structural dependence on external markets
Market-maker protection vs. taker rights	Leans toward market makers	Lower barrier to market making, increased liquidity supply	Takers face a structural disadvantage, arbitrage space compressed
Transparency vs. strategy privacy	Leans toward transparency	Auditability, trust, decentralized verification	Market-maker strategies exposed, intensified analytical competition
Security vs. performance	Balances in the middle	A 200–400 ms block time balances security and user experience	Not as fast as a CEX, not as secure as Ethereum
Rule certainty vs. flexibility	Leans toward certainty	Rules hardcoded at the consensus layer, not unilaterally modifiable	Slower to adapt to novel attacks, as in the JELLY incident

Table 9-9. The five core trade-offs of Hyperliquid's mechanism (Data source: compiled by the author)

Hyperliquid's design philosophy explicitly chooses a value orientation of *liquidity first, market-maker friendly, and transparently verifiable*, accepting in return such costs as *externally dependent price discovery, disadvantaged takers, and a lack of strategy privacy.* Understanding these inherent, structural trade-offs is key to assessing whether the Hyperliquid model can be adopted more widely and how stable it will be over the long term.

9.5 Microstructure effects and the liquidity flywheel

The preceding analysis focused on the internal logic and trade-offs of the microstructure system; this section shifts the perspective from mechanism design to the systemic consequences at the market level. Semantic ordering, cancellation priority, and discrete-time matching do not act in isolation on individual traders; by reshaping the cost structure of market making, opening up channels for liquidity supply, and forming a self-reinforcing feedback loop, they exert a profound influence on Hyperliquid's overall market microstructure. The combined effect of these influences forms the basis of the core competitiveness that distinguishes Hyperliquid from both the CEX and the general-purpose DEX.

9.5.1 The restructuring of market-making costs and spread compression

Using the Glosten-Milgrom spread-decomposition framework established in Section 9.1.3 [5], the effect of the cancellation priority mechanism on the cost structure of market making can be analyzed quantitatively through its two dominant components: adverse selection costs and infrastructure costs. In the CEX environment, market makers must continually contend with the risk of informed traders arbitraging against stale quotes while also investing heavily in low-latency infrastructure to stay competitive in

the speed race; Hyperliquid's cancellation priority mechanism produces a direct and quantifiable compression of both components.

On the adverse selection dimension, because Cancel instructions enjoy absolute priority within a block, a market maker can, with high probability, successfully withdraw a stale quote within the one-block-time window (200–400 ms) after an external price move. This reduces the probability of an adverse selection event from a competitive outcome that, in the CEX environment, depends mainly on physical latency, to a risk confined to a very small window near the block boundary. On the infrastructure dimension, discrete-time matching raises the effective competitive time granularity from the nanosecond to the hundred-millisecond scale, so that costly infrastructure investments such as dedicated microwave towers and colocated servers are no longer a necessary condition for the market-making business. The simultaneous decline in these two costs enables market makers to quote narrower spreads while remaining profitable. Empirical data show that, during calm market periods, the bid-ask spreads on Hyperliquid's major trading pairs are typically comparable to, or even better than, those of top-tier CEXs such as Binance (the measurement basis and the static-versus-effective-depth qualifications are detailed in Section 9.5.4), while the teams engaged in market making invest far less capital in infrastructure than comparable teams in the CEX environment. This restructuring of the cost structure is precisely the cost-side root of the shift in the focus of market-making competition described in Section 9.2.3 — *from a capital-intensive speed race to knowledge-intensive competition on model quality.*

9.5.2 The HLP Vault and the democratization of liquidity supply

Hyperliquid's microstructural protection not only lowers the operating costs of professional market makers but also provides the basic conditions for an entirely new mode of liquidity supply. The HLP Vault (Hyperliquidity Provider) is a liquidity-provider vault built into the Hyperliquid protocol that allows ordinary users to deposit USDC into the vault, where automated strategies carry out market-making operations on their behalf, and users share in the market-making returns in proportion to their holdings. The core premise of this design is that cancellation priority and discrete-time matching provide the vault's market-making strategy with the same protocol-level protection as a professional market maker: the vault's Cancel instructions likewise enjoy the highest execution priority within a block, and the probability that its stale quotes are filled by arbitrageurs is likewise compressed to a structural-residual level. Without these microstructural protections, a large, publicly visible liquidity pool with a predictable strategy would become an ideal attack target for informed traders, and market-making losses would far exceed returns.

The economic significance of the HLP Vault is that it transforms market making from an exclusive activity requiring professional infrastructure and quantitative-modeling capability into a source of return in which ordinary users can participate indirectly through capital contributions. In traditional financial markets and on CEXs, market-making returns are captured almost entirely by professional institutions with speed and model advantages. Hyperliquid's mechanism protection lowers the technical barrier to market making, and the HLP Vault further lowers the capital barrier, allowing dis-

persed small amounts of capital to pool into a liquidity pool with an impact on market depth. The narrative of democratized market making, however, also requires scrutiny of participants' behavioral characteristics. Barber and Odean (2000) show that retail investors systematically exhibit return-chasing bias, overconfidence, and recency bias [19]. In the context of the HLP Vault, its substantial historical returns (promoted during some periods as a high annualized figure, though independent estimates put the 2025 annualized return at about 11%–12% and the full-period compound annual growth rate (CAGR) at about 42%) may have attracted many participants who do not understand the underlying leveraged market-making risk and who may view it as a yield-farming product rather than an investment that bears the tail risk of market making. The pattern after the JELLY incident, in which assets under management (AUM) drew down by about 15% and then gradually recovered, is highly consistent with a typical *panic-and-forget* cycle, suggesting that participants as a group may systematically underestimate risk. Moreover, the structure of the HLP Vault (an investment of money, in a common enterprise managed by others, with an expectation of profit) formally satisfies the four elements of the Howey test [20]. If it were deemed a security in a major jurisdiction, its issuance and sale would require registration or an exemption, which would fundamentally constrain the scaling of democratized market making.

The JELLY incident also revealed an inherent fragility of this model: as a large, behaviorally predictable public liquidity pool, the HLP Vault's liquidation-absorption rules can be exploited in a targeted manner by attackers. There is an inherent tension between the vault's advantage of scale and its visibility as an attack target.

9.5.3 The liquidity flywheel and ecosystem sustainability

The two levels of influence described above — the compression of professional market-making costs and the opening of retail liquidity supply — together drive a self-reinforcing liquidity flywheel. The flywheel's operation relies on two coupled engines: professional market makers, drawn by lower adverse selection costs, provide tighter quotes, while HLP Vault users, drawn by a substantial share of market-making returns, deposit capital. Professional market makers benefit from the order-book depth thickened by the vault's capital (greater depth lowers the probability that a large order will move against their positions), while vault depositors benefit from the more efficient market conditions that professional market makers create (a more efficient market means more stable market-making returns). The positive feedback between the two accelerates the accumulation of liquidity: deeper liquidity brings narrower spreads, narrower spreads attract more trading volume, more trading volume generates more market-making profit, and that profit in turn attracts further entry by professional market makers and vault depositors.

A positive feedback loop, however, inherently possesses a symmetry between self-reinforcement and self-unraveling. The noise-trader model of DeLong, Shleifer, Summers, and Waldmann (1990) and Soros's theory of reflexivity both hold that the very mechanism by which a positive feedback loop reinforces itself in the upswing also reinforces itself in the downswing [21]. The JELLY incident provides an empirical test of this symmetry: in the positive phase, the HLP Vault's high returns attracted

inflows, the inflows thickened order-book depth, and the greater depth further raised market-making returns, forming an upward spiral; in the negative phase (after JELLY), losses triggered outflows, an AUM drawdown of about 15% weakened order-book depth, and the decline in depth could further worsen market-making returns, forming a downward spiral. The recovery of AUM after JELLY suggests that the flywheel currently runs above critical mass, but the size of the safety margin is difficult to determine a priori: whether there exists a threshold of TVL or trading volume below which the flywheel would collapse irreversibly is an important question that remains to be tested.

The flywheel's scaling itself also faces endogenous constraints. The HLP Vault's current AUM of about $270 million (with a historical peak of about $500 million in mid-2025) operates within an average daily trading volume of $6.9 billion, and changes in its positions already carry significant market influence. As AUM grows, the vault faces a threefold capacity constraint: larger position adjustments produce greater price impact against limited order-book depth; a larger-scale quoting pattern is more easily identified and countered by competitors in a fully transparent on-chain environment; and a larger-scale withdrawal during emergency de-risking would intensify the market shock. These constraints imply a theoretical upper bound on AUM, beyond which market-making returns would be eroded by the vault's own scale effects.

The sustainability of this flywheel depends on the balance between protection and openness maintained by the microstructure system. If market-maker protection is too strong, tilting market conditions excessively toward market makers at the expense of takers' experience, trading volume may shrink as execution costs rise too high; if protection is too weak, market makers' profits are eroded, and both professional institutions and vault depositors may withdraw. The JELLY incident also revealed a third failure mode: the flywheel can be interrupted by an attack on the vault's infrastructure itself, even when the transaction-ordering and matching mechanisms function normally. The long-term stability of the flywheel therefore depends not only on the design quality of the microstructure mechanisms but also on the robustness of supporting systems such as liquidation rules, the margin system, and vault governance. The directions in which the Hyperliquid model evolves — the ongoing optimization of block time, the extension of semantic ordering rules, and the governance of validators' informational advantage — will all affect the flywheel's operating efficiency by altering this balance.

9.5.4 Empirical data and market performance

The theoretical predictions established in the preceding three subsections — that the compression of market-making costs narrows spreads, that the vault mechanism attracts retail capital, and that the flywheel effect drives the accumulation of liquidity — can be preliminarily validated against publicly available market data. As of March 2026, Hyperliquid's L1 TVL was about $4.4 billion and its average daily trading volume about $6.9 billion — roughly 15% of Binance Futures' daily volume — giving it about a 70% share of the on-chain perpetual futures market [22]. In terms of open interest, Hyperliquid's open interest was about $6.5 billion, roughly 30% of Binance Futures' (about $21.2 billion), while reaching more than 70 times that of dYdX (about $82

million) and more than 50 times that of GMX (about $124 million), far exceeding other on-chain competitors [23].

Before analyzing the specific data, the measurement methodology should be explained. The *spread* cited in this section is defined as the time-weighted average of the best bid and offer (BBO) spread — that is, the mean, weighted by time, of the difference between the best bid and the best offer at each moment within the observation window. *Depth* is defined as the total volume of resting orders within a certain percentage band of the mid price (typically 0.1–0.5%). The data source is public market data from CoinGecko and DefiLlama [22][23], with a sampling window of February–March 2026 and a sampling frequency of hourly snapshots. The choice of these measurement parameters directly affects the comparability and interpretability of the data.

In the BTC perpetual futures market, some snapshot data show that Hyperliquid's top-of-book bid-ask spread can be narrower than Binance's. The commonly cited specific figures (for example, a spread of about $1 versus $5.5, or mid-price depth of about 140 BTC versus 80 BTC) are difficult to trace to an independent primary source, and independent slippage estimates show that at larger trade sizes (for example, around $1 million) Binance typically offers deeper executable depth. The judgment that *Hyperliquid has better depth* should therefore be confined to the top of the book and read cautiously in light of the effective-depth discount discussed below [23]. This comparison must be qualified along two dimensions.

The first dimension is the distinction between static depth and effective depth. The 140 BTC shown in an order book snapshot is *static depth* — the total volume displayed on the order book at that moment. However, under the cancellation priority mechanism, market makers can, with high probability, withdraw their quotes within the same block before a large market order executes. Therefore, when a 100 BTC market buy order actually executes, the *effective depth* available for filling may be significantly lower than the 140 BTC statically displayed. The upper bound of effective depth equals static depth (no market maker cancels), while the lower bound equals the residual depth after all market makers simultaneously exercise cancellation priority — and this residual depth may include only passive liquidity, such as the HLP Vault, that cannot respond with a cancellation within the same block. This distinction cross-corroborates the *depth illusion* phenomenon analyzed in Section 9.2.3 and the analysis of apparent order-book liquidity in Section 8.2.

The second dimension is the HLP Vault's subsidizing effect on the spread. In his classic model of informed trading, Kyle (1985) distinguished apparent liquidity from equilibrium liquidity [24]. As a liquidity pool that aims at sharing market-making returns (rather than pure profit maximization), the HLP Vault may systematically make markets at spreads below the equilibrium level, and its narrow spreads may partly reflect an implicit subsidy rather than a pure efficiency gain. Comparing spreads across periods with high and low vault contributions would help separate the subsidy effect from the mechanism-efficiency effect, but public data for such period-by-period comparisons are currently lacking.

The growth metrics above must also be interpreted within the macro environment. The years 2024–2025 fell within the bull-market cycle following the approval of the BTC

exchange-traded fund (ETF), during which the TVL and trading volume of almost all crypto infrastructure grew substantially. To separate Hyperliquid's platform-specific excess growth from systemic bull-market beta, its growth rate over the same period should be compared with on-chain competitors such as dYdX and GMX. Preliminary data show that Hyperliquid's growth rate is significantly higher than that of comparable platforms, but the contribution of HYPE token incentives within it remains difficult to fully isolate.

The HLP Vault's operating data provide a basis for testing the theory of democratized liquidity. The vault's AUM peaked at about $510 million (mid-2025) and currently stands at about $270 million (DefiLlama, mid-2026), with cumulative net profit of about $121.8 million (roughly 50% of the total return over the period, using the vault's average AUM over the period as the denominator; as of October 2025, on the DefiLlama HLP dashboard basis, consistent with Chapter 21); according to independent estimates from OAK Research, OnchainTimes, and others, its annualized return was about 11%–12% (2025), its full-period CAGR about 42%, and its trailing-12-month return about 22%, with returns that are highly event-driven and volatile rather than a stable annualized level of 65%–70% [22]. This annualized return, however, lacks a decomposition, and its composition may include several sources: market-making spread income, HYPE token incentives, directional beta exposure (in a bull market, market makers typically hold a net long position), and profits from absorbing liquidations. Citing the 65%–70% annualized figure without decomposing the sources of return can easily give readers the mistaken impression of *65%–70% from pure market making*, when in reality a considerable proportion of the return may come from token incentives and directional exposure — sources that are not sustainable across cycles. After the JELLY incident, the vault's AUM drew down by about 15% around the event (the vault stood at about $230 million when JELLY occurred, consistent with the figure in Section 9.4.3, and drew down to about $195 million; the roughly $500 million was the historical peak in mid-2025, not the level at the time of JELLY), and then gradually recovered after the protocol repaired its liquidation-absorption rules. This data point validates both the flywheel's positive transmission and its fragility under extreme events. At the validator level, Hyperliquid currently runs about 16–24 active validators (roughly 21 as of early 2026), among which the nodes controlled by the Hyper Foundation hold a majority of the staked share [3]. Single-block finality (the time required for a transaction to become irreversible) is about 70 ms, distinct from the block time of about 200 ms; the median end-to-end order-confirmation latency is about 200 ms, with a 99th-percentile latency of about 900 ms (for a colocated client). Overall, these data support this chapter's theoretical analysis, but their limitations should be noted: the market data are snapshots at specific points in time, they cannot fully separate the contribution of mechanism effects from that of token-incentive effects to liquidity growth, and, as a rapidly evolving system, Hyperliquid's market metrics may change significantly over the short term.

9.6 Chapter summary

Taking Hyperliquid as a case study, this chapter applied the order-semantics analytical framework established in Chapter 7 and the order-book microstructure theory revealed in Chapter 8 to a specific application-chain exchange, systematically analyzing how the two mechanism innovations of semantic ordering and cancellation priority form a self-consistent microstructure system on top of the discrete-time property inherent to on-chain matching. Semantic ordering classifies the economic meaning of transactions into four tiers (System, High, Standard, and Low) at the consensus-protocol level, placing the priority of liquidity-protecting actions above that of liquidity-consuming actions, while assigning liquidation orders and system-maintenance transactions to separate priority tiers. Section 9.1.7 further analyzed the interactions of semantic ordering with the margin system, funding rate settlement, and the liquidation engine, revealing how the ordering mechanism is embedded in the broader contract-settlement system. Cancellation priority concretizes the classification rules into a protocol-level right of last look for market makers; the discrete-time property inherent to on-chain matching compresses the precision of the speed race from the nanosecond to the hundred-millisecond scale, creating the time window necessary for implementing semantic ordering and cancellation priority — a window that Hyperliquid did not invent but that is an inherent structural feature of any on-chain matching system; Hyperliquid's innovation lies in recognizing and exploiting this feature to build differentiated ordering rules. Placing Hyperliquid's semantic ordering within the broader spectrum of MEV mitigation approaches — comparing it with alternatives such as Flashbots' private mempool, the threshold-encrypted mempool, intent solving, and dYdX v4's distributed ordering rights — shows that, among the approaches surveyed in this chapter, semantic ordering is a rare one that makes an explicit value judgment at the protocol level, at the cost of a strong dependence on an application-specific chain.

The core effect of this microstructure system is to systematically reduce the adverse selection risk and infrastructure investment costs faced by market makers. Empirical data provide preliminary validation: some snapshots show that the top-of-book bid-ask spread on Hyperliquid's BTC perpetual futures can be narrower than Binance's (the commonly cited figures of about $1 versus about $5.5 and static depth of about 140 BTC versus about 80 BTC are difficult to trace to an independent primary source, and Binance's effective depth is typically better at large trade sizes, so these should be read cautiously and confined to the top of the book), while the roughly 70% share of the on-chain perpetual futures market is a more firmly established figure. The interpretive qualifications — the gap between static and effective depth, the HLP Vault's subsidizing effect on the spread, and the macro bull-market cycle — are all detailed in Section 9.5.4. With a peak AUM of about $500 million, the HLP Vault has validated the feasibility of democratized liquidity supply, but its commonly cited 65%–70% annualized return is in fact a figure that includes token incentives and directional exposure; independent estimates put it at about 11%–12%, with returns that are highly event-driven and volatile, and whose sustainability remains to be verified.

At the level of liquidity, cancellation priority exhibits a *liquidity paradox*: in normal times it lowers market-making costs and improves liquidity supply, but the same mech-

anism may accelerate liquidity evaporation during periods of market stress, and the procyclical effect of market makers exercising their cancellation right with zero friction merits the attention of market designers. The in-depth analysis of the JELLY incident revealed, along the three dimensions of attack mechanism, risk transmission, and legal governance, systemic fragilities lying beyond microstructural protection: the breach of the liquidation waterfall, the HLP Vault's forced absorption, and the validators' forced settlement at a non-market price — all of which occurred beyond the boundary of the ordering layer's protection. The limited size and rapid governance capacity of the validator set appear as advantages in crisis response but also give rise to tensions among concentrated operational risk, moral hazard, and the narrative that *rules are hardcoded at the consensus layer*. The positive feedback loop of the liquidity flywheel possesses a symmetry between self-reinforcement and self-unraveling, and its sustainability depends not only on the design quality of the microstructure mechanisms but also on the robustness of supporting systems such as liquidation rules, the margin system, and vault governance. As an adjustable hyperparameter, block time directly defines the trade-off surface between market-maker protection and price discovery efficiency, but the quantitative calibration of the U-shaped curve requires estimation of parameters specific to the crypto perpetual futures market.

As Budish et al. (2015) [14] noted, the core of market design is not to have participants reach an equilibrium under given rules but to design rules that produce a better equilibrium. Hyperliquid's experiment shows that blockchains offer the possibility of encoding market rules in a consensus protocol in a way that is transparent, verifiable, and difficult to tamper with unilaterally; and its explorations in microstructural protection, the democratization of liquidity supply, and the construction of an ecosystem flywheel, together with the boundary conditions and failure modes those explorations exposed, will provide the necessary theoretical foundation for later chapters' analysis of the concrete operating mechanisms of the HLP Vault.

References

[1] Aquilina, M., Budish, E., & O'Neill, P. (2022). Quantifying the high-frequency trading "arms race". *The Quarterly Journal of Economics*, *137*(1), 493–564. https://doi.org/10.1093/qje/qjab032

[2] Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., & Juels, A. (2020). Flash Boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. *2020 IEEE Symposium on Security and Privacy (SP)*, 910–927. https://doi.org/10.1109/SP40000.2020.00040

[3] Hyperliquid. (n.d.). *Hyperliquid docs*. Retrieved February 28, 2026, from https://hyperliquid.gitbook.io/hyperliquid-docs

[4] Hyperliquid. (2024, January 6). *Latency and transaction ordering on Hyperliquid*. Medium. https://hyperliquid.medium.com/latency-and-transaction-ordering-on-hyperliquid-cf28df3648eb

[5] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*,

14(1), 71–100. https://doi.org/10.1016/0304-405X(85)90044-3

[6] Harris, L. (2003). *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press.

[7] Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, *70*(3), 393–408.

[8] Flashbots. (n.d.). *Flashbots Protect documentation*. Retrieved March 15, 2026, from https://docs.flashbots.net/flashbots-protect/overview

[9] CoW Protocol. (n.d.). *CoW Protocol documentation*. Retrieved March 15, 2026, from https://docs.cow.fi

[10] Lewis, M. (2014). *Flash Boys: A Wall Street revolt*. W. W. Norton & Company.

[11] Foucault, T., Kadan, O., & Kandel, E. (2005). Limit order book as a market for liquidity. *The Review of Financial Studies*, *18*(4), 1171–1217. https://doi.org/10.1093/rfs/hhi029

[12] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

[13] Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, *3*(2), 5–39. https://doi.org/10.21314/JOR.2001.041

[14] Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, *130*(4), 1547–1621. https://doi.org/10.1093/qje/qjv027

[15] Kirilenko, A. A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The Flash Crash: High-frequency trading in an electronic market. *The Journal of Finance*, *72*(3), 967–998. https://doi.org/10.1111/jofi.12498

[16] Arkham Intelligence. (2025). *JELLYJELLY exploit on Hyperliquid: On-chain attribution*. Arkham Research. https://intel.arkm.com

[17] Halborn. (2025). *Explained: The Hyperliquid JELLY incident (March 2025)*. Halborn Blog. https://www.halborn.com/blog

[18] Diamond, D. W., & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, *91*(3), 401–419. https://doi.org/10.1086/261155

[19] Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The Journal of Finance*, *55*(2), 773–806. https://doi.org/10.1111/0022-1082.00226

[20] SEC v. W. J. Howey Co., 328 U.S. 293 (1946).

[21] DeLong, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, *98*(4), 703–738. https://doi.org/10.1086/261703

[22] DefiLlama. (n.d.). *Hyperliquid L1 statistics*. Retrieved March 15, 2026, from https://defillama.com/chain/hyperliquid-l1

[23] CoinGecko. (n.d.). *Hyperliquid futures statistics*. Retrieved March 15, 2026, from

https://www.coingecko.com/en/exchanges/hyperliquid

[24] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

[25] Sun, X. (2022). *Short note on {Monarch, Moloch, Mafia} extractable value*. HackMD. https://hackmd.io/@sxysun/short-note-ext

Part IV
Pricing Mechanisms

Core question: How is the price of a perpetual future anchored to reality? How do funding rates, liquidation, and margin jointly maintain market order, and how do they fail under extreme conditions?

A Game-Theoretic Analysis of Funding Rates

On March 12, 2020, the cryptocurrency market suffered one of the most violent single-day crashes in its history, a day that came to be known as Black Thursday [1]. Within just 24 hours, the price of Bitcoin fell by nearly 50%, plunging from about $7,900 to a low of $3,867 [2]. This macro shock, triggered by global panic over the COVID-19 pandemic, set off an unprecedented liquidation cascade of leveraged positions in the crypto derivatives market; on BitMEX alone, liquidations exceeded $700 million [3]. Yet amid this collapse, in which evaporating liquidity intertwined with panic selling, the funding rate mechanism of the derivatives market malfunctioned severely: the price of BitMEX's XBTUSD perpetual futures traded at one point at a discount of more than 10% to spot, and the funding rate touched an extreme level of -0.375% in a single period [4].

At that particular moment, the price of the perpetual futures had fallen far below the spot price—BitMEX's XBTUSD perpetual traded at one point more than $500 below spot—exhibiting an extreme negative basis [5]. According to the standard theoretical description of the funding rate mechanism, such a negative basis should trigger a negative funding rate, under which shorts pay longs, thereby incentivizing traders to buy the perpetual and pushing its price back toward spot value. Yet even after the rate had turned extremely negative, the price failed to converge. Arbitrageurs, unable to deposit funds into BitMEX because of congestion on the blockchain network, could not execute the arbitrage [5]; the forced sell orders generated by the liquidation cascade continued to depress the perpetual price, forming a positive feedback loop of "extreme rate $\rightarrow$ still no correction $\rightarrow$ widening basis." Under conditions of evaporating liquidity, the funding rate's anchoring mechanism nearly failed entirely.

This extreme anomaly raises three fundamental questions about the underlying mechanism of perpetual futures. Where does the anchoring force of the funding rate actually come from? Is it a passive mathematical formula for correcting the price spread, or the equilibrium outcome of a complex, friction-laden game among multiple participants? Building on this, why does a mechanism that has functioned well more than 90% of the time break down so severely at exactly the moment when the market most needs it to anchor prices? And taking the analysis together, does the extreme behavior of the funding rate itself encode a deeper signal about the market's leverage state and systemic fragility?

To answer these questions, this chapter moves beyond the surface-level description of a "negative feedback loop of algorithmic anchoring" and builds a systematic game-theoretic analytical framework. It uses a three-layer decomposition model to deconstruct the economic components of the funding rate, employs an anchoring strength spectrum to measure the effectiveness boundary of the anchoring mechanism across different market states, and constructs a three-layer game model to analyze the interaction structure among longs and shorts, arbitrageurs, and exchanges. It then examines how rate polarization positive feedback explains the mechanism that generates extremes, and finally advances the funding rate as state variable hypothesis, which repositions the funding rate from a passive market "output" to a "predictor" of future volatility and systemic risk. Across these five nested frameworks, the chapter builds a chain of reasoning from foundational concepts to frontier analysis, arguing that the funding rate is a core variable for understanding the microstructure and game dynamics of the crypto derivatives market.

10.1 Mechanisms and definitions

Before turning to the game-theoretic analysis, we need a clear and rigorous shared vocabulary. The three-price system of perpetual futures, the internal structure of the funding rate formula, and the parameter differences across exchanges together set the initial conditions and payoff structure for every game that follows. They determine the profit margin available to arbitrageurs, the holding cost borne by speculators, and the resilience of the entire system under extreme stress. Before analyzing participants' strategies, we must define the rules of the game precisely.

10.1.1 The three-price system

The ability of perpetual futures to operate independently, free from the constraint of a delivery date, rests on a three-price system. Each of the three serves a distinct function, and their interaction keeps the market running. Understanding the distinctions and connections among them is a prerequisite for grasping the funding rate mechanism.

The index price is the objective value benchmark for the entire system. It is typically computed as a volume-weighted average of the latest traded prices across several major global spot exchanges, such as Coinbase, Kraken, and Bitstamp [6]. The index price is designed to provide a global-consensus value that is difficult for any single entity or large trade to manipulate. To ensure its robustness, exchanges usually build multiple anomaly-detection mechanisms into the calculation: automatically discarding price sources that deviate excessively, excluding data from periods of liquidity drought, and dynamically adjusting weights. For example, when one spot exchange's quote deviates from the median of the others by more than a certain threshold, that exchange's weight is automatically reduced or even set to zero. This design makes the index price a relatively "exogenous" variable: it reflects the aggregate state of the global spot market rather than the internal supply-and-demand dynamics of any single derivatives exchange. However, the exogeneity assumption for the index price holds only if its constituent exchanges have sufficient liquidity depth. For mainstream assets such as

BTC and ETH, the weighted average across many highly liquid exchanges makes manipulation extremely costly, so the exogeneity assumption holds reasonably well. For long-tail assets with low circulating market capitalization, however (such as ALPACA, analyzed later in this chapter), the constituent exchanges may be extremely thin, and a single medium-sized market order can move the spot price significantly and thereby affect the index price. At that point, the "exogenous anchor" assumption of the three-price system partly breaks down: a manipulator can influence both the index price and the perpetual price at once, fundamentally weakening the anchoring foundation of the rate mechanism. The Mango Markets exploit of October 2022 is a canonical case: the attacker, Avraham Eisenberg, manipulated the price of the MNGO token in a low-liquidity spot market, successfully affecting the oracle price feed and ultimately causing protocol losses of over $110 million (per CFTC and DOJ figures). This risk dimension is developed further in the analysis of funding-rate-manipulation feasibility in Section 10.6.3.

The perpetual price—the latest traded price on the derivatives order book—is a direct indicator that reflects, in real time, the internal supply and demand, leverage direction, and trading sentiment of the derivatives market. It forms directly through the bidding of longs and shorts within the exchange and is therefore highly "endogenous." When large numbers of leveraged longs pour in, the perpetual price is driven above the index price, forming a positive premium; conversely, when panic selling or short-side force dominates, the perpetual price falls below the index price, forming a negative premium (i.e., a discount). Yet precisely because of this endogeneity, the perpetual price is highly vulnerable to the impact of a single large order or short-term manipulation, and one sufficiently large market order can create a significant price deviation during a thinly traded period.

The mark price, in turn, is a design product that strikes a balance between anchoring to true value and reflecting market reality. It is typically composed of a robust index price plus a moving-averaged basis (the difference between the perpetual price and the index price) [7]. The mark price calculation differs fundamentally across exchanges: BitMEX uses a composite calculation based on the order-book mid-price and the index price, with an exponential moving average (EMA) window length that varies by contract; Binance takes a weighted average of the index price and the best-bid/best-ask mid-price, with weights that adjust dynamically to a contract's liquidity conditions; and Hyperliquid uses the oracle median price plus a small-window EMA basis. These differences directly determine the moment at which liquidation is triggered: under identical market conditions, the same trader may be liquidated on one exchange but not on another. The mark price serves two core functions. First, it is the sole basis for computing unrealized profit and loss (P&L) and assessing margin levels. Second, during extreme volatility it serves as the warning line that triggers forced liquidation. The mark price is designed to filter out short-term noise in the perpetual price, providing a buffer layer for the market. If the perpetual price were used directly to trigger liquidations, a brief wick—a sharp price deviation over a very short interval that quickly reverts—could unnecessarily liquidate large numbers of innocent positions. Through its smoothing mechanism, the mark price minimizes this risk.

The interaction among these three prices constitutes the infrastructure on which the

funding rate operates: the basis, time-weighted and averaged, becomes the premium component that drives the rate calculation, while the mark price maintains the system's solvency floor, preventing over-leveraged positions from triggering systemic risk over transient price disturbances.

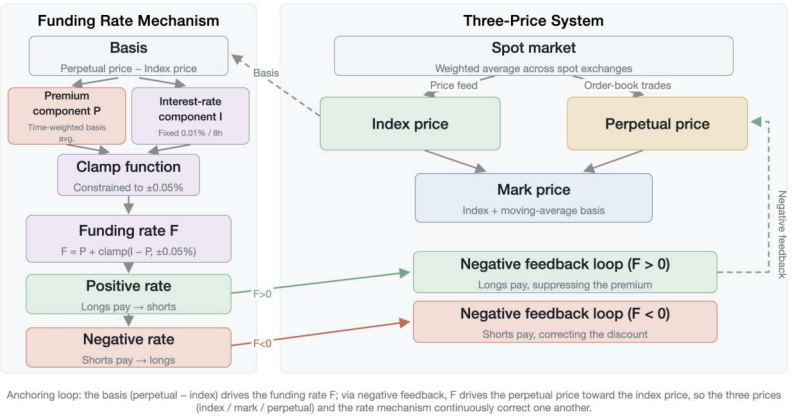


Figure 10-1. The three-price system of perpetual futures and the funding rate anchoring mechanism

10.1.2 Deconstructing the funding rate formula

The details of the funding rate calculation may vary slightly across exchanges, but the core logic is highly consistent. Taking the common formula used by major exchanges, the funding rate F can be deconstructed into the following form [8]:

$$F = P + \text{clamp}(I - P,\ -0.05\%,\ +0.05\%)$$

where F is the funding rate, P is the premium component, and I is the interest rate component.

This formula carries several layers of economic meaning: the premium component captures the market's directional preference, the interest rate component embeds the capital cost of holding a position, and the clamp function introduces nonlinear institutional friction through its boundary constraints. Their interaction gives the funding rate markedly different response characteristics across market states: it is locked to a fixed baseline under mild deviations, and only under extreme deviations does it exhibit linear sensitivity to the premium.

The premium component is the most information-dense dynamic part of the rate, directly quantifying the degree to which the perpetual price deviates from the index price. It is usually the time-weighted average of the basis over a window (such as the 8-hour settlement period) rather than a simple instantaneous snapshot. Specifically, BitMEX uses a time-weighted average price (TWAP) over the past minute, Binance uses an arithmetic mean sampled every minute within the settlement period, and Hyperliquid uses an average over a 1-hour window of samples taken every 5 seconds. These sampling differences directly affect how sensitive the rate is to short-term price shocks. This time-weighting has a direct effect: it prevents traders from manipulating the rate

by momentarily creating a large basis just before settlement, and it also makes the rate reflect a sustained directional preference over a period rather than transient noise. When P is positive, the perpetual has sustained a premium over the settlement period and long demand is strong; when P is negative, a sustained discount prevails and short-side force dominates. In short, the premium component prices the directional exposure that participants are willing to pay for, which makes it the most information-rich part of the rate.

The interest rate component reflects the capital cost of holding a synthetic leveraged position. When BitMEX pioneered perpetual futures in 2016, it set this component at a fixed 0.01% every 8 hours (approximately 10.95% simple annualized) [9]; the institutional rationale for this interest rate component—why it is not a precise macro risk-free rate, and how it was set as an institutional benchmark to incentivize arbitrageurs to provide liquidity—is discussed in the financing cost layer of Section 10.2.1. As perpetual futures were widely adopted by major exchanges worldwide, this 0.01% interest rate component became a de facto industry standard that profoundly shaped the structural characteristics of the entire market.

The clamp function is an often-underestimated part of the formula, and its influence extends far beyond what it appears to have on the surface. It is a boundary-constraint function that forces the value of $(I - P)$ into the interval $[-0.05\%,\ +0.05\%]$. Its mathematical definition is:

$$\text{clamp}(x,\ a,\ b) = \max(a,\ \min(x,\ b))$$

The clamp function does more than limit extreme fluctuations. It creates an asymmetric effect that forms the basis for understanding the structural characteristics of the funding rate. To see this intuitively, we can analyze the response curve of the funding rate F to the premium component P. When the interest rate component $I = 0.01\%$, the entire response relationship exhibits a three-segment piecewise-linear character (Table 10-1):

Range of premium component P	Value of clamp	Final rate F = P + clamp	Economic meaning
$P < -0.04\%$	$+0.05\%$ (upper bound)	$F = P + 0.05\%$, declining linearly with P	Deep-discount zone: the rate falls as the discount deepens, but is always "held up" by the $+0.05\%$ correction term
$-0.04\% \le P \le +0.06\%$	$I - P$ (no clamping)	$F \equiv 0.01\%$ (constant)	Stable band: the clamp offsets fluctuations in P, locking the rate
$P > +0.06\%$	-0.05% (lower bound)	$F = P - 0.05\%$, rising linearly with P	Deep-premium zone: the rate rises as the premium deepens, but is "slowed" by the -0.05% correction term

Table 10-1. The three-segment piecewise-linear response of the funding rate to the premium compo-

nent (interest rate component I = 0.01%) (Data source: compiled by the author)

Empirical data confirm the effect of this design [8]: BitMEX's Q3 2025 report shows that the BTC perpetual rate was exactly equal to 0.01% for 78.19% of the time and the ETH rate for 87.52% of the time. This does not mean the market was continuously at equilibrium—the premium component P in fact fluctuated continuously—but rather reflects a "stability illusion" created by the clamp function. This stable band thereby gives the funding rate a built-in positive bias: even when the market is slightly in discount, longs must still pay shorts, constituting a non-negligible "institutional friction layer." The full argument for this structural positive bias—including the width of the stable band, the critical example at $P = -0.03\%$, and data on the share of positive rates—is formalized in the "three-layer decomposition model" of Section 10.2, specifically in Section 10.2.3.

10.1.3 Cross-exchange parameter differences

The funding rate is not a single, unified standard but a design space with multiple parameters. Differences across exchanges in settlement frequency, the interest rate component, the clamp bounds, the composition of the index price, and rate caps may look like mere "technical implementation details" on the surface, but they significantly alter participants' cost-benefit structures and thereby shape markedly different game environments. Understanding these differences is key to understanding why the funding rate for the same asset can behave so differently across platforms.

Table 10-2 compares the core design parameters of three representative exchanges—BitMEX, Binance, and Hyperliquid—across five dimensions: settlement frequency, the interest rate component, the clamp range, the rate cap, and the source of the index price. These parameter differences directly determine the response speed of arbitrage capital on each platform, the feasible cost of a manipulation attack, and the magnitude of rate polarization during extreme conditions.

Design dimension	BitMEX	Binance	Hyperliquid
Settlement frequency	Every 8 hours (fixed)	Dynamic (8h / 4h / 1h) [10]	Every 1 hour (fixed) [11]
Interest rate component	0.01% / 8h	0.01% / 8h	0.01% / 8h (actual payment of $F/8$ per 1h) [11]
Clamp range	±0.05%	±0.05%	±0.05%
Rate cap	±0.75% / 8h	Major assets such as BTC/ETH: ±0.75%/8h; long-tail assets can widen to ±2% to ±4% once dynamic settlement is triggered	±4% / 1h (uniform cap)
Index price source	Volume-weighted average across multiple centralized exchanges (CEXs)	Volume-weighted average across multiple CEXs	Validator oracle (weighted median of CEX prices; permissioned validator set, execution by a centralized sequencer) [11]

Table 10-2. Comparison of core funding-rate design parameters across major exchanges (Data source: compiled by the author from each platform's public documentation)

Settlement frequency is the core variable that changes the time scale of the game. BitMEX has long adhered to settlement every 8 hours (at 04:00, 12:00, and 20:00 UTC), a classic design that leaves arbitrageurs ample buffer time to deploy capital and execute strategies. In 2025, Binance introduced a dynamic settlement-frequency mechanism that allows flexible switching among 8-hour, 4-hour, and even 1-hour intervals depending on market volatility and the characteristics of a specific contract [10]. This adjustment was a direct response by Binance to rate-manipulation events such as ALPACA: a higher settlement frequency means a manipulator must maintain its position over a shorter time window, sharply raising the cost of manipulation. Decentralized exchanges (DEXs) such as Hyperliquid, meanwhile, adopt hourly settlement [11]. In theory, the higher the settlement frequency, the tighter the price anchoring, but it also significantly increases the ongoing cost pressure on position holders in one-sided trending markets. In certain extreme cases (such as the high-frequency-settlement squeeze analyzed in detail in Section 10.7.3), shorts can lose nearly 20% of their notional principal to funding payments alone within 24 hours.

Differences in the interest rate component and the clamp range directly determine the rate's baseline and volatility. Both BitMEX and Binance use a fixed 0.01% interest rate component and a ±0.05% clamp range, which creates the structural positive bias and stable band described earlier. Hyperliquid's rate calculation adopts a "compute first, then scale" normalization architecture [11]: the formula first computes a nominal rate F using the same parameters as 8-hour-settlement exchanges (interest rate component $I = 0.01\%$, clamp bounds $\pm 0.05\%$), then sets the actual hourly payment to $F/8$, so that within the clamp-bound stable band the default hourly rate is $0.01\%/8 = 0.00125\%$. Although the formula parameters match those of 8-hour exchanges, Hy-

perliquid's premium-index sampling window is 1 hour (sampled every 5 seconds and then averaged) rather than 8 hours. The shorter sampling window makes the premium component more sensitive to short-term price shocks, which partly explains the statistical feature of higher rate volatility on Hyperliquid. More importantly, Hyperliquid's rate cap is set at 4% per hour, far above the 0.75% per 8 hours typical of centralized exchanges (BitMEX). As a result, under extreme conditions, rates on Hyperliquid can reach levels far exceeding those on centralized exchanges, giving speculators and manipulators greater strategic room.

Differences in the source of the index price affect the reliability of the "anchor" itself. Centralized exchanges typically use a volume-weighted average price across several spot exchanges, supplemented by anomaly detection and dynamic weight adjustment. As an on-chain protocol, Hyperliquid relies on price feeds provided by its validator network, with each validator independently computing the weighted median of spot prices across multiple CEXs [11]. This design has advantages in decentralization and censorship resistance, but it also introduces an additional risk dimension of oracle latency and update frequency. On closer inspection, Hyperliquid's actual degree of centralization is higher than its decentralization narrative suggests: its validator set is permissioned (about 16–24 nodes, roughly 21 as of early 2026), and the Hyperliquid team controls a substantial share of the staking weight; order matching and rate calculation are executed by a centralized sequencer, making validators closer to "confirmers" than to independent consensus participants; and actual price-feed update latency is constrained by sequencer throughput. The JELLY incident of March 2025 exposed the fragility of this architecture, as the lag in the oracle price feed was exploited by a manipulator, causing significant market distortion.

These parameter differences are clearly reflected in empirical data. BitMEX's Q3 2025 data vividly illustrate how different design choices shape markedly different rate behavior [8]:

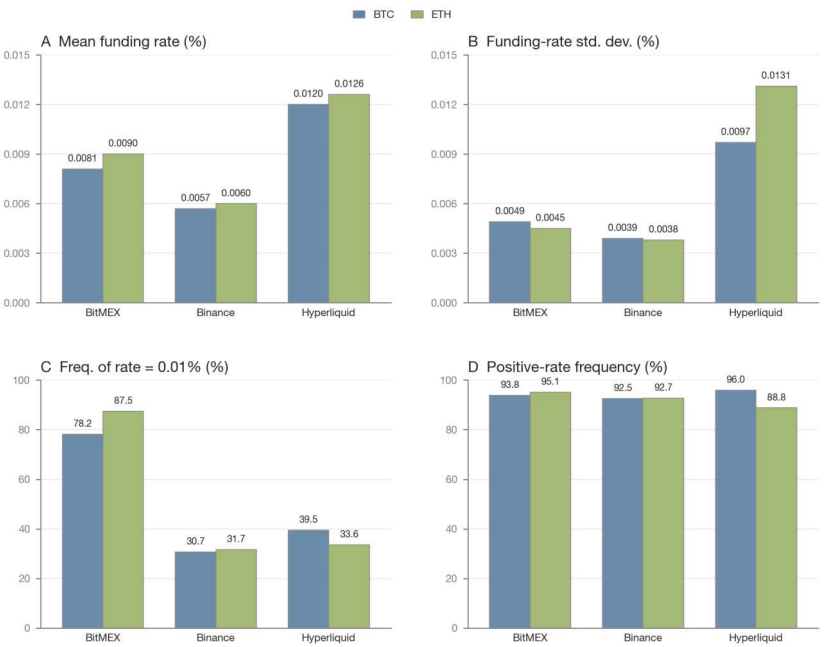


Figure 10-2. Comparison of the statistical characteristics of BTC and ETH funding rates on major exchanges in Q3 2025 [8]

As Figure 10-2 shows, the rate behavior of the three exchanges differs markedly. BitMEX's BTC rate was stable at 0.01% for 78.2% of the time (panel C), with a standard deviation of only 0.0049%, exhibiting high predictability. Binance, the world's largest exchange, had a mean rate (0.0057%) that was actually lower than BitMEX's (0.0081%), reflecting persistent short-side pressure on the Binance platform—possibly from the large numbers of hedge funds and market makers that establish short positions on Binance to hedge long exposure on other platforms. Hyperliquid's mean rate (0.0120%, panel A) and standard deviation (0.0097%, panel B) were both significantly higher than those of the two centralized exchanges, with a maximum of 0.0672%, exhibiting higher volatility and greater sensitivity to market sentiment [8].

Exchange	BTC mean rate	BTC standard deviation	Frequency of BTC rate = 0.01%	Frequency of positive BTC rate
BitMEX	0.0081%	0.0049%	78.19%	93.83%
Binance	0.0057%	0.0039%	30.70%	92.54%
Hyperliquid	0.0120%	0.0097%	39.45%	95.98%

Table 10-3. Core statistical indicators of BTC funding rates on three exchanges in Q3 2025 [8] (Data source: compiled by the author based on [8])

One shared feature stands out: despite the vast differences in design parameters and rate volatility across the three exchanges, the frequency of positive BTC rates exceeded 92% on all of them (Table 10-3). This means that throughout Q3 2025, on every platform, longs paid shorts more than 90% of the time. This phenomenon is not

accidental but the joint result of the institutional benchmark set by the interest rate component $I = 0.01\%$ and the structural positive bias of the clamp function: even in periods when the market's directional preference is relatively neutral, the "default state" of the rate is positive. In its report, BitMEX calls this phenomenon the "formula's secret positive bias" [8].

These cross-exchange parameter differences and statistical patterns are not minor technical details. They determine the thickness and response speed of arbitrage capital on each platform, the cost and feasibility of a "rate squeeze" attack by a manipulator, and the speed and magnitude of rate polarization during extreme conditions. More importantly, they constitute the initial conditions for all the analysis that follows in this chapter: the "three-layer decomposition model" of Section 10.2 explains why these parameter differences produce institutional friction layers of differing sizes; the "anchoring strength spectrum" of Section 10.3 argues how different designs affect the effectiveness boundary of anchoring; the "three-layer game model" of Section 10.4 analyzes how parameter differences alter the equilibrium structure of each layer of the game; and Section 10.7 examines settlement frequency as a core design dimension.

10.2 The three-layer decomposition model

In the crypto derivatives market, traders often treat the funding rate simply as a "directional indicator of long-short forces" or a "market sentiment gauge." This one-dimensional understanding, however, obscures the microstructural complexity of the funding rate. To accurately parse the market signal the funding rate conveys, we must abandon the view of it as a single variable and instead adopt a three-layer decomposition model. The model deconstructs the observed surface funding rate into three decomposable components with distinct economic meanings: the financing cost layer, the sentiment premium layer, and the institutional friction layer. The institutional friction layer, however, is not statistically orthogonal to the sentiment premium layer P but mechanically dependent on it through the clamp function.

10.2.1 The financing cost layer

The bottom layer of the funding rate is the financing cost layer, which represents the capital cost implied by holding a derivatives position under a risk-free-arbitrage assumption. This layer makes the funding rate of perpetual futures not merely a tool for adjusting a relative price, but also an expression of an absolute rate of return.

Before analyzing each layer, consider the architecture of the three-layer decomposition model as a whole. Figure 10-3 arranges the three components of the rate from lowest to highest by their frequency of variation.

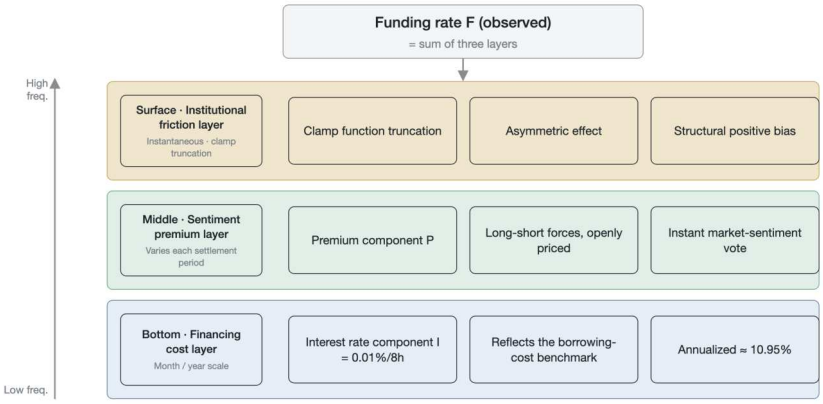


Figure 10-3. The three-layer decomposition model of the funding rate

As Figure 10-3 shows, the financing cost layer forms the base benchmark of the rate, maintaining a default state near 0.01% in the absence of extreme market deviation; the sentiment premium layer is superimposed on the financing cost layer and fluctuates in real time with shifts in long-short forces; and the institutional friction layer imposes a nonlinear truncation on the combined result of the first two layers through the clamp function, producing the structural positive bias of the rate distribution. The interaction of the three layers determines the actual observed rate at any moment.

In the funding rate formulas of most major crypto exchanges (such as BitMEX and Binance), the financing cost layer is hard-coded as a fixed interest rate component (usually denoted I). For example, BitMEX initially set the benchmark rate at 0.01% every 8 hours, equivalent to an annualized rate of about 10.95% (simple annualized, i.e., $0.01\% \times 3 \times 365$; compounded, it is about 11.57%, i.e., $(1 + 0.01\%)^{1095} - 1$) [8]. This setting is not an arbitrary mathematical constant but an empirical estimate based on the difference in borrowing costs between the U.S. dollar (or stablecoins) and Bitcoin in the early crypto market. It constitutes a form of institutional rigidity: after the Federal Reserve raised its benchmark rate to 5.25% in 2023, the premium of the crypto market's fixed 10.95% interest rate component over the traditional risk-free rate was sharply compressed (from more than 10 percentage points in the zero-rate era to about 5.7 percentage points), yet the formula itself was not dynamically adjusted.

The borrowing cost hypothesis holds that a buyer of perpetual futures is effectively borrowing the quote currency (such as USDT) to purchase the underlying asset (such as BTC), while a seller is borrowing the underlying asset to obtain the quote currency. The benchmark of the funding rate should therefore reflect the interest rate spread between these two assets in the spot lending market. Because demand for stablecoin borrowing in the crypto market has long exceeded demand for borrowing native crypto assets, this structural supply-demand imbalance has made stablecoin lending rates significantly higher than crypto-asset lending rates, which is then embedded at the formula level as a persistently positive interest rate component.

10.2.2 The sentiment premium layer

Above the financing cost layer sits the highly dynamic sentiment premium layer, the part of the funding rate that traders watch most closely. The sentiment premium layer is composed of the premium component (usually denoted P), which measures in real time the deviation between the perpetual futures price and the spot index price. The sentiment premium layer is the direct result of the game between long and short forces in the market microstructure and represents the open-market pricing of risk. When the market expects prices to rise, speculative long forces strengthen and push up the perpetual price, so that $P > 0$; conversely, when the market expects a decline, short forces dominate, so that $P < 0$. During the bull phase of November 2024, when Bitcoin broke \$100,000, the premium component of BTC perpetual futures stayed persistently in the elevated range of 0.05% to 0.15%, reflecting a concentrated influx of leveraged long demand; whereas during the AI-sector turbulence triggered by DeepSeek in late January 2025, the premium component of some altcoins plunged below -0.3%, revealing the impact of panic selling on the microstructure of the derivatives market.

In a normal market environment with sufficient arbitrage capital, the sentiment premium layer reflects the risk compensation that arbitrageurs demand for absorbing one-sided market imbalances. In their study of crypto carry, Schmeling et al. at the Bank for International Settlements note that the arbitrage returns of crypto futures are essentially a highly volatile "convenience yield," one arising from the trend-following behavior of retail investors chasing leveraged upside exposure and from the relative scarcity of arbitrage capital willing to bear basis risk [12]. The sentiment premium layer is therefore not merely a measure of sentiment but the pricing that market liquidity providers (arbitrageurs) place on the use of their capital.

10.2.3 The institutional friction layer

The component of the three-layer decomposition model that is often overlooked yet plays a significant role in actual operation is the institutional friction layer. This layer arises from the mathematical truncation mechanism—the clamp function—that exchanges deliberately introduce to prevent the funding rate from fluctuating excessively. BitMEX introduced this mechanism when it designed perpetual futures in 2016, with the aim of preventing extreme single-period rate fluctuations from triggering chain liquidations, while also providing arbitrageurs with a predictable return floor to sustain their participation. The final funding rate formula of most exchanges can be written as $F = P + \text{clamp}(I - P, -0.05\%, +0.05\%)$. This seemingly simple smoothing mechanism creates a significant asymmetric effect at the microstructural level.

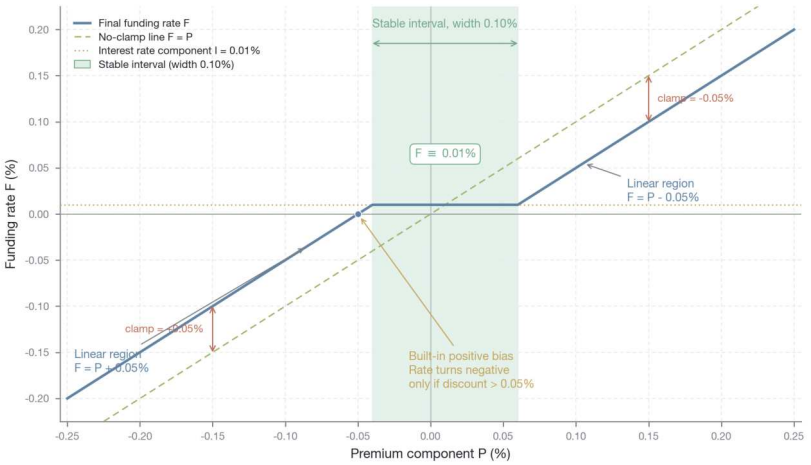


Figure 10-4. The asymmetric truncation effect of the clamp function on the funding rate and the resulting stable band (a conceptual illustration derived from the formula, not empirical data; the clamp thresholds of about ±0.05%, the interest rate component of 0.01%, and the stable-band width of 0.10% are all formula constants [8])

As Figure 10-4 shows, the clamp function creates a "stable band" 0.10% wide (when the premium component P lies between −0.04% and +0.06%). Within this band, the truncation term in the formula offsets fluctuations in the premium, forcing the final funding rate to be anchored at $I = 0.01\%$.

This mechanism produces a structural positive bias in the funding rate distribution. Even when the perpetual futures are at a slight discount (for example, $P = -0.03\%$), the intervention of the clamp function still requires longs to pay shorts a funding rate of 0.01%. Only when the discount exceeds −0.05% does the funding rate fall to zero; only when the discount is deeper still does the rate turn negative.

BitMEX's Q3 2025 derivatives report corroborates this structural feature (Figure 10-2): the BTC funding rate was exactly equal to 0.01% for as much as 78.19% of the time and remained positive 93.83% of the time [8], and even the more volatile Hyperliquid had a positive-rate share as high as 95.98%. This shows that the widespread "persistently positive rate" phenomenon in the market is largely not the result of sustained market optimism but of the institutional friction layer—that is, the formula's built-in positive bias.

10.2.4 Macro interest rate resonance and the economics of arbitrage

Having decomposed the funding rate, we can compare its financing cost layer with the macroeconomic environment to reveal the economic drivers of cross-market arbitrage. Figure 10-5 compares the divergence between the Federal Reserve's benchmark rate and the implied rate of perpetual futures from 2020 to 2025, along with the evolution of the resulting cross-market arbitrage spread.

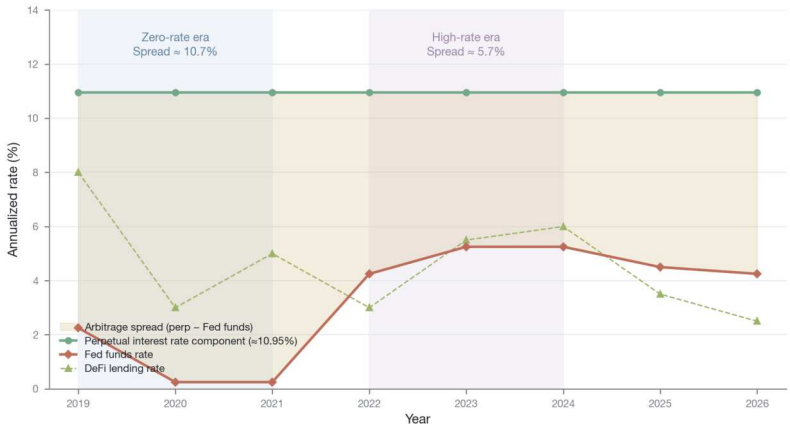


Figure 10-5. The divergence between the macro interest rate environment and the implied rate of perpetual futures, and the resulting arbitrage space (an illustrative trend chart, not a true month-by-month series; the Federal Reserve path is public macro data, the perpetual implied rate of about 10.95% annualized is an arithmetic formula constant, and the narrowing-spread trend is consistent with Bloomberg reporting [13])

As Figure 10-5 shows, the interest rate component of perpetual futures (about 10.95% annualized) was fixed at the outset. During the "zero-rate era" of 2020–2021, the Federal Reserve's benchmark rate was near zero and risk-free returns in traditional financial markets were extremely low. At that time, the crypto market's benchmark funding rate of more than 10% created a large cross-market arbitrage spread (exceeding 10%). This large spread attracted substantial traditional institutional capital (such as early crypto hedge funds) into the crypto market to execute cash-and-carry arbitrage, providing ample liquidity to the market.

Entering the "high-rate era" of 2023–2024, however, the Federal Reserve's benchmark rate climbed above 5.25% and traditional risk-free yields rose sharply. At this point, the advantage of the crypto market's fixed interest rate component over the traditional dollar risk-free rate was greatly compressed, and the arbitrage spread narrowed to around 5%. This change in the macro environment raised the marginal opportunity cost of arbitrage capital. As Bloomberg reported in early 2026, with basis yields falling to around 5%, Wall Street hedge funds began retreating from the Bitcoin basis trade [13].

Beyond the price signal of a narrowing spread, rising macro interest rates also weaken the anchoring strength of perpetual futures through two overlooked channels. The first is **structural diversion through the ETF/CME channel**: after the approval of BTC spot ETFs in January 2024, institutions could construct arbitrage using ETFs plus CME futures without bearing the counterparty risk of crypto-native exchanges. This is not merely a total retreat driven by a narrowing spread but a structural migration of arbitrage capital from crypto-native platforms to traditional financial infrastructure; institutions no longer need to bear exchange credit risk on Binance or Bybit to cap-

ture basis yield. The second is **quantity contagion and risk-appetite contagion**: rising macro interest rates weaken anchoring strength not only through the price signal (a narrowing spread) but also directly through quantity shocks (redemptions by liquidity providers, which force compliant funds to withdraw arbitrage positions) and risk-appetite contagion (a macro risk-off that drives down crypto prices and triggers a liquidation cascade). These three contagion channels—a narrowing spread, structural diversion of capital, and risk-appetite transmission—together explain why the arbitrage depth of the perpetual futures market has declined structurally since 2023.

This resonance and divergence of macro interest rates explains why the liquidity depth and the abundance of arbitrage capital in the crypto derivatives market differ significantly across macro cycles. Now that the Federal Reserve has begun its rate-cutting cycle in the second half of 2025, the decline in traditional risk-free yields will again widen the crypto market's arbitrage spread and may attract a return of the institutional capital that had previously retreated. Meanwhile, on-chain DeFi lending rates (such as the floating borrowing rate for USDC on Aave) provide a more real-time dynamic proxy for the financing cost layer, and the spread between them and the funding rate of perpetual futures constitutes a driver of cross-protocol arbitrage.

10.2.5 Empirical methods for separating the three layers

In empirical research and quantitative trading, accurately separating these three components is a prerequisite for constructing effective strategies. Quantitative researchers typically use the following methods to separate them (Table 10-4):

Component	Observation and separation method	Economic meaning	Trading strategy application
Financing cost layer (I)	The exchange's publicly disclosed fixed parameter (usually 0.01%/8h), or a dynamic proxy estimate via a DeFi stablecoin lending rate (such as the Aave USDC borrowing rate).	The time value of capital and the cross-market risk-free interest rate benchmark.	Serves as the benchmark return for cross-market arbitrage and helps assess the opportunity cost of arbitrage capital.
Sentiment premium layer (P)	Computed from high-frequency order-book data as the real-time basis between the perpetual futures (traded price/mid-price) and the spot index price.	Pure market-microstructure imbalance and one-sided speculative momentum.	Building momentum or mean-reversion strategies; identifying extreme sentiment-reversal points.
Institutional friction layer	In the frictionless limit, $F = P + (I - P) = I$ (the premium P cancels out), so the institutional friction term should be measured by the amount of truncation: friction = clamp($I - P$, -0.05%, $+0.05\%$)$-$ $(I - P)$, i.e., the extent to which the clamp truncates $(I - P)$.	The intervention and distortion of the market's natural clearing process by exchange rules.	Finding structural mismatches where the market is "at a discount but the rate is still positive" and building statistical-arbitrage strategies with a positive expectation.

Table 10-4. Empirical methods for separating the three components of the funding rate (Data source: compiled by the author)

The three-layer decomposition model is not the only framework for explaining the drivers of the funding rate. In the traditional finance literature, several competing theories exist for the pricing of the futures basis. The convenience yield hypothesis holds that holding the spot asset itself carries an implicit value (such as voting rights, staking rewards, or immediate availability), and this value should be reflected in a futures discount. In the crypto context, the convenience yield of holding spot BTC may include the opportunity cost of on-chain staking or DeFi participation. The hedging pressure hypothesis emphasizes that when producers (analogous to miners or token issuers in the crypto market) systematically short futures to lock in forward prices, the concentrated supply of shorts depresses the futures price and produces a negative basis unrelated to speculative demand. The market segmentation hypothesis notes that if participants in the spot and derivatives markets differ systematically in capital, information, or access, the prices of the two markets can diverge persistently without being arbitraged away. The three-layer decomposition model adopted in this chapter does not reject these theories but incorporates part of their insights into different layers:

the effects of convenience yield and hedging pressure are absorbed by the sentiment premium layer, while market segmentation is reflected indirectly through the arbitrage-impediment conditions (Section 10.3.3). A rigorous empirical test, however (such as a regression that decomposes the rate into convenience-yield and speculative-sentiment components), is left to future research.

Through the three-layer decomposition model, we recognize that the funding rate is not a freely floating price determined purely by market supply and demand but a hybrid indicator deeply shaped by institutional rules. In interpreting the funding rate signal, a trader must strip away the "pseudo-steady state" caused by the institutional friction layer and the "positive bias" caused by the financing cost layer in order to reach the true market-microstructure state reflected in the sentiment premium layer.

10.3 The anchoring strength spectrum

In traditional futures markets, the physical or cash delivery mechanism at expiry provides a kind of "hard anchor." No matter how far a contract deviates from the spot price during its life, at the moment of expiry the futures price and the spot price must forcibly converge, or else a risk-free, deterministic arbitrage opportunity would arise. Perpetual futures, however, have no expiry date, and the convergence of their price toward the spot price depends entirely on the funding rate mechanism. This mechanism, which relies on continuous payments, is called a "soft anchor."

The key to understanding the perpetual futures market is to recognize that this soft anchoring is not a binary, black-or-white state (either anchored or unanchored) but a continuous anchoring strength spectrum.

10.3.1 Defining anchoring strength

We can introduce a parameter α (anchoring strength) between 0 and 1 to quantify this concept. Operationally, α can be estimated from the half-life of the deviation of the perpetual futures price from the spot price: the shorter the half-life, the faster the deviation is corrected and the higher the anchoring strength. When $\alpha = 1$, anchoring is perfect, equivalent to the hard-delivery state of traditional futures at expiry, with the futures price F_t strictly equal to the spot price S_t. When $\alpha = 0$, there is no anchoring, and the perpetual futures degenerate into a purely speculative instrument entirely unconstrained by the spot price, with its price determined solely by the internal long-short game of the derivatives market. In actual perpetual futures markets, anchoring strength always lies on a continuous spectrum with $0 < \alpha < 1$. Operationally, α can be estimated with an error correction model. Let $b_t = P_t^{\text{perp}} - P_t^{\text{spot}}$ be the basis at time t; the dynamics of the basis can be modeled as:

$$\Delta b_t = -\kappa \cdot b_{t-1} + \varepsilon_t$$

where $\kappa > 0$ is the mean-reversion speed parameter and ε_t is a white-noise disturbance. The larger κ is, the faster the basis is corrected and the tighter the anchoring. The operational proxy for α is $\hat{\alpha} = 1 - e^{-\kappa}$, with a corresponding half-life of $h = \ln 2/\kappa$. In a normal market environment, based on high-frequency basis data from BitMEX and Binance, a preliminary estimate from an error correction model suggests that h is about

2 to 4 hours—that is, half of the basis is corrected by arbitrage activity within 2 to 4 hours (corresponding to a cumulative correction of about 75%–94% of the basis within 8 hours and about 94%–99.6% within 16 hours). During extreme events (such as Black Thursday in March 2020), however, h can exceed 48 hours (corresponding to $\alpha < 0.2$), indicating that the anchoring mechanism has almost completely failed. The estimated value of κ depends heavily on data frequency, and during extreme events ε_t exhibits fat tails, which may bias the OLS estimate. A more refined Markov regime-switching model can capture the switching of κ across three states—normal, stress, and crisis—though its estimation and testing are left to future research. The half-life parameters above should be understood as preliminary estimates based on descriptive analysis in this book, not as rigorous econometric test results.

In the overwhelming majority of normal market periods, α is close to 1 (for example, $\alpha \approx 0.9$), and the funding rate effectively constrains the perpetual futures price to near the spot price.

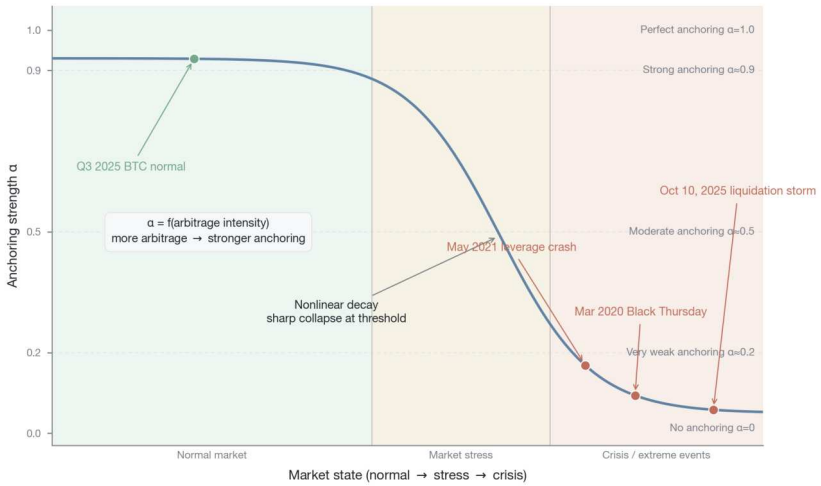


Figure 10-6. The anchoring strength spectrum and deviations during extreme events (a conceptual illustration, not empirical data; the anchoring strength α is a model parameter defined in this book and an author's estimate)

As Figure 10-6 shows, anchoring strength remains in the high range of 0.8 to 0.95 in a normal market environment, and the deviation between the perpetual futures price and the spot price is usually no more than 0.1%. When the market enters an extreme event (such as Black Thursday in March 2020 or the leverage crash of May 2021), anchoring strength plunges below 0.2, and the perpetual futures price deviates persistently from the spot price by hundreds of dollars. This nonlinear decay characteristic shows that the anchoring efficacy of the funding rate does not decline along a linear path but collapses abruptly at a particular threshold.

10.3.2 Arbitrage activity and anchoring strength

The funding rate formula is merely a calculation rule and has no direct power to pull prices. The formula does not create anchoring; it only creates the profit space for arbitrage. What truly creates anchoring is the arbitrage capital drawn in by that profit. When the perpetual futures are at a premium ($P > 0$), the funding rate is positive. Arbitrageurs construct a market-neutral portfolio by "shorting the perpetual + buying spot," thereby collecting the funding rate risk-free. This operation generates sell pressure in the derivatives market (depressing the contract price) and buy pressure in the spot market (raising the spot price), physically forcing the two prices to converge.

Ackerer, Hugonnier, and Jermann (2025) propose the AHJ framework [14] and prove this under the assumption of a frictionless, continuous-trading market. They show that the price of perpetual futures can be expressed as the risk-neutral expectation of the spot price sampled at random time points that reflect the strength of price anchoring. This anchoring strength, at the microstructural level, depends primarily on the smoothness of dynamic trading and the replication of primitive securities—that is, on the efficiency of arbitrage activity.

We can therefore draw a core conclusion: α is a monotonically increasing function of the intensity of arbitrage activity. This conclusion is highly consistent with the intermediary asset pricing theory of Gabaix and Maggiori (2015) [15]. In their framework, the degree of asset-price deviation depends not on changes in fundamentals but on the capital adequacy of financial intermediaries (in this context, arbitrageurs): when intermediary capital contracts, price deviations widen sharply even if fundamentals are unchanged. The anchoring strength of perpetual futures is precisely a direct mapping of this theory onto the crypto derivatives market: the capital condition of arbitrageurs determines the effectiveness boundary of the soft anchor. This can be formalized as $\alpha = g(A, L, \sigma)$, where A is the total amount of active arbitrage capital in the market, L is the liquidity depth for cross-market execution, and σ is the current volatility level. As long as arbitrage channels are unobstructed, capital is abundant, and volatility is within a controllable range, the funding rate can anchor prices effectively; but once A plummets because of capital retreat, L dries up because market makers pull their orders, or σ spikes above arbitrageurs' risk-control thresholds, anchoring strength slides rapidly toward 0.

10.3.3 Conditions that impede arbitrage

If arbitrage activity is the sole source of anchoring strength, then the question of "why the rate fails" becomes "why arbitrage is impeded." In extreme market turmoil, arbitrageurs often face the following six impediment conditions, which prevent them from executing the arbitrage that should flatten the basis and thereby trigger an anchoring collapse.

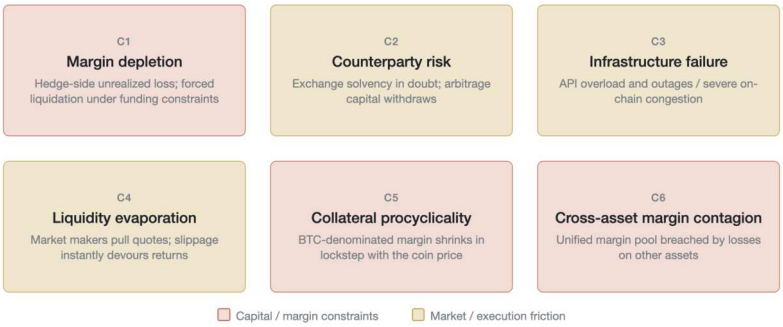


Figure 10-7. The six arbitrage-impediment conditions that cause the anchoring of perpetual futures to collapse

Margin depletion is the primary impediment. In a violent one-sided market, an arbitrageur's hedged position may face large paper losses. Although the overall portfolio is market-neutral, the paper losses on the derivatives leg require continuous margin top-ups. If an arbitrageur faces a liquidity constraint and cannot replenish fiat or stablecoin margin in time, it risks forced liquidation. This mechanism corresponds closely to the spiral between market liquidity and funding liquidity proposed by Brunnermeier and Pedersen (2009) [16]. In their theoretical framework, a decline in asset prices tightens traders' margin constraints, forcing them to cut positions; the position-cutting further depresses prices, forming a self-reinforcing deleveraging loop. Margin depletion among arbitrageurs in the crypto market is precisely the concrete manifestation of this funding-liquidity spiral in the perpetual futures market: arbitrageurs exit not because their strategy has failed but because their funding constraint has been breached. At the same time, counterparty risk intensifies. When systemic risk emerges, arbitrageurs worry about the solvency of the exchange itself. If they expect the exchange might go bankrupt or restrict withdrawals, arbitrage capital will choose to withdraw to protect principal no matter how high the funding rate, creating a vacuum of arbitrage force. In addition, infrastructure failures tend to erupt at critical moments. Arbitrage relies on high-frequency, low-latency cross-market operations, and in extreme conditions an exchange's API may become overloaded and go down, or the underlying blockchain network may become severely congested. Liquidity evaporation further worsens the arbitrage environment: as panic spreads, market makers cancel orders to avoid risk, leaving the order book extremely thin. At this point, arbitrageurs face enormous slippage costs when establishing or unwinding spot and derivatives positions, which can instantly wipe out months of funding-rate returns and make arbitrage economically infeasible. Collateral procyclicality further aggravates the situation: in the early, BTC-margined perpetual futures, when the BTC price crashed, short arbitrageurs faced not only the pressure of a falling contract price but also a simultaneous shrinkage in the fiat value of the BTC they used as margin. This double effect significantly amplified liquidation risk, forcing arbitrageurs to close positions passively at the very moment the market most needed liquidity.

Beyond these five conditions, **cross-asset margin contagion** constitutes a sixth, often-overlooked arbitrage-impediment condition. Under the unified margin mode of exchanges such as Binance and Bybit, an arbitrageur's BTC cash-and-carry position may share a margin pool with positions in other assets. When another asset (such as LUNA in 2022) collapses, losses on that asset erode the unified margin, forcing the arbitrageur to close out an otherwise healthy BTC arbitrage position. This contagion path means that the anchoring strength of BTC can fall unexpectedly during a non-BTC event: arbitrageurs exit not because of pressure in the BTC market itself but because they are breached by risk spillover from other assets. During the LUNA/UST collapse of May 2022, some arbitrage funds operating under the unified margin mode were forced to liquidate their market-neutral positions in both BTC and ETH simultaneously, providing empirical support for this contagion path.

10.3.4 The AHJ framework: applications and limitations

The AHJ framework lays a solid mathematical foundation for the no-arbitrage pricing of perpetual futures in theory, proving that under the assumption of a perfect (frictionless, continuous-trading) market, a specific funding rate setting makes the perpetual futures price converge to the spot price [14].

When this theoretical framework is applied to the real crypto market, however, its limitations become fully apparent. The core assumptions of the AHJ model are a frictionless market and continuous trading capability. As the arbitrage-impediment conditions analyzed above show, the crypto market exhibits precisely the opposite at extreme moments: extremely high friction (slippage, congestion) and discontinuity of trading (outages, circuit breakers).

The AHJ framework therefore depicts the ideal state of $\alpha \approx 1$ in Figure 10-6. It explains why the funding rate can operate effectively during the 99% of the time that is normal, but it cannot explain why a deep price dislocation arises during the 1% "crisis moments." The no-arbitrage condition exhibits significant fragility when confronting a systemic liquidity shock.

10.3.5 The anchoring paradox

Taken together, these observations point to a core microstructural paradox of the crypto derivatives market: the anchoring paradox. When the market is calm and volatility is low, the natural deviation between the perpetual futures price and the spot price is already small. At such times, arbitrage channels are unobstructed, liquidity is ample, the funding rate mechanism operates effectively, and anchoring strength is extremely high. Yet at such times the market does not truly "need" a powerful anchoring mechanism, because prices themselves have no impetus to derail.

Conversely, when the market suffers a black swan event, volatility spikes, and derivatives prices deviate sharply from spot prices out of panic or greed, this is precisely the moment when the market most needs the funding rate to anchor. But it is exactly at this moment that, because the six arbitrage-impediment conditions are triggered simultaneously, arbitrage capital retreats on a large scale or is even forcibly liquidated,

causing the transmission mechanism of the funding rate to break down entirely and anchoring strength to fall to its lowest level.

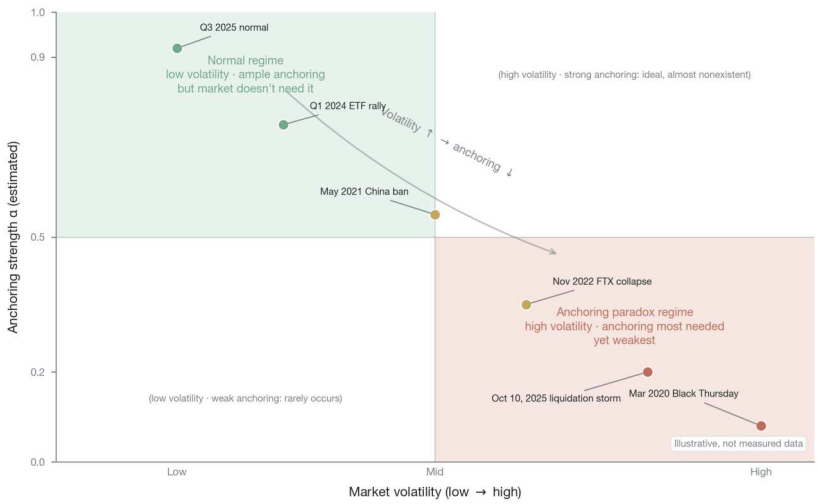


Figure 10-8. The inverse relationship between market volatility and anchoring strength (a conceptual illustration, not empirical data; the anchoring strength α is an estimate, the volatility axis is a qualitative scale of low, medium, and high, and the event points are qualitatively positioned)

As Figure 10-8 shows, the historical extreme events of the crypto market fully validate this paradox.

During the Black Thursday crash of March 12, 2020, the price of Bitcoin was cut in half in a single day. A post-mortem report by Multicoin Capital notes that, because of extreme congestion on the blockchain network and evaporated liquidity on BitMEX, arbitrageurs could not deposit BTC into BitMEX to flatten the spread [5]. Only about $20 million in bids remained on the entire BitMEX order book, yet it faced more than $200 million in long liquidations. The funding rate mechanism failed completely, causing the perpetual futures price on BitMEX to fall hundreds of dollars below other spot exchanges at one point (a discount of more than $500 at its peak, as noted in the chapter introduction) and nearly flash-crashing to zero.

Similarly, during the $19 billion liquidation storm of October 10, 2025, geopolitically induced panic caused liquidity across the entire market to evaporate instantly. Arbitrage capital was powerless in the face of extreme slippage and high network gas fees, causing the anchoring function of the funding rate to fail briefly and the basis to widen substantially [17].

The anchoring paradox reveals an inherent flaw of the funding rate as a "soft anchor": its effectiveness depends on market liquidity, yet it cannot itself create liquidity when liquidity dries up. This paradox is, in essence, a precise mapping of the "limits of arbitrage" theory of Shleifer and Vishny (1997) onto the crypto derivatives market [18]. Shleifer and Vishny note that real-world arbitrageurs are not the unconstrained agents of textbooks; they rely on external capital, and capital providers tend to withdraw funds precisely when arbitrage is most needed, because paper losses are most severe at that

time. The anchoring paradox of perpetual futures is exactly the crypto version of this logic: arbitrageurs easily earn rate income when the market is calm but cannot perform their price-correction function under extreme deviations because of capital constraints and execution friction. This is the fundamental game-theoretic reason why the funding rate "works most of the time but occasionally fails at the most critical moment."

Several mechanism-design paths might alleviate the anchoring paradox, though each introduces new trade-offs. At least three are worth discussing. The first is **the insurance fund as arbitrageur of last resort**: the exchange actively executes cash-and-carry arbitrage during extreme basis events (as Hyperliquid's HLP vault plays a similar role in normal markets), but this turns the exchange from a neutral rule-setter into a direct market participant, introducing new conflicts of interest, and the size of the insurance fund may be far too small to fill the arbitrage vacuum in a systemic crisis. The second is **an extreme-basis circuit breaker**: pausing the liquidation engine when the basis exceeds a threshold, but pausing liquidations may cause greater risk accumulation (the paper losses of unliquidated positions keep expanding), and an intervention requiring human judgment is difficult to implement in a decentralized protocol. The third is **dynamic rate-cap increases**: automatically raising the rate cap in extreme states to strengthen negative feedback. But as the ALPACA case in Section 10.6.3 shows, an extreme rate can itself become a manipulation tool rather than a corrective force. Fully resolving the anchoring paradox may require introducing additional stabilization mechanisms beyond the rate mechanism, which constitutes an open research frontier in derivatives mechanism design.

10.4 The three-layer game model

The funding rate is not merely the result of a mathematical calculation but the product of a continuous game among multiple participants within a specific rule framework. To understand deeply how this mechanism operates in the crypto derivatives market, we construct a "three-layer game model." The model systematically deconstructs the formation and fragility of the funding rate's dynamic equilibrium at three levels: the microstructural confrontation of positions, the meso-level of cross-market arbitrage, and the macro-level of rule attack and defense.

10.4.1 The anchoring game

At the most microstructural level, the funding rate is essentially a near-zero-sum "war of attrition" between the long and short camps. When the perpetual futures price exceeds the spot index price, the funding rate is positive and longs must pay shorts periodically; the reverse holds otherwise. This mechanism is designed to force prices back to their spot anchor by raising the holding cost in the direction of the deviation.

Figure 10-9 arranges the three games in a concentric-circle structure by participant scope and time scale.

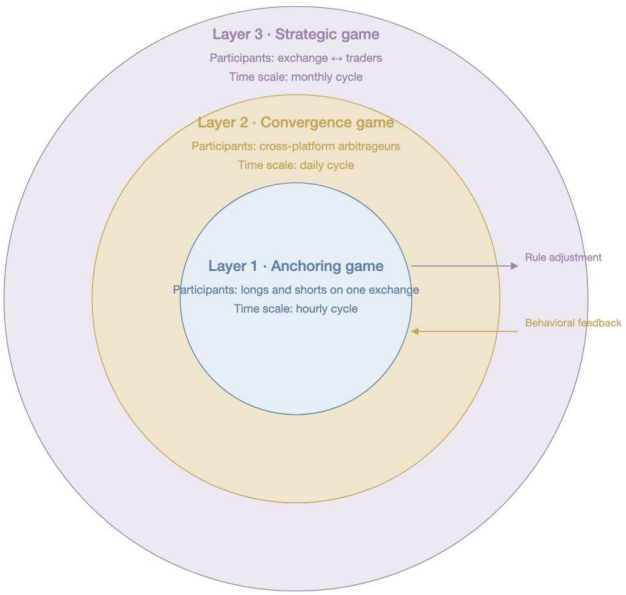


Figure 10-9. The nested structure of the three-layer game model of the funding rate

As Figure 10-9 shows, the three games exhibit a nested structure from the inside out: the innermost anchoring game determines the balance of long-short forces and the direction of the basis within a single exchange; the middle convergence game pulls prices and rates across platforms toward consistency through cross-market arbitrage; and the outermost strategic game determines the evolutionary path of the rules themselves. Feedback runs bottom-up among the three layers: imbalance in an inner game propagates outward, while rule adjustments in an outer layer alter the game structure of the inner layers.

To characterize the equilibrium structure of the anchoring game precisely, we model it as a 2×2 normal-form game. Consider a simplified single-period game in which longs and shorts each choose to "hold" or "exit." Let F be the current funding rate (paid by longs to shorts when positive), Δ the directional payoff expectation (positive favors longs), and λ the liquidation-risk loss triggered by a counterparty's exit. The payoff matrix is shown in Table 10-5:

	Short: hold (H)	**Short: exit (E)**
Long: hold (H)	$(\Delta - F,\ -\Delta + F)$	$(\Delta + \lambda,\ -\lambda)$
Long: exit (E)	$(-\lambda,\ F + \lambda)$	$(0,\ 0)$

Table 10-5. The 2×2 normal-form payoff matrix of the anchoring game (Data source: compiled by the author)

The Nash equilibrium of this game depends on the size of the rate F relative to the directional payoff Δ and the liquidation risk λ. When $F - \lambda \leq \Delta \leq F + \lambda$ (i.e., $|\Delta - F| \leq \lambda$), longs and shorts are mutual best responses, and (hold, hold) constitutes the unique pure-strategy Nash equilibrium (the upper bound $\Delta \leq F + \lambda$ is the short's tolerance condition and is the economically more binding side); both sides have an economic incentive to maintain their positions, and the anchoring game is in a steady state. When the rate rises to $F > \Delta + \lambda$, the long's best response to the short holding becomes to exit (because $-\lambda > \Delta - F$), while the short's best response to the long exiting is to hold (because $F + \lambda > 0$); the two are mutual best responses, and the equilibrium jumps to (exit, hold), which is precisely the condition under which the funding rate's negative feedback mechanism takes effect. When the directional payoff expectation becomes extreme ($\Delta \quad F$), however, the long's incentive to hold becomes unshakable, and even a persistently rising rate cannot trigger an equilibrium jump. In this state, the game slides from the "both hold" steady state into a positive-feedback regime, and the funding rate loses its corrective function. The analysis above assumes a single-period static game; in a multi-period dynamic game, reputation effects, learning, and strategic adaptation would further enrich the equilibrium structure, though its core dependence on the rate threshold would remain unchanged.

The analysis above also treats the directional payoff expectation Δ as an exogenous, commonly known parameter, but in practice longs and shorts hold systematically biased estimates of it. The heterogeneous-beliefs model of Scheinkman and Xiong (2003) [19] shows that overconfident longs systematically overestimate Δ, while shorts in a panic systematically underestimate it. More importantly, Δ is itself endogenous to the game: longs adding leverage and driving up the price raises the realized value of Δ ex post, forming a self-fulfilling prophecy. Treating Δ as an exogenous parameter obscures this reflexivity: longs and shorts do not face the same Δ but different values Δ_L and Δ_S weighted by their respective subjective probabilities, and these subjective beliefs are partly self-validating through their effect on the market price. This methodological qualification does not affect the qualitative conclusions of the equilibrium analysis above, but it indicates that a complete theoretical treatment requires a dynamic game framework with heterogeneous beliefs.

In actual operation, however, this war of attrition often exhibits strong reflexivity [20]. The noise-trader-risk model of De Long, Shleifer, Summers, and Waldmann (1990) provides a theoretical basis for understanding this phenomenon [21]. In their framework, even when rational arbitrageurs know that a price has deviated from fundamentals, the irrational behavior of noise traders may push the deviation further in the short run, exposing arbitrageurs to the risk of "being liquidated while on the right side." In the perpetual futures market, this noise-trader risk is sharply amplified by high leverage: arbitrageurs must bear not only the risk that the price deviates further but also the cash-flow pressure of the rate each settlement period, which can lead rational arbitrageurs to stay out even when the rate is extreme. In a bull cycle, longs expect the asset's price appreciation to far exceed the funding-rate cost they pay. Thus, even facing high holding costs, longs not only refuse to close positions but may use paper profits to add further leverage. This behavior widens the contract premium further, which in turn pushes the funding rate higher. At this point, the high funding rate is no longer

a resistance that suppresses speculation but instead becomes a "confirmation signal" of the market's extreme optimism, attracting more speculative capital and forming a self-reinforcing loop of "rising price → rising rate → reinforced expectations → further position-adding."

The outcome of this war of attrition depends on the limits of each side's tolerance for cost. Although shorts can continuously collect risk-free rate income, they must bear the liquidation risk from further price increases. When market sentiment is extremely euphoric, the short camp may be forced to close positions because it cannot withstand the pressure of margin calls; this not only causes the price to spike instantly but also leaves the funding rate without a "counterparty" in the short term, further distorting the anchoring mechanism.

10.4.2 The convergence game

Without the intervention of an external force, the long-short war of attrition of the first layer easily falls into extreme imbalance. The core participants of the second-layer game are funding rate arbitrageurs, who, through cross-market operations, objectively bring about a dual equilibrium—price convergence and rate convergence—while pursuing risk-free returns.

Price convergence is achieved mainly through "cash-and-carry arbitrage." When the perpetual premium is too high (a positive funding rate), an arbitrageur buys the asset in the spot market while shorting an equal-size position in the perpetual futures market. This operation not only locks in the risk of price fluctuations but also steadily earns the funding rate. The arbitrageur's short-selling directly adds short-side force to the contract market, depressing the contract price and pushing it back toward the spot price.

Rate convergence is achieved through "cross-exchange arbitrage." Because different exchanges differ in liquidity depth, user composition, and rate-calculation details, the funding rate for the same asset often differs across platforms. Arbitrageurs short on the exchange with the higher rate and go long on the exchange with the lower rate, earning the rate spread between them. This capital flow ultimately flattens the rate differences across platforms. This convergence is limited, however, by the friction of transferring funds across platforms: transfers between CEXs involve withdrawal fees, on-chain confirmation delays (BTC takes 30–60 minutes), and anti-money laundering (AML) screening; transfers between CEXs and DEXs additionally involve the smart-contract risk of cross-chain bridging. These frictions determine the existence of a "no-arbitrage band": rate differences within the cost of cross-platform transfer cannot be eliminated by arbitrage, constituting the equilibrium width of the rate difference rather than a transient state converging to zero. According to Boros Finance's 2025 empirical research, cross-exchange rate arbitrage in BTC and ETH between Hyperliquid and Binance can achieve weighted-average fixed annualized returns of 5.98% to 11.4%, with peak returns in certain extreme periods exceeding 23% [22]. Because Boros is a provider of funding rate swap products, it reports gross returns before execution costs; actual net returns must deduct cross-exchange fund-transfer costs, the capital-efficiency loss from locking margin on two exchanges, slippage, and position-rebalancing friction, and

may be significantly lower than the nominal figures.

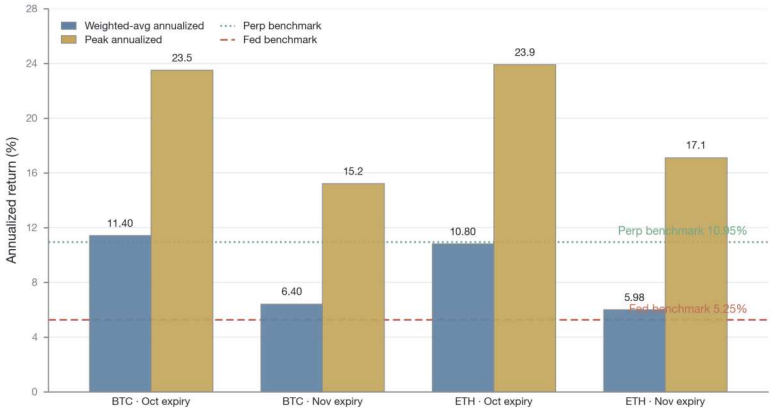


Figure 10-10. Annualized returns of cross-exchange funding rate arbitrage for BTC and ETH (gross returns before execution costs; BTC and ETH correspond to different maturities and are not a single-asset constant) [22]

As Figure 10-10 shows, the cross-exchange rate arbitrage returns for BTC and ETH exhibit pronounced time variation. The weighted-average annualized return for BTC was 11.4% (for the October 31 maturity) and 5.98% for ETH (for the November 28 maturity); note that the two correspond to different maturities and are not a single-asset constant, and the report shows gross returns before execution costs. Returns peaked in certain periods (exceeding 23%), typically corresponding to a widening cross-platform rate spread driven by extreme market sentiment. These high-return windows provide an economic incentive for arbitrage capital to enter, objectively promoting cross-platform convergence of the rate.

10.4.3 The strategic game

The third-layer game takes place between the rule-setters (exchanges) and the rule-adapters (traders). The exchange's goals are to maintain stable anchoring of the contract price, guard against systemic risk, and maximize trading volume; the trader's goal is to find loopholes within the rule framework to maximize profit.

To prevent excessive funding-rate fluctuations from causing large-scale chain liquidations, major exchanges generally set a rate clamp function and rate caps/floors. For example, BitMEX's empirical data show that under its clamp mechanism, the funding rate was fixed at the default 0.01% (every 8 hours) for as much as 78.19% of the time, equivalent to a benchmark annualized rate of about 10.95% [8]. Although such rules reduce day-to-day fluctuations, they also create space for rule arbitrage.

As the market has matured, this strategic game has intensified. For example, when some exchanges raised the settlement frequency from every 8 hours to every 1 hour or even real-time settlement, high-frequency traders quickly developed a "second-scale front-running" strategy, entering positions momentarily before settlement and exit-

ing immediately afterward to capture rate income at no cost. As a countermeasure, exchanges (such as Binance in its September 2025 algorithm update) introduced a frequency normalization factor to eliminate this time-arbitrage space by smoothing the calculation window. This dynamic game of "rule-loophole discovery → strategy exploitation → rule patching" constitutes the underlying driver of the funding rate mechanism's evolution.

10.4.4 State dependence and the decay of anchoring efficacy

The validity of this three-layer game model is not fixed but highly dependent on the current market state. As the market moves from calm to extreme, the anchoring efficacy of the funding rate exhibits a pronounced nonlinear decay (Table 10-6).

Market state	Basis characteristics	Core dominant force	Assessment of anchoring efficacy	Representative historical case
Normal state	Very small ($< \pm 0.1\%$)	Rate arbitrageurs, market makers	Strong: arbitrage capital is ample, small deviations are quickly flattened, and the funding rate stays near its default benchmark.	Most of the calm trading periods in 2024–2025
Stress state	Significant ($\pm 0.1\%$ to $\pm 1\%$)	Speculators (one-sided sentiment) vs. arbitrageurs	Weakened: one-sided sentiment widens the arbitrage space, but the execution constraints on arbitrage capital begin to appear, and the rate shows sustained positive or negative polarization.	The August 2024 yen carry-trade unwind shock; the AI-sector turbulence triggered by DeepSeek in late January 2025
Crisis state	Extreme ($> \pm 1\%$)	Liquidation mechanism (forced liquidation)	Failed: arbitrage channels break down entirely because of liquidity drought or network congestion, the liquidation mechanism takes over the market, and the rate mechanism no longer anchors but may instead accelerate the collapse.	Black Thursday on March 12, 2020; the leverage crash of May 19, 2021 [23]

Table 10-6. State-dependent decay of the funding rate's anchoring efficacy (Data source: compiled by the author)

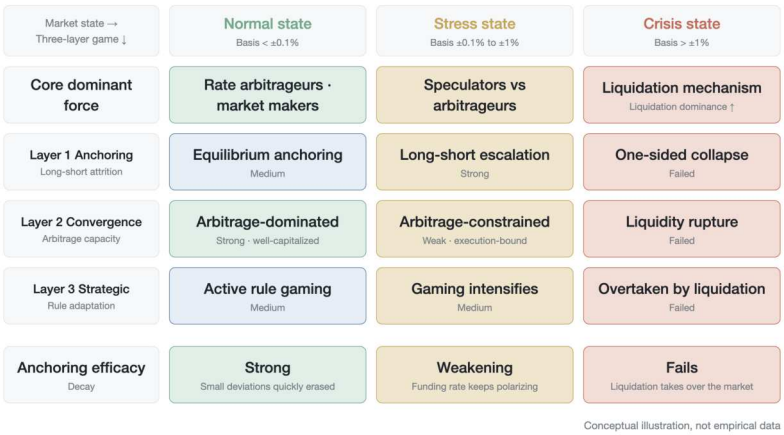


Figure 10-11. The relative strength of the three layers of game forces and the decay of anchoring efficacy across market states (a conceptual illustration, not empirical data; the strength of each layer is a qualitative rating of strong, medium, weak, or failed, and the basis thresholds are drawn from Section 10.4.4)

In the normal state, the second-layer convergence game (arbitrageurs) dominates, and the funding rate effectively anchors. When the market enters a crisis state, however (such as the May 19, 2021 crash in which the crypto market plunged 30% and more than $8 billion was liquidated market-wide [23]), arbitrageurs cannot execute cross-market operations because of spot-liquidity drought and severe network congestion. At this point, the second-layer game fails entirely, the first-layer long-short war of attrition devolves into one-sided chain forced liquidations, and the funding rate not only fails to pull prices back but its extreme negative values instead become a gauge of market panic.

In the crisis state, when the funding rate's anchoring mechanism fails and the liquidation cascade unfolds fully, the exchange's insurance fund and auto-deleveraging mechanism become the last line of defense for maintaining market continuity. The core function of the insurance fund is to absorb shortfall losses: when a position's liquidation price is worse than its bankruptcy price, the difference is borne by the insurance fund, thereby avoiding passing the shortfall on to the winning side. During Black Thursday on March 12, 2020, BitMEX's insurance fund drew down by about $8 million to $13 million (roughly 1,600 to 2,600 BTC) within a few hours, yet this was still insufficient to cover all shortfall losses [5]. When the insurance fund is exhausted, the auto-deleveraging (ADL) mechanism is triggered: the system automatically reduces counterparty positions according to a specific ranking algorithm to make up the shortfall. The ADL ranking algorithms of different exchanges differ significantly: BitMEX and Binance use a two-factor ranking of profit percentage and effective leverage,

whereas Hyperliquid first uses HLP vault funds to absorb shortfall losses and triggers ADL against traders only when the HLP is insufficient. Systematic public data on the historical trigger frequency of ADL are currently lacking (flagged as an empirical-data gap), but Hyperliquid's on-chain transparency provides a data foundation for future empirical research.

The ADL mechanism fundamentally changes the payoff function of the game. In an ideal world without ADL, a trader's return is a deterministic function of directional judgment and leverage choice. But the existence of ADL means that even a trader who is right about direction and maintains prudent leverage may have profitable positions forcibly reduced during a liquidation cascade. The shortfall losses of a "neighbor who is right about direction but over-leveraged" are transmitted to the winning side through ADL. This creates a unique negative externality: the risk-taking of high-leverage speculators threatens not only their own survival but also, through the ADL mechanism, the return certainty of prudent traders. For the equilibrium of the anchoring game, the existence of ADL weakens arbitrageurs' incentive to participate: although arbitrageurs construct market-neutral portfolios, their profit leg still faces the tail risk of being forcibly reduced by ADL. This mechanism-design tension—the conflict between protecting system continuity and incentivizing arbitrage participation—remains an unresolved open problem in the mechanism design of perpetual futures.

Moreover, under a unified margin system, the rate polarization of one asset affects not only the game equilibrium of that asset itself but may also be transmitted to other assets through a shared margin pool. For example, when extreme negative rates on SOL cause SOL shorts to be liquidated, if those shorts hold ETH longs in the same margin pool, the ETH longs may also be liquidated by association, even if the ETH market itself shows no anomaly. This cross-asset contagion mechanism is developed further in the liquidation-cascade analysis of Chapter 11.

10.4.5 The structural boundaries of arbitrageurs

If arbitrageurs are the core force sustaining the funding rate's anchoring mechanism, why do they collectively "disappear" during extreme conditions? This stems from three structural boundaries that arbitrageurs face. As Figure 10-12 shows, all three constraints are loose in a normal market and arbitrage activity is unimpeded; when the market enters an extreme state, all three constraints tighten simultaneously and arbitrage capacity shrinks sharply.

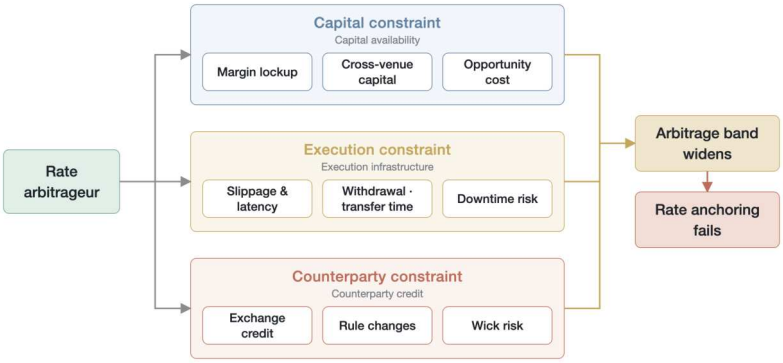


Figure 10-12. The three structural constraints faced by funding rate arbitrageurs

Gromb and Vayanos (2002) rigorously prove, from a general-equilibrium perspective, the systemic impact of capital constraints on arbitrageur behavior [24]. In their model, when arbitrageurs face funding constraints, the degree to which market equilibrium deviates from no-arbitrage pricing depends on arbitrageurs' wealth level: the poorer the arbitrageurs, the larger the price deviation and the more severe the welfare loss. This conclusion applies directly to the perpetual futures market: the persistent deviation of the funding rate (i.e., the decline in anchoring strength) essentially reflects the inadequacy of arbitrage capital relative to speculative demand.

These three constraints correspond mechanistically to the six arbitrage-impediment conditions of Section 10.3.3; rather than restate each one's operational details, we note only the core of each constraint. The capital constraint stems from the capital-intensity of cash-and-carry and cross-exchange arbitrage: arbitrageurs must reserve large redundant margins across multiple platforms to guard against one-sided liquidation, so capital efficiency is significantly reduced, and when the arbitrage space widens while idle capital runs out, the rate deviation persists. The execution constraint manifests as rate-limiting, latency, or even outages of exchange APIs during extreme conditions (such as the service disruptions at some leading exchanges on May 19, 2021 [25]), as well as delayed fund transfers caused by on-chain congestion and spiking gas fees. The counterparty constraint is the arbitrageur's passive exposure to the credit risk of centralized exchanges and to unilateral rule changes (such as forcibly rolling back trades, adjusting rate caps, or suspending withdrawals). These tail risks lead rational arbitrageurs to withdraw and wait on the sidelines precisely when the market most needs liquidity supply.

Together, these three constraints determine the capacity ceiling of the funding rate's anchoring mechanism. In quantitative terms, this ceiling depends on the total capital arbitrageurs can deploy, the speed of cross-platform fund transfers, and the credit rating of exchanges. In the BTC perpetual futures market, thanks to the participation of systematic arbitrage entities such as Ethena, arbitrage capacity is usually ample, whereas in the long-tail altcoin market, arbitrage capacity may be only a few million dollars, and a single medium-sized directional trade can breach this ceiling. When the

corrective demand generated by a market deviation exceeds arbitrageurs' structural capacity, the equilibrium of the three-layer game model tends to collapse.

10.5 Rate polarization and positive feedback

In classical financial theory, the funding rate is designed as a negative feedback system: when the contract price deviates from spot, a rising rate increases the holding cost on the deviating side, forcing it to close positions or attracting arbitrageurs to enter, ultimately returning the price to spot and pulling the rate back down. In the extreme conditions of the crypto market, however, this negative feedback mechanism can flip and evolve into a self-reinforcing positive feedback system, driving the rate into extreme polarization.

10.5.1 Trigger conditions for positive feedback

The mathematical trigger condition for the flip from negative to positive feedback can be formalized as the following proposition.

Proposition (necessary condition for continued position-holding). Let $E_t[\Delta P_{t+1}]$ be a trader's conditional expectation at time t of the next-period price change, F_t the current funding rate, γ the trader's risk-aversion coefficient, and $\mathrm{Var}_t[\Delta P_{t+1}]$ the conditional variance of the price change. When the following condition holds, a rational trader chooses to maintain the position and the funding rate's negative feedback mechanism fails:

$$E_t[\Delta P_{t+1}] > F_t + \gamma \cdot \mathrm{Var}_t[\Delta P_{t+1}]$$

The left side of this inequality is the expected return of holding the position, and the right side is the total cost of holding, comprising the deterministic rate outlay F_t and the risk-aversion-adjusted directional risk premium $\gamma \cdot \mathrm{Var}_t[\Delta P_{t+1}]$. In the simplified version (setting $\gamma = 0$, i.e., the risk-neutral assumption), the condition reduces to $E_t[\Delta P_{t+1}] > F_t$—that is, the expected directional profit merely needs to exceed the rate cost. The full version introduces the risk-aversion coefficient, meaning that even if the expected return exceeds the rate, a risk-averse trader may still choose to exit if the directional risk is too large (high Var_t). When the market narrative is highly certain (such as the ALPACA delisting announcement), traders tend to establish positions immediately, and positive feedback is more easily triggered. This analysis requires a finer distinction, however: a delisting announcement provides directional certainty (the price should fall) but increases uncertainty in magnitude and time path, and the price path before delisting may exhibit violent fluctuations and a short squeeze (as events indeed proved). The "focusing illusion" of behavioral finance (Kahneman 2011 [26]) may better explain this phenomenon: traders focus excessively on the directional narrative ("delisting = crash") while underestimating path risk, causing them to systematically overestimate the net expected return of establishing a position immediately.

The inequality above characterizes only the static decision condition under which a holder does not exit; it is necessary but not sufficient for positive feedback to form. Positive feedback also requires a dynamic signal loop: position-holding sustains the premium → the high rate is read by the market as confirmation of trend continuation

→ more same-direction capital is drawn in → the premium widens further. A complete formalization of this loop requires a multi-period model with information feedback and is left to future research. A final note, on the dimension of γ: it has the dimension of the reciprocal of the variance of the price change (i.e., if ΔP is measured in percent, the unit of γ is $\%^{-2}$), and its economic meaning is the additional risk compensation a trader demands per unit of variance risk exposure.

For a concrete numerical example, if a trader expects a token to rise 20% over the next three days, while the funding rate is 0.3% every 8 hours (about 0.9% cumulative per day), then the total rate cost over three days is about 2.7%, far below the 20% expected return. Under such a return expectation, even if the rate rises further to 1% every 8 hours, the trader still has an economic incentive to maintain the long position.

Consider the ALPACA token delisting event of April–May 2025. Binance announced that it would delist ALPACA spot and settle its perpetual futures on May 2 [27]. The market widely expected the delisting to cause a price crash, so large numbers of speculators poured into the perpetual futures to short, driving the funding rate quickly to a negative extreme (−1.24% every 8 hours, with some reports of a peak of −2.0%).

By negative-feedback logic, such a high shorting cost should have forced shorts to close, thereby relieving downward price pressure. In reality, however, the shorts expected the price drop upon spot delisting to far exceed the funding-rate cost of roughly 3.72% per day (−1.24% × 3). As long as the expected directional profit is large enough, speculators are willing to endure an extreme, sustained rate outlay. This "the math still works out" game led shorts not to close but instead to add further, ultimately producing a large-scale short squeeze: the ALPACA price did not crash before delisting but instead surged more than 1,100% (from about $0.024 to about $0.29), and large numbers of shorts were liquidated after paying extremely high funding rates [27].

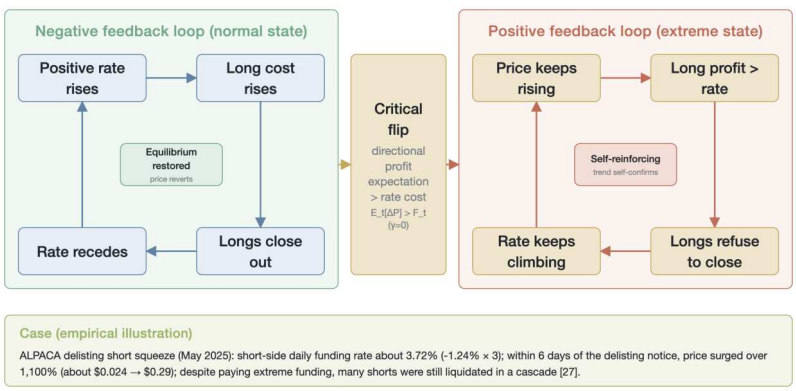


Figure 10-13. The flip of the funding rate mechanism from negative to positive feedback (a mechanism illustration based on a conceptual process, not empirical data; in the ALPACA event, −1.24%/8h is the peak negative rate, the daily total of 3.72% is extrapolated over three periods, and the price range of $0.024 to $0.29 is a conservative figure [27])

As Figure 10-13 shows, the flip proceeds in clear stages. In the normal state, a con-

tract premium induces a positive rate, the rate cost suppresses longs and attracts arbitrageurs, forming a negative feedback loop of price convergence. Once the expected directional profit exceeds the rate cost, holders no longer close positions under rate pressure; instead, the rising rate is read as confirmation of trend continuation, drawing in more same-direction capital and flipping negative feedback into positive feedback. In the ALPACA event, this flip completed within six days of the delisting announcement, accompanied by a price increase of more than 1,100% and chain forced liquidations of shorts [27].

10.5.2 A behavioral finance explanation

In the positive-feedback phase, the market exhibits a counterintuitive phenomenon: "the higher the funding rate, the more the paying side refuses to close." Beyond the pure mathematical profit calculation above, the behavioral-finance research of scholars such as Skwarek (2025) offers a deeper explanation for such irrational behavior in the crypto market [28].

The sunk cost fallacy and the disposition effect are the primary factors. After paying substantial funding rates and enduring paper losses, traders are often unwilling to accept the loss and exit. The disposition effect shows that investors tend to sell winning assets too early and hold losing assets too long. In a high-rate environment, closing a position means converting the paper loss and the rate already paid into a realized loss, and this psychological pain leads traders to keep holding. Confirmation bias reinforces the same behavioral pattern. In a polarized market, traders tend to selectively seek information supporting their position direction while ignoring the risk signal sent by a high funding rate. For example, at a bull-market peak, longs rationalize the extreme positive rate they pay with narratives of institutional entry or technical breakouts. Herding intensifies this dynamic further. The crypto market has pronounced social and community-driven characteristics. When the funding rate is extremely skewed to one side, it actually reflects the highly homogeneous expectations of market participants. Retail traders often believe the crowd's judgment is more accurate and thus establish positions in the crowd's direction, further intensifying rate polarization. At the theoretical level, two herding mechanisms must be distinguished: the rational information cascade described by Banerjee (1992) [29], in which each trader infers others' private information from their behavior and rationally imitates it, versus pure social conformity. The on-chain position transparency of the crypto market (such as Hyperliquid's public whale-position leaderboard) significantly accelerates the information-cascade effect by lowering the cost of observing behavior: traders can observe the direction and size of large positions almost in real time, which has no direct counterpart in the traditional-finance herding literature. Together, these factors point to gambling-like sensation-seeking behavior; Delfabbro and colleagues note that the behavior of some crypto traders is structurally similar to pathological gambling [30]. The extreme high-leverage, high-rate environment not only fails to deter but instead provides the high volatility and thrill they seek.

10.5.3 Termination paths of polarization

Once the funding rate falls into a polarized positive-feedback loop, it usually terminates in only one of two ways, whose effects on market structure differ starkly. Along the two dimensions of leverage concentration and liquidity depth, Figure 10-14 compares the two termination paths of polarized positive feedback: mild mean reversion and liquidation cascade.

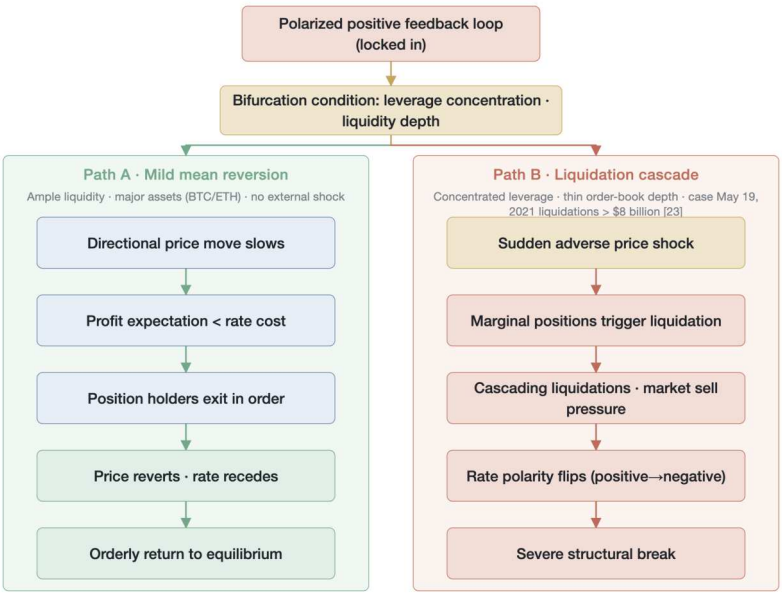


Figure 10-14. A comparison of the two termination paths of funding rate polarization positive feedback (a conceptual illustration based on synthetic trajectories, not empirical data)

In the case of mild mean reversion, the one-sided movement of the asset price begins to slow or stall, the expected directional profit falls accordingly, and the mathematical trigger condition reverses (expected profit $<$ rate cost). Holders realize that continuing to pay a high rate is no longer worthwhile and begin to close positions actively, in an orderly, staged manner. As one-sided positions decline, the contract price moves back toward the spot price and the funding rate gradually falls to normal levels. This path usually occurs in mainstream assets with good liquidity and ample arbitrage capital (such as BTC and ETH), absent a major external macro shock.

By contrast, the liquidation cascade path is far more destructive. When polarized one-sided positions (usually accompanied by high leverage) encounter a sudden adverse price move, some marginal positions are triggered into forced liquidation. In the case of positive-rate polarization (a crowded long side), long liquidations generate large numbers of market sell orders that instantly depress the contract price, in turn triggering long stop-losses and liquidations at lower price levels. This chain of forced long liquidations drives the contract price far below the spot price in a very short time, and the funding rate not only instantly returns to zero but may even flip polarity, from

an extreme positive rate to an extreme negative rate. The crash of May 19, 2021 is a canonical case: more than $8 billion in leveraged longs were liquidated within about 24 hours, and the funding rate fell from a sustained high positive value to an extreme negative value [23].

10.5.4 Rate polarization and the leverage cycle

Rate polarization is not an isolated event; it is the crypto derivatives market's counterpart to Hyman Minsky's "financial instability hypothesis" [31]. Minsky argues that prolonged stability breeds overconfidence, which in turn drives the accumulation of debt leverage and ultimately leads to collapse (the "Minsky moment").

The mapping is not exact, and its limits are worth stating precisely. The core causal chain of the Minsky framework is the gradual deterioration of the debt-to-income ratio and the hard constraint of repaying principal at maturity. Perpetual futures holders face no maturity-repayment constraint, and what triggers deleveraging is not "cash flow failing to cover interest" but the forced liquidation of margin. This changes the temporal profile of the crisis: the traditional Minsky cycle is gradual (usually unfolding over months to years of leverage accumulation), whereas a liquidation cascade in the crypto market can complete within minutes, so the transition from "boom" to "bust" is several orders of magnitude faster. The crypto market's leverage cycle is therefore more accurately described as an "accelerated Minsky," which retains the leverage-driven positive-feedback core while compressing the unfolding of the crisis from the quarterly scale of the macroeconomy to the minute scale of the microstructure.

In the crypto market, the evolutionary trajectory of the funding rate exhibits a close mapping to the leverage cycle. Figure 10-15 decomposes this mapping into four phases: leverage accumulation, criticality, deleveraging, and recovery.

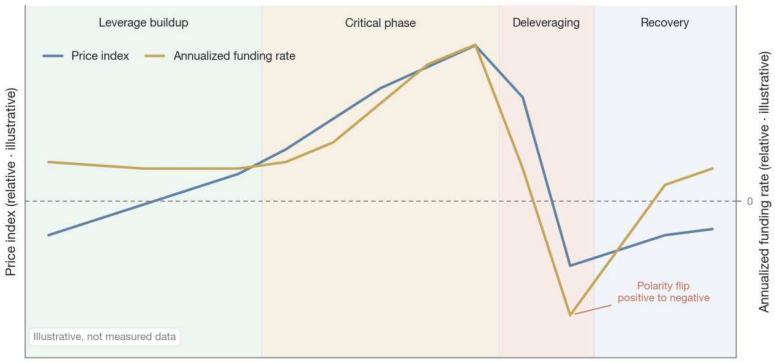


Figure 10-15. The mapping between the crypto-market leverage cycle and the evolution of the funding rate (a conceptual illustration, not empirical data; the horizontal axis shows qualitative cycle phases and the vertical axis shows relative levels, not measured scales)

The leverage accumulation phase (hedge finance) marks the start of the cycle: the market rises steadily, the funding rate stays at a mild positive value slightly above 0.01%,

arbitrageurs participate actively, and the market appears to be in a stable equilibrium. As the market advances, it enters the criticality or euphoria phase (speculative and Ponzi finance). Prices accelerate upward, speculators use paper profits to add positions and become desensitized to the rate cost. The funding rate soars and stays elevated, arbitrageurs' capital is exhausted, and negative feedback flips into positive feedback. At this point, an extreme funding rate is a clear signal that market leverage is over-concentrated and the system is highly fragile. When this fragility peaks, it triggers the deleveraging phase (the Minsky moment). A small external shock can set off a liquidation cascade, causing a price collapse. The funding rate flips polarity, from positive to negative, deeply reflecting the market's panic and liquidity drought. After violent turmoil, the market finally enters the recovery phase. Leverage is thoroughly washed out, the market enters a low-volatility phase, and the funding rate slowly returns toward zero or its default benchmark, quietly awaiting the gestation of the next cycle.

A persistently polarized funding rate is therefore not only a cost indicator but also a state variable measuring the market's systemic risk and leverage fragility. In practice, quantitative traders and risk managers can incorporate the duration and magnitude of rate polarization into early-warning models: when the funding rate stays above its historical 90th percentile for more than 48 hours, the market is in a high-probability window for a deleveraging event. Section 10.8 tests this hypothesis systematically and empirically.

10.5.5 The reconfiguration of participant behavior

When the market enters an extreme rate environment, the original three-layer game ecosystem breaks down, and the behavioral patterns and risk-return profiles of each participant are reconfigured. Figure 10-16 compares the changes in profit impact and risk exposure of five participant types under an extreme rate environment, revealing the transition from an arbitrageur-dominated orderly equilibrium to a liquidator-dominated disorderly state.



Figure 10-16. The reconfiguration of profit impact and risk exposure for each type of participant under an extreme funding rate environment (an illustrative relative assessment, not empirical data; the relative impact is assigned by the author)

Speculators and retail traders come under pressure first in rate polarization, becoming the net payers of the high rate, facing the largest profit drawdown and extremely high liquidation-risk exposure; their irrational persistence is precisely the driver that sustains the extreme rate. At the same time, the position of rate arbitrageurs also changes subtly. Although they can in theory earn excess rate income, they are constrained by the three boundaries of capital, execution, and counterparty, so their actual participation declines sharply, and they instead must bear extremely high exchange credit risk and wick risk. In this environment, liquidators replace arbitrageurs as the dominant market force. They earn substantial liquidation-penalty income by taking over and liquidating bankrupt positions, at relatively low risk. Market makers' behavior further aggravates the turmoil: amid violent fluctuations, they typically widen the bid-ask spread or pull orders outright to protect their inventory, which further worsens liquidity and intensifies the extremes in price and rate. By contrast, informed traders act more strategically. They use rate polarization as a contrarian indicator, positioning in the opposite direction ahead of a violent liquidation cascade, or providing liquidity to buy the dip after a cascade occurs.

In sum, the flip of the funding rate from negative to positive feedback marks a state transition of the market microstructure from orderly arbitrage to disorderly deleveraging. Understanding this polarization process directly informs the design of resilient crypto-asset trading strategies and risk-management frameworks.

10.6 The spectrum of strategic behavior

The funding rate mechanism was originally designed to anchor perpetual futures to spot prices, but in actual operation this mechanism inevitably breeds a complex game around its rules. Market participants are no longer merely price-takers but players of the mechanism. As the derivatives market's microstructure has matured, trading strategies around the funding rate have grown from simple cash-and-carry arbitrage into a full spectrum of more complex behaviors. This section systematically reviews this strategy spectrum and, through data and empirical cases, analyzes the impact of different strategies on market equilibrium and the internal logic of their evolution toward systemic manipulation.

10.6.1 A three-tier classification

By their contribution to market pricing efficiency and their manner of exploiting trading rules, funding-rate-related trading strategies can be divided into three tiers, forming a continuous spectrum from compliant arbitrage to systemic manipulation. Figure 10-17 arranges these three tiers along the two dimensions of compliance and market impact.

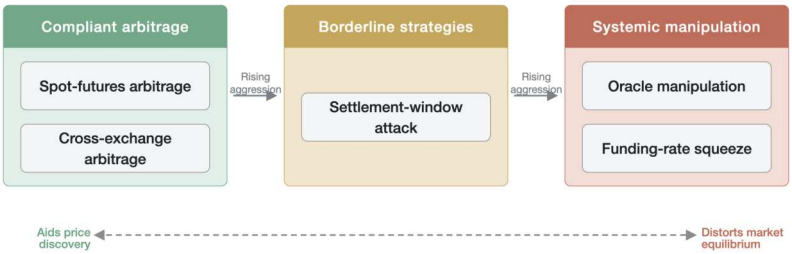


Figure 10-17. The spectrum of funding rate strategic behavior

The first tier is compliant arbitrage, mainly comprising traditional cash-and-carry arbitrage and cross-exchange rate arbitrage (the specific operating mechanisms of both are described in Section 10.4.2). These strategies earn rate income while bearing very low market risk, supply important liquidity, and perform the core function of eliminating price deviations and promoting price discovery, all of which supports the healthy functioning of the market.

The second tier comprises edge strategies. These strategies begin to operate in the gray area of the rules, their main feature being to profit by exploiting microstructural flaws in the mechanism design. A typical edge strategy is the "settlement window attack." Under a fixed-time settlement mechanism (such as traditional 8-hour settlement), a trader closes positions a few seconds before settlement to avoid the rate payment and re-establishes them immediately afterward. Although this behavior does not violate the trading rules, it causes severe liquidity fluctuations during the settlement window and reduces the effectiveness of the funding rate's anchoring mechanism.

The third tier is systemic manipulation. When edge strategies combine with a specific market structure (such as low liquidity and high concentration), they can evolve into systemic manipulation. Systemic manipulation no longer passively exploits rate differences but actively controls the spot or contract price to artificially create an extreme funding rate environment, forcing counterparties to be liquidated because they cannot bear the enormous funding cost or run out of margin. Typical systemic manipulation includes oracle manipulation and rate squeezes. In these strategies, the funding rate has been transformed from a tool for maintaining price anchoring into a tool for attacking counterparties. This chapter treats rate squeezes as game-theoretic cases for descriptive analysis, but such behavior may legally constitute market manipulation. In traditional securities markets, using an information advantage to control supply and engineer a forced-liquidation cascade in the derivatives market is a criminal offense in most jurisdictions. The U.S. Department of Justice filed criminal charges against Avraham Eisenberg in January 2023 (commodities fraud, commodities manipulation, and wire fraud), with the CFTC and SEC filing parallel civil enforcement actions at the same time; in April 2024 a jury found him guilty on three counts [32] (the criminal conviction was vacated in May 2025 by a federal judge on grounds including venue, and prosecutors have appealed). This shows that such behavior faces substantial enforcement risk in the digital-asset market, even though its ultimate legal characterization remains uncertain. In addition, an exchange's changing the settlement frequency dur-

ing the life of a contract (such as the shift from 8 hours to 1 hour in the ALPACA event) constitutes, under contract-law frameworks, a unilateral modification of a core contractual term. Although an exchange's terms of service usually reserve such modification rights, their legal validity and their legitimacy with respect to user interests remain contested across jurisdictions.

10.6.2 Equilibrium characteristics under normal markets

In a normal market where no systemic manipulation occurs, the funding rate exhibits specific statistical distribution characteristics. These characteristics reflect the equilibrium state of market participants within a compliant-arbitrage framework. Through a statistical analysis of 2,190 8-hour settlement rate data points for Binance's BTCUSDT perpetual futures from January 2024 to December 2025 ($365 \times 3 \times 2 = 2{,}190$, excluding the data points generated at higher frequencies during the 2025 dynamic-settlement-frequency period), we can clearly observe this equilibrium characteristic.

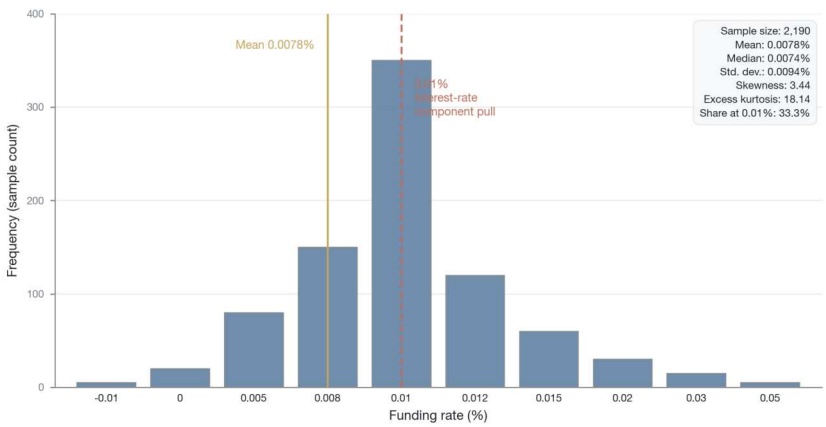


Figure 10-18. The distribution characteristics of BTCUSDT perpetual futures funding rates (computed by the author from the Binance BTCUSDT 2024–2025 funding rate history; descriptive statistics; the histogram is drawn in equal-width bins with stylized bar heights, and the excess kurtosis is highly sensitive to extreme observations)

As the figure above shows, the funding rate distribution under normal conditions has a pronounced leptokurtic (sharp-peak, fat-tail) character. The data's excess kurtosis reaches 18.14, far above the normal-distribution baseline of 0 (corresponding to a raw kurtosis of 21.14). This means the funding rate is highly concentrated at a specific level most of the time but occasionally exhibits extreme deviations. Note that the funding rate time series has a significant autocorrelation structure (volatility clustering), and the excess kurtosis of 18.14 may be partly attributable to this time dependence rather than to true fat tails under an i.i.d. assumption. After removing the autocorrelation structure (for example, via GARCH filtering), the excess kurtosis of the residuals may decline somewhat but still deviate significantly from normality.

Moreover, the distribution of the funding rate is markedly right-skewed (skewness of 3.44), with a mean of 0.0078%, slightly above zero. This structural positive deviation

reflects the long-standing long premium in the cryptocurrency market. Because market participants generally tend to use leverage to go long, the contract price is often above the spot price, and longs must pay shorts the funding rate.

The data show one more prominent feature: in fully 33.3% of all samples, the funding rate stays exactly at 0.01%. This phenomenon is not the result of natural market trading but a direct expression of the "interest rate component" in the mechanism design. In Binance's older formula, the base rate was fixed at 0.03% per day (i.e., 0.01% every 8 hours), and when the market premium was within the clamp band, the funding rate was dominated by this fixed rate. This "gravitational effect" shows that under normal conditions, the mechanism rules often determine the rate more than actual market supply and demand do.

10.6.3 The ALPACA funding rate squeeze case

When fragility appears in the market microstructure, the normal equilibrium is broken and the funding rate can become a tool for systemic manipulation. The ALPACA rate squeeze that occurred on Binance in April–May 2025 is a canonical case of this phenomenon. The event exhibited the contract paradox of "a negative catalyst triggering an extreme positive price reaction."

On April 24, 2025, Binance announced that it would delist the ALPACA spot trading pair on May 2 and settle the perpetual futures on April 30. In traditional market logic, a delisting announcement is a strong bearish signal. This news quickly formed a consensus expectation among retail traders that "the price will inevitably go to zero," causing large amounts of short capital to pour into ALPACA perpetual futures and forming an extremely crowded short camp.

The manipulator, however, recognized this structural weakness. At the time, ALPACA's circulating market capitalization had shrunk to about $5 million, and the token was highly concentrated. The manipulator used very little capital to buy continuously in the spot market, driving up the spot price, while establishing long positions in the contract market. Because of the large presence of shorts, the contract price was severely depressed, causing the funding rate to take an extreme negative value (shorts paying longs).

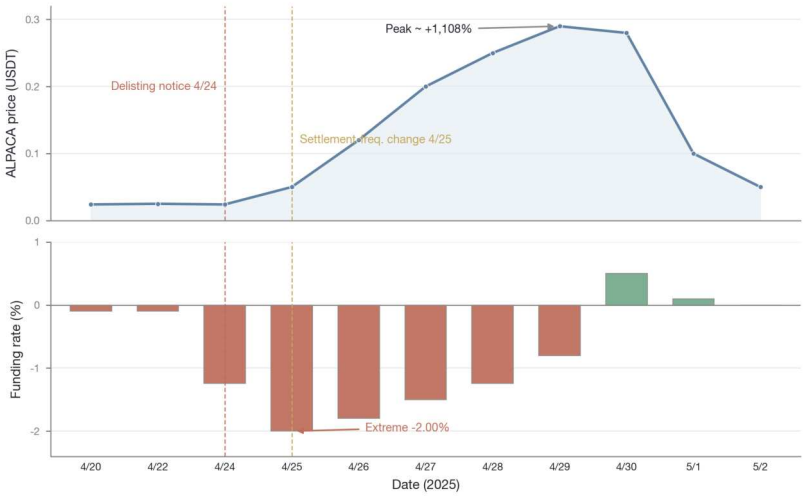


Figure 10-19. Price and funding rate movements during the ALPACA funding rate squeeze (illustrative movements, event reconstruction); the price range of \$0.024 to \$0.29 is a conservative figure, the rate cap moved from ±2% to ±4%, and the total liquidations of about \$50 million (of which about \$43 million were shorts) have been verified value by value [27]

As the figure above shows, after the delisting announcement, the ALPACA price did not crash but instead surged violently, with a peak gain of more than 1,100%. At the same time, the funding rate hit its floor deeply. To respond to the extreme market conditions, Binance activated the dynamic settlement-frequency mechanism on April 25, gradually shortening ALPACA's settlement frequency from every 8 hours to every 1 hour and adjusting the rate cap to ±2% (and subsequently widening it to ±4%).

This mechanism adjustment was intended to raise the cost of manipulation, but objectively it sharply accelerated the depletion of short margin. Under 1-hour settlement and a ±2% rate cap, a short position had to pay as much as 2% in funding every hour, amounting to a daily cost as high as 48% (this calculation applies to the old algorithm before the September 2025 normalization update. The ALPACA event occurred in April–May 2025, when Binance had not yet introduced the frequency normalization factor; note also that the rate is calculated on the notional position value, and because the ALPACA price rose more than 1,100% during this period, the absolute amount of each later period's rate was far larger than that of the early periods, a nonlinear amplification effect). At this rate level, the funding rate is no longer a marginal trading cost but a cash-flow burden the holder cannot bear. Facing enormous funding costs and a continuously rising price, shorts were forced to close (buy), and this chain of forced short liquidations pushed the price higher still, ultimately producing total liquidations of about \$50 million (of which about \$43 million were shorts). In this case, the actual effect of the funding rate mechanism ran counter to its corrective intent, becoming a tool for the manipulator to amplify counterparty losses.

From the ALPACA case, we can distill a more general model of rate-manipulation feasibility. The manipulator's expected profit can be expressed as:

$$\Pi = V_{\text{squeeze}} - C_{\text{spot}} - C_{\text{funding}}$$

where V_{squeeze} is the gain from squeezing the counterparty (including the price-impact profit from the counterparty's forced closing and the rate income), C_{spot} is the capital cost of establishing and maintaining control in the spot market, and C_{funding} is the funding-rate outlay the manipulator's own side bears over the manipulation period. This simplified model omits nonlinear execution costs: on an extremely thin order book, the manipulator's position-building self-inflates the price, and the closing slippage of the squeezed shorts amplifies actual losses; a more complete model should add a market-impact function (such as the Almgren-Chriss framework) to capture these nonlinear frictions. Even so, the feasibility of manipulation still depends on three key parameters: the circulating market capitalization M (the lower M, the smaller C_{spot}), the short concentration ρ (the higher ρ, the larger V_{squeeze}, because crowded shorts are more easily liquidated in a chain), and the settlement frequency f (under the un-normalized old algorithm, the higher f, the faster the counterparty's cumulative rate burden $C_{\text{funding}}^{\text{opponent}}$ grows). ALPACA happened to satisfy all three conditions simultaneously—low M (circulating market capitalization of about \$5 million), high ρ (delisting expectations made shorts extremely crowded), and high f (Binance raised the settlement frequency from 8 hours to 1 hour)—making the expected profit of manipulation far exceed its cost. This framework indicates that in assessing a token's manipulation risk, regulators and exchanges should systematically monitor the combined state of these three parameters.

The ALPACA event and similar high-frequency-settlement squeezes exposed a systemic vulnerability in early funding rate mechanisms under extreme conditions. In particular, when the settlement frequency is dynamically raised, if the calculation of the per-settlement rate remains unchanged, the daily cumulative rate cost is amplified many times over, giving manipulators an opening.

To patch this vulnerability, Binance made a major update to the funding rate formula on September 18, 2025. The core of this update was the introduction of a "frequency normalization factor."

In the old algorithm before the update, the per-settlement funding rate formula was:

$$F = \bar{P} + \text{clamp}(I - \bar{P},\ -0.05\%,\ +0.05\%)$$

where $\bar{P}$ is the average premium index over the settlement period. Under the old algorithm, when the settlement frequency was shortened from 8 hours to 1 hour, although each settlement covered a shorter time, the part of the formula determined by the premium index did not scale down proportionally. This means that if the market has a persistent extreme premium, the per-settlement rate under 1-hour settlement can be of the same order of magnitude as the per-settlement rate under 8-hour settlement, amplifying the daily cumulative rate eightfold.

The updated new algorithm introduces the frequency normalization factor, where N is the number of hours in the settlement period:

$$F = \frac{N}{8} \cdot \left[\bar{P} + \text{clamp}(I - \bar{P},\ -0.05\%,\ +0.05\%)\right]$$

The logic of this formula is to first compute a nominal rate using the standard 8-hour formula, then multiply by $N/8$ to scale. Note that the clamp bounds $(-0.05\%, +0.05\%)$

do not scale with the settlement frequency in the formula; the scaling applies only to the overall result. For a concrete numerical example: assume the average premium index $\bar{P} = 0.5\%$ and the interest rate component $I = 0.01\%$. When $N = 8$ (the old algorithm), $F = 1 \times [0.5\% + (-0.05\%)] = 0.45\%$, and the daily cumulative rate is $0.45\% \times 3 = 1.35\%$; when $N = 1$ (the new algorithm with 1-hour settlement), $F = \frac{1}{8} \times 0.45\% = 0.05625\%$, and the daily cumulative rate is $0.05625\% \times 24 = 1.35\%$—the two daily cumulative rates are exactly identical.

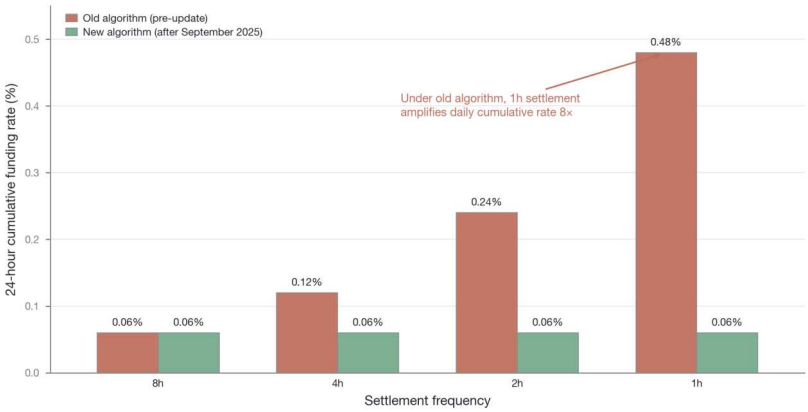


Figure 10-20. A comparison of daily cumulative rates before and after Binance's funding rate algorithm update (a formula-derived illustration with a real mechanism; the example values are constructed and are not the 0.45%/1.35% worked example in the main text [10])

As Figure 10-20 shows, by introducing the frequency normalization factor, the new algorithm ensures that at the same market premium level, regardless of how the settlement frequency changes (8 hours, 4 hours, 2 hours, or 1 hour), the theoretical upper bound of the cumulative funding rate over 24 hours remains consistent. This mechanism evolution effectively seals off the attack path by which manipulators used high-frequency settlement to amplify counterparty costs, and it is a concrete improvement of the funding rate mechanism toward manipulation resistance. Binance also shortened the calculation baseline of the mark price from 1 minute to 30 seconds, improving the system's responsiveness to abnormal price fluctuations and further compressing the space for manipulation.

10.7 The settlement frequency game

In the early design of perpetual futures, the settlement frequency of the funding rate was usually static and fixed (such as every 8 hours). This design had advantages in reducing the system's computational load and providing predictability. As market volatility increased and strategies grew more complex, however, static settlement frequency gradually revealed its fragility under extreme market conditions. Beginning in 2025, major exchanges successively introduced dynamic settlement-frequency mechanisms, a shift that is less a technical upgrade than a change in mechanism-design philosophy, from passive reaction to active management.

10.7.1 From passive reaction to active management

The traditional 8-hour static settlement mechanism is essentially a "passive reaction" system. It assumes that the market has enough time (8 hours) to digest information and restore equilibrium, with the funding rate serving merely as a mild corrective tool. In the crypto market, however, asset prices can fluctuate violently within minutes, and an 8-hour settlement period means the system is in a "blind spot" most of the time, providing ample room for edge strategies such as the settlement window attack.

The introduction of the dynamic settlement-frequency mechanism marks the beginning of exchanges adopting an "active management" posture. Take the dynamic mechanism that Binance fully implemented in May 2025: the system automatically switches the settlement frequency among 8 hours, 4 hours, and even 1 hour according to the real-time state of the market premium. Binance's dynamic settlement-frequency mechanism is not fully algorithmic, however: the decision to switch frequency involves a manual or semi-automated approval step, and the trigger condition depends not only on the rate itself but also on a multidimensional assessment of the absolute value of the premium index, its duration, and the overall market risk situation. The mechanism's transition "from passive to active" therefore spans two levels—algorithmic automation and human risk-control judgment—and the latter has inherent limitations in transparency and predictability.

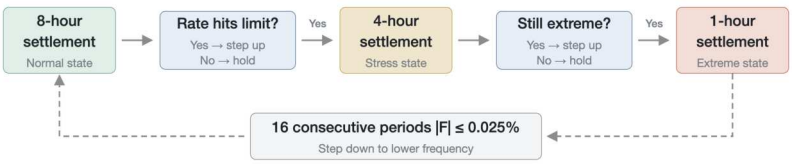


Figure 10-21. The state-transition logic of the dynamic settlement frequency mechanism (a mechanism illustration; the recovery threshold of 16 consecutive periods with $|\text{rate}| \leq 0.025\%$ is the author's illustrative rendering of the mechanism's logic, not a published specification)

As Figure 10-21 shows, the core logic of the mechanism is that when the funding rate repeatedly hits the set upper and lower bounds, the system automatically raises the settlement frequency (for example, from 8 hours to 4 hours, then to 1 hour). Only when the market returns to calm (for example, when the absolute value of the funding rate is less than or equal to 0.025% for 16 consecutive periods) does the system gradually restore a lower settlement frequency. This design turns the settlement frequency itself into an endogenous game variable, enabling the mechanism to adaptively adjust the strength of correction according to market stress.

10.7.2 The game-theoretic implications of different settlement frequencies

Different settlement frequencies have starkly different game-theoretic implications for participant behavior and market microstructure. Table 10-7 compares in detail the

market characteristics under three main settlement-frequency modes:

Dimension	8-hour static settlement	1-hour high-frequency settlement	Continuous settlement
Arbitrageur behavior	Concentrated around settlement (settlement window attack), causing periodic liquidity fluctuations	Trading is distributed more evenly, arbitrage opportunities are fleeting, and high automated-trading capability is required	Arbitrage behavior is fully smoothed, eliminating the traditional settlement window attack, but the EMA reference window it relies on (such as dYdX v4's 8-hour EMA) itself creates a new microstructural manipulation dimension, as an instantaneous price shock can affect the rate throughout the entire EMA window
Funding cost accumulation	Step-wise jumps; each payment is large in amount, but there is ample time to close and avoid it	Rapid accumulation; continuous cash-flow pressure; high holding cost	Real-time interest accrual; continuous deduction; extremely high margin requirements
Suppression of one-sided markets	Weaker; within 8 hours the price may already have deviated irreversibly	Stronger; the high holding cost forces the disadvantaged side to surrender quickly	Extremely strong; a price deviation is immediately converted into a cash-flow penalty
System-performance requirements	Low; suitable for large-scale concurrent computation	Higher; requires more frequent snapshots and liquidations	Extremely high; usually implemented only on-chain in perpetual futures protocols or specific architectures

Table 10-7. Comparison of the game-theoretic implications of three settlement-frequency modes (Data source: compiled by the author)

Under the old algorithm system, because of the lack of frequency normalization, 1-hour settlement not only increased the frequency of payments but also amplified the cumulative funding cost many times over (with a premium index of 0.5%, under the old algorithm 1-hour settlement amplified the daily rate about eightfold; the mechanism and worked example are given in Section 10.6.3). Even after the introduction of the normalization factor, high-frequency settlement still significantly changes the game equilibrium between longs and shorts by increasing friction costs and reducing decision time. The compression of decision time forces trend traders to set tighter stop-losses, because the faster depletion of margin means that a countertrend position that could originally be held for several days may hit the liquidation line within hours under 1-hour settlement.

10.7.3 The PIPPIN case analysis

The PIPPIN perpetual futures event of December 2025 is a canonical case of how a high-frequency-settlement mechanism reshapes the long-short game under extreme conditions. As an emerging AI-agent-themed token, PIPPIN attracted large amounts of short capital trying to "call the top" as its price climbed rapidly and repeatedly set new highs (a closing high of about $0.54 within December). Because shorts were overcrowded, the perpetual futures price fell far below the spot price, triggering Binance's dynamic settlement mechanism and adjusting the settlement frequency to once every hour.

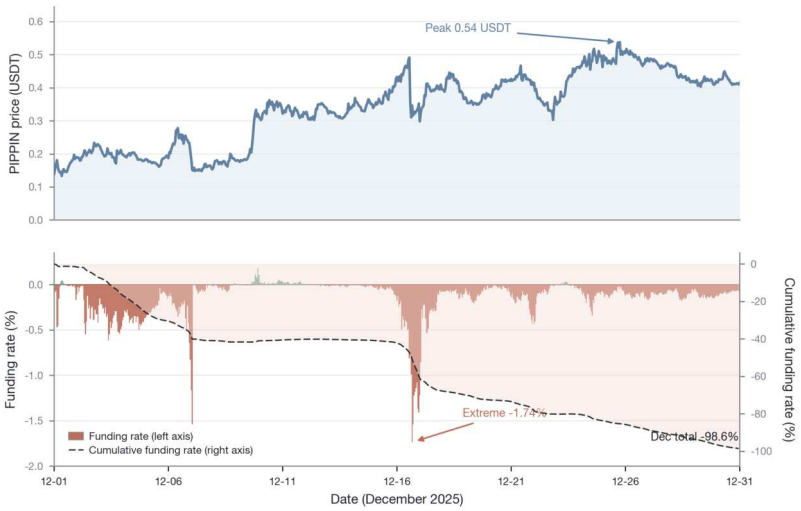


Figure 10-22. Price and funding rate dynamics for PIPPIN under extreme conditions (Data source: Binance PIPPINUSDT perpetual, 1-hour settlement, hourly observations for December 2025; as of June 30, 2026; the extreme single-period rate within the month was −1.74%, and the cumulative funding rate for the full month was about −98.6%)

As Figure 10-22 shows, during the extreme conditions of mid-December, PIPPIN's funding rate stayed for a long time in a deeply negative range. In particular, from December 16 to 17, there were dozens of hours of continuous high negative rates. According to live data, in the most extreme 24 hours (18:00 on December 16 to 17:00 on December 17), a short position's cumulative funding-rate outlay was as high as −19.77%. Over the whole of December, PIPPIN experienced 26 consecutive high-negative-rate cycles, the longest lasting 80 hours, during which the cumulative funding rate reached −27.81%.

Such high-frequency, high-magnitude funding rates dealt a devastating blow to shorts. For a trader holding a short with a notional value of about 3.53 million USDT, one day's funding-rate outlay alone could approach 700,000 USDT.

From the operational-risk dimension, the PIPPIN event also exposed two problems that are easily overlooked but critically important for live trading in a high-frequency-settlement environment. The first is **liquidation-engine pressure**: when large numbers of shorts trigger forced liquidation simultaneously, the liquidation engine must

execute large numbers of buy market orders in a very short time, placing extreme pressure on order-book depth and matching-system throughput and potentially causing shortfall losses and accelerated insurance-fund depletion. The second is **the margin-management dilemma**: under an hourly deduction mechanism, traders face multiple execution dilemmas including congested deposit channels, delays in cross-exchange fund transfers, and the automatic reduction of notional position size through partial liquidations. In PIPPIN's extreme rate environment, these operational-level frictions may be the decisive factor in a trader's final P&L, rather than merely a minor assumption in a theoretical model.

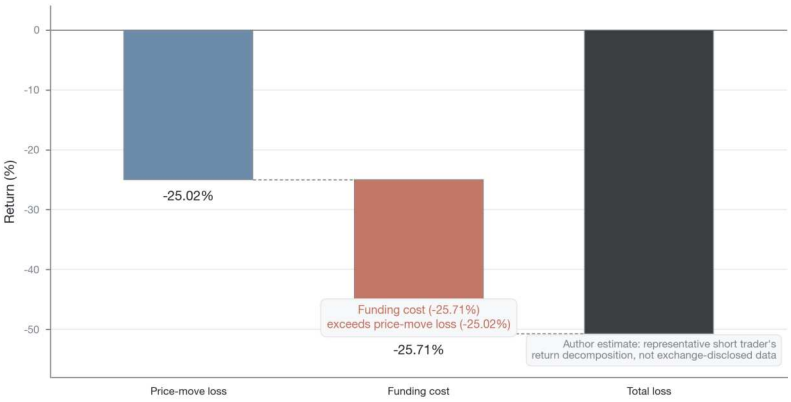


Figure 10-23. A decomposition of the return of a PIPPIN short trader (a representative/illustrative decomposition, estimated by the author, not measured from a specific account)

Through a decomposition of the return of a typical short trader (Figure 10-23), a counterintuitive pattern emerges: of the total loss of −50.73%, the loss from price movement was −25.02%, while the loss from funding-rate outlays reached −25.71%. The cost of the funding rate had exceeded the loss caused by price movement itself. This decomposition demonstrates forcefully that under a high-frequency-settlement mechanism, the funding rate is no longer a secondary friction cost but the core variable determining a trade's P&L outcome.

10.7.4 Microstructure effects of dynamic settlement

The introduction of dynamic settlement frequency has had a profound systemic impact on the microstructure of the derivatives market, and different types of participants have been forced to adjust their behavioral patterns:

Market makers and liquidity providers are directly affected: high-frequency settlement requires market makers to have stronger inventory-management capability and faster hedging speed. On on-chain derivatives DEXs, because there is no centralized risk control, market makers often must bear large unrealized profits and extremely high funding-rate outlays. This forces them to complete their pump and distribution within a very short time, or else the enormous interest outlays will rapidly erode their operating capital. The strategies of trend traders also face severe challenges. For a trend

trader holding a one-sided position, the dynamic mechanism significantly increases the time cost of holding. When holding a countertrend position, high-frequency settlement accelerates the depletion of margin, so that the risk of the traditional hold-and-endure strategy rises significantly, forcing traders to stop out earlier, which may in turn intensify one-sided price movement in the short run. The survival space of arbitrageurs is likewise reconfigured. Dynamic settlement compresses the space for the settlement window attack but also creates new arbitrage opportunities based on expectations of frequency switching. Arbitrageurs must build more complex models, forecasting not only the price's return but also the state transition of the settlement frequency.

10.7.5 The costs of high-frequency settlement

Although the dynamic high-frequency-settlement mechanism demonstrates strong corrective power when addressing extreme market deviations, it is not without cost. From a game-theoretic perspective, the widespread application of this mechanism prompts deep reflection on the boundaries of market design.

The first problem is intensified liquidity friction. Although high-frequency settlement smooths the payment of funding, the frequent cash-flow transfers themselves constitute a systemic friction. When market liquidity is insufficient, a high funding rate may cause liquidity providers to exit because they cannot bear the interest cost, which in turn widens the bid-ask spread and worsens market depth.

The risk of mechanism perversion also arises. As the ALPACA and PIPPIN cases show, when the market has a structural flaw (such as low circulating market capitalization or highly concentrated holdings), the funding rate mechanism—designed to penalize deviation—may instead be exploited by manipulators to become a tool for engineering short or long squeezes. High-frequency settlement objectively shortens the disadvantaged side's reaction time and accelerates the onset of liquidation.

In sum, the settlement frequency of the funding rate is no longer a purely technical parameter but a game lever for regulating participant behavior. Future mechanism design must find a finer balance between "timely correction" and "preventing abuse of the mechanism." For example, the frequency normalization factor introduced by Binance is a successful attempt that limits the unbounded amplification of the single-day cost while retaining the deterrent power of high-frequency settlement, marking a step toward a more mature and resilient derivatives market.

From a broader mechanism-design perspective, the linear rate function currently used by most major exchanges ($F = P + \text{clamp}(I - P, a, b)$) is not the only design choice, and its alternatives merit systematic exploration. A quadratic rate function ($F \propto P^2 \cdot \text{sgn}(P)$) can impose a mild penalty for small deviations and a sharply growing penalty for large deviations, thereby strengthening corrective force under extreme states without adding friction in normal markets. A dynamic-clamp-bounds scheme allows the width of the clamp band to adjust adaptively with market volatility or open interest: narrowing the stable band in a low-volatility environment to increase the rate's sensitivity to the premium, and widening it in a high-volatility environment to prevent the rate from over-amplifying liquidation pressure. In addition, Drift Protocol (Solana) has deployed a nonlinear rate function based on the degree of long-short imbalance in open interest

(rather than the price premium). This design directly targets the asymmetry of long-short open interest: even if a manipulator widens the premium by pumping the spot price, the rate does not move substantially as long as the ratio of long-short open interest does not change significantly, which in theory offers better resistance to price manipulation.

In theory, continuous settlement is the most thorough way to eliminate the time-arbitrage space of discrete settlement windows, and it brings the rate closer to a real-time feedback signal. Yet the EMA reference window it relies on introduces a new manipulation dimension. Continuous settlement faces severe engineering constraints, however. The current major on-chain perpetual futures protocols (Hyperliquid, dYdX v4) no longer run on Ethereum mainnet. Hyperliquid runs on its own L1 chain (subsecond block time, zero gas fees), and dYdX v4 is a standalone chain based on the Cosmos SDK. In these app-chain environments, the engineering constraints on continuous settlement come mainly from state consistency (all validators must reach consensus on the rate calculation at the same instant) and sequencer throughput, rather than from gas fees in the traditional sense; centralized exchanges have no such limitation, but continuous settlement requires real-time margin snapshots and equity calculations, placing far higher demands on back-end infrastructure than discrete settlement. Whether these alternatives can improve the anchoring performance and manipulation resistance of the rate mechanism while maintaining computational efficiency constitutes an open research frontier in derivatives mechanism design.

10.8 Funding rate as a state variable

In traditional microstructure analysis, the funding rate is usually treated as an output of the market state: a passive result of the imbalance between long and short forces, a friction cost the anchoring mechanism produces to correct a price deviation. When we move from the microstructural view of mechanism design to the macro view of market dynamics, however, the nature of the funding rate changes fundamentally: it is no longer merely a passive result but an information-rich state variable and a forward-looking predictor. This section advances the chapter's fifth core theoretical contribution: the funding rate as state variable hypothesis. The hypothesis holds that, because of the absolute dominance of perpetual futures in the crypto derivatives market, the funding rate not only records historical trading traces but also encodes participants' directional preferences about the future, the fragility of leverage accumulation, and the distribution of systemic risk. This shift in thinking is analogous to the paradigm evolution in traditional finance from viewing the interest rate as a passive output of monetary policy to viewing the yield curve as an active predictor of the economic outlook. The sections below deconstruct the information content of the funding rate along three dimensions—rate level, rate volatility, and the cross-asset rate surface—and draw out its fundamental differences from traditional sentiment indicators.

10.8.1 The information content of the rate level

The absolute level of the funding rate is the most intuitive signal in the market. At the first order, it conveys directional information: a positive rate means the market is net

long, and a negative rate means the market is net short. The deeper value of the rate level, however, lies in its second-order information—that is, it quantifies the intensity of leverage accumulation.

Applying the three-layer decomposition model of Section 10.2, after stripping away the financing cost layer (the institutional benchmark, such as BitMEX's 0.01% every 8 hours, about 10.95% annualized) and the institutional friction layer (the truncation effect of the clamp function), the absolute value of the remaining sentiment premium layer directly measures the "overheating" of market leverage. The higher the sentiment premium layer, the deeper the leverage accumulation, and the closer the "leverage cycle" discussed in Section 10.5 is to its critical point.

In a fully efficient market, the presence of arbitrageurs should quickly return the funding rate to a neutral level. When the funding rate deviates from the neutral level for an extended period, however, it reveals the depletion of arbitrage capital or the overwhelming advantage of directional speculative force. Such a persistent deviation usually means the market is entering a fragile one-sided state. When the rate stays elevated for a long time (an extreme positive rate), longs are not only bearing a large holding cost, but the accumulation of this cost continuously erodes longs' margin buffer. This state is analogous to "potential-energy accumulation" in physics: a high rate does not directly cause a price crash, but it creates an extremely fragile system state in which any small negative shock can trigger a chain of long deleveraging. Conversely, when the rate is at an extreme negative value, shorts are paying a high price to maintain their positions, and the market is likewise extremely taut, liable to a short squeeze at any moment.

A high positive rate does not equal "market consensus bullishness"; it means only that leveraged longs are overcrowded. When crowding reaches its limit and begins to collapse, relying on the rate signal to make directional decisions faces serious risk. November 2021 is a canonical case: the funding rate of BTC perpetual futures stayed above 0.1% every 8 hours around the time the price hit its $69,000 all-time high, appearing on the surface to send a strong bullish signal; in reality, this signal reflected the over-crowding of leveraged longs rather than fundamental support. Over the subsequent two months, BTC fell more than 40%, and the long positions established during the high-positive-rate period suffered the double loss of directional loss and rate outlay. The rate mainly reflects leverage sentiment rather than a market view, and this distinction is fundamental to correctly interpreting the rate signal.

This chapter validates this mechanism with empirical data. Based on measured 8-hour funding rate data for Bitcoin and Ethereum perpetual futures on Binance from 2024 to March 2026, we divide the absolute rate value into three ranges (low rate, medium rate, high rate) and observe their power to predict rate volatility over the next 7 days.

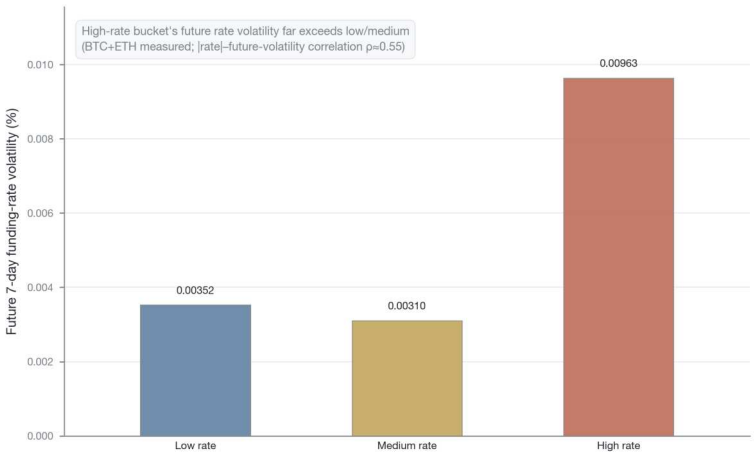


Figure 10-24. The relationship between the current absolute rate level and the funding rate volatility over the next 7 days (Data source: measured 8-hour funding rates for BTC and ETH perpetuals on Binance, 2024 to March 2026; as of June 30, 2026; the future 7-day rate volatility is averaged within tertile buckets, and the correlation between |rate| and future volatility is about 0.55)

The data show that the degree to which the current rate level is extreme is positively correlated with future volatility. When the market is in the "high rate" range (whether positive or negative), the rate volatility over the next 7 days is higher than in the "low rate" and "medium rate" ranges. This confirms that a high-rate state is essentially a high-potential-energy, high-fragility market structure: the rate level is not only a static snapshot of long-short forces but also a forward-looking indicator of future market turmoil.

10.8.2 The information content of rate volatility

If the rate level measures the degree of leverage accumulation, then rate volatility directly measures the uncertainty at the market-microstructure level and the degree of disagreement in consensus. Rate volatility is an information dimension independent of price volatility; it captures the frequency and magnitude of the swings in long-short forces within the derivatives market.

In a calm market environment, the funding rate usually fluctuates within a narrow band (dominated mainly by the benchmark rate). When the market faces a major information shock, liquidity drought, or an impeded arbitrage mechanism, however, the funding rate oscillates violently. This oscillation reflects the chaotic state of market participants repricing risk, as well as the higher risk compensation that liquidity providers (arbitrageurs) demand in the face of high volatility. Specifically:

When rate volatility is low, market consensus on direction tends to be stable. This may mean a trend has been established, or it may mean the market has fallen into complacency—the calm-before-the-storm trap. Conversely, when rate volatility is high, directional preference is switching frequently, meaning the market is full of high uncertainty. This is often a leading indicator of a market state transition (such as from

trend to range, or from range to reversal). Further, a sharp rise in rate volatility is a clear signal that the market is entering a new state, analogous to a jump in VIX in traditional markets, but it more directly reflects the sudden change in uncertainty within the leveraged market.

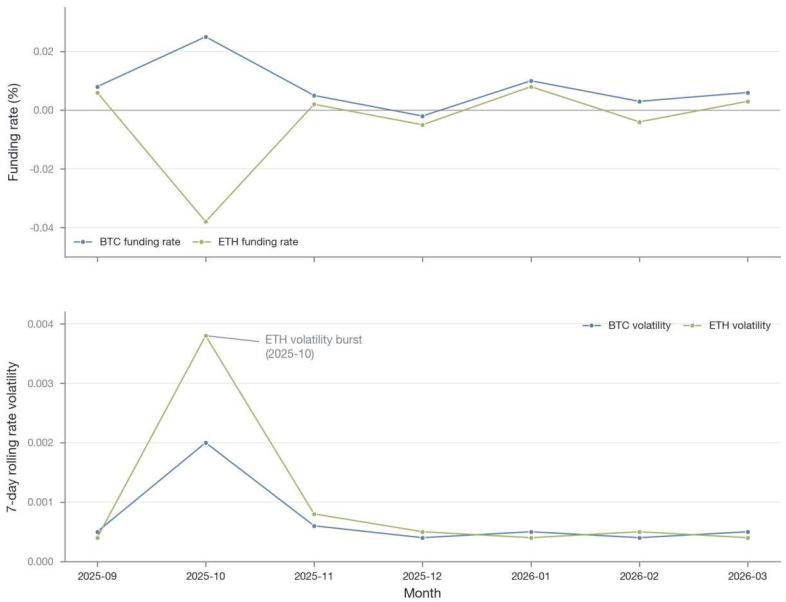


Figure 10-25. Time series of BTC and ETH funding rates and 7-day rolling rate volatility (computed by the author from the Hyperliquid API; descriptive; sample period September 2025 to March 2026, single exchange) [33]

As Figure 10-25 shows, the rate volatility of BTC and ETH is not uniformly distributed but exhibits pronounced episodic clustering. In particular market phases, rate volatility suddenly spikes, usually accompanied by violent spot-price movements or major macro events. ETH's rate volatility is significantly higher than BTC's in several periods, reflecting the structural feature of relatively shallower arbitrage depth and relatively stronger speculative force in the ETH derivatives market.

At a deeper level, a rise in rate volatility often signals a weakening of the "anchoring strength" discussed in Section 10.3. When arbitrage capital retreats because of risk aversion or capital constraints, even a small change in speculative force can cause large ripples in the rate. Rate volatility can therefore be viewed as an inverse indicator of the effectiveness of the arbitrage mechanism: the higher the volatility, the more fragile the arbitrage mechanism and the more vulnerable the market is to one-sided force.

To further quantify the impact of an extreme rate state on subsequent market behavior, we analyze the change in market activity after extreme-rate days.

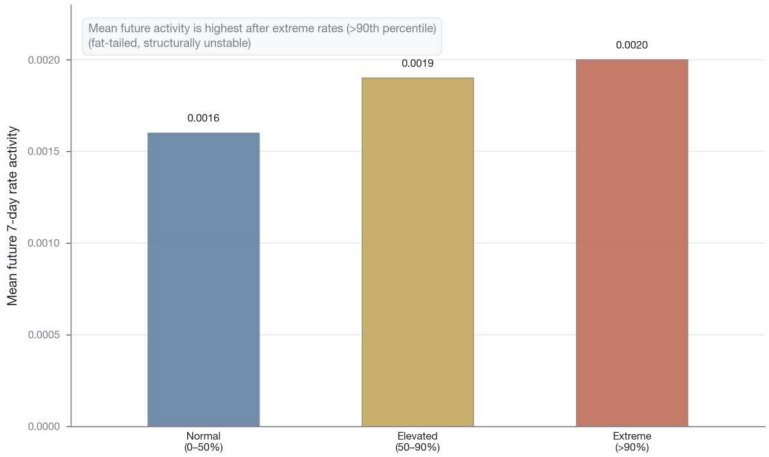


Figure 10-26. The distribution of funding rate activity over the next 7 days following an extreme rate state (computed by the author from the Hyperliquid API; descriptive; sample period September 2025 to March 2026, single exchange) [33]

Figure 10-26 clearly shows that when the market is in an "extreme rate" state (above the 90th percentile of the historical absolute-value distribution), the rate activity over the subsequent 7 days (i.e., the sum of absolute rate changes) exhibits a pronounced fat-tailed distribution, with a mean significantly higher than in the normal and elevated rate states. This means that an extreme rate not only foreshadows future high volatility but also foreshadows a structural instability. Once the rate enters the extreme range, the market rarely calms down quickly; instead, it undergoes a sustained period of turmoil until leverage is thoroughly cleared and a new equilibrium is established.

10.8.3 The cross-asset funding rate surface

Extending the view from a single asset to the entire market, the funding rate constructs a multidimensional cross-asset surface. Strictly speaking, perpetual futures have no "term structure" in the traditional sense (because there is no expiry date), but we can construct an analogous information surface along two dimensions: the cross-asset dimension and the cross-exchange dimension. This surface is one of the high-frequency indicators for characterizing the cross-asset distribution of speculative heat.

Along the cross-asset dimension, the rates of BTC, ETH, SOL, and various altcoins reflect the intensity of directional preference for different assets. Altcoin rates are usually far higher than BTC's, reflecting stronger speculative preference and lower arbitrage efficiency. Along the cross-exchange dimension, the rate differences for the same asset across platforms such as Binance, Bybit, Hyperliquid, and dYdX reflect differences in each platform's user composition and liquidity structure. Boros Finance's data show that the rate difference between Binance and Hyperliquid can reach a significant level, providing considerable return space for cross-exchange arbitrage (the specific return ranges are given in Section 10.4.2 and Figure 10-10) [22].

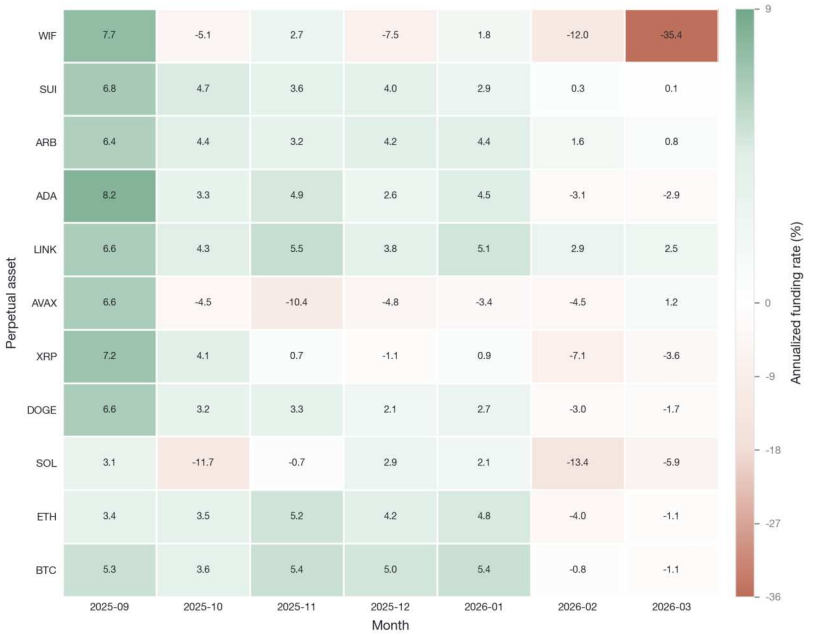


Figure 10-27. A monthly heatmap of cross-asset funding rates (all values are measured funding rates for Binance USDT perpetuals, annualized from each asset's mean per-period rate for the month; sample period September 2025 to March 2026, as of June 30, 2026)

The cross-asset heatmap in Figure 10-27 reveals several patterns. The cornerstone assets exhibit high stability: the rates of BTC and ETH are relatively stable, staying in a mild positive range in most months (about 3%–5% annualized) and turning slightly negative only in early 2026, with a volatility clearly smaller than that of altcoins, reflecting their status as market benchmarks and the ample arbitrage liquidity in their derivatives markets. Long-tail assets fluctuate more widely: relative to BTC, long-tail assets (such as WIF, SOL, and OP) exhibit extreme rate fluctuations. For example, according to measured Binance data, WIF's annualized rate plunged to about −35.4% in March 2026 and SOL's fell to about −13.4% in February 2026, indicating that these assets were undergoing intense shorting pressure or a liquidity crisis. Sector rotation is the final dimension: by watching which assets carry high rates from month to month, one can track the movement of speculative capital. In September 2025, most altcoin rates rose collectively (ADA and WIF reached about 8.2% and 7.7%, respectively), but by early 2026 most turned negative (for example, AVAX fell from about 6.6% in September to about −10.4% in November), while LINK maintained a stable positive rate throughout the observation period. When the assets of a sector collectively exhibit high positive rates, this is often a clear signal that the sector is overheated.

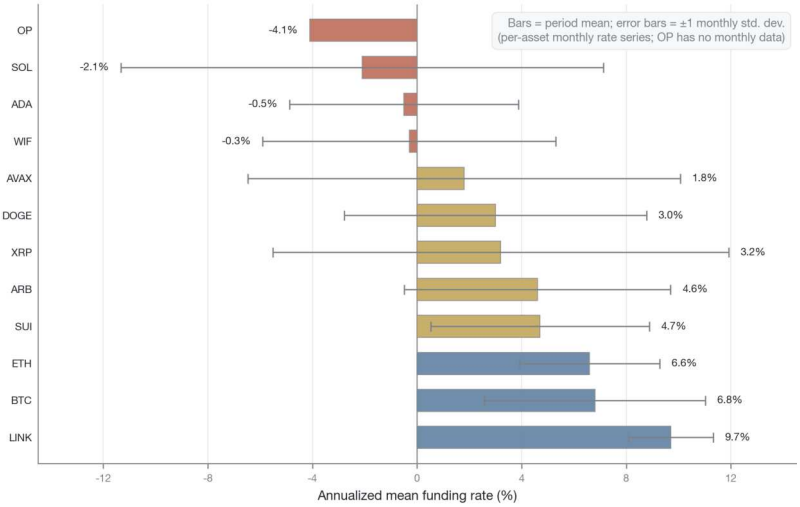


Figure 10-28. A comparison of cross-asset annualized average funding rates with standard-deviation ranges (computed by the author from the Hyperliquid API; descriptive; sample period September 2025 to March 2026, single exchange) [33]

Figure 10-28 further shows the overall rate characteristics of different assets over the observation period. LINK (9.7%) and BTC (6.8%) maintained relatively high positive rates, while OP (−4.1%), SOL (−2.1%), and ADA (−0.5%) exhibited a structural negative rate. Such a long-term structural difference often stems from microstructural friction in a particular asset's spot lending market, or from the specific role the asset plays in hedging strategies (such as cash-and-carry arbitrage or staking-mining hedges).

In addition, the rate correlation between BTC and ETH constitutes an independent macro state variable. When their rates are highly positively correlated, the market is being driven by a unified macro force (such as Federal Reserve policy expectations or ETF inflows) and systemic risk is highly concentrated; when the correlation falls or even turns negative, capital is being allocated differentially based on asset-specific narratives (such as an Ethereum upgrade or a Bitcoin halving), and the market is in a phase of structural divergence. For a trader holding a multi-asset portfolio, a sudden change in rate correlation is a key signal for adjusting the hedge ratio.

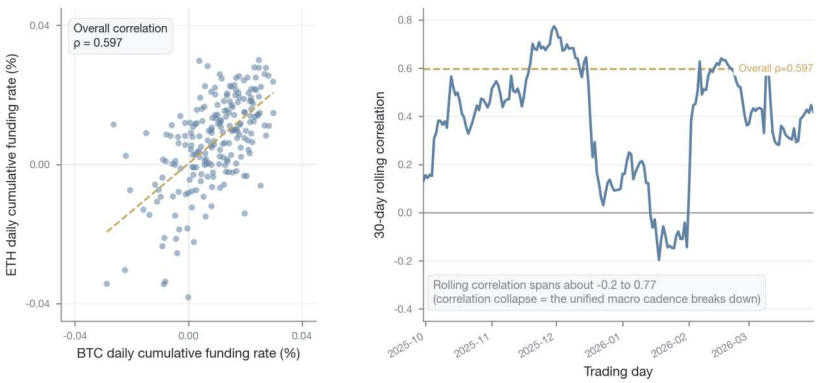


Figure 10-29. A scatter plot of daily BTC and ETH funding rates and their 30-day rolling correlation (Data source: measured daily cumulative funding rates for BTC and ETH perpetuals on Binance, September 2025 to March 2026; as of June 30, 2026)

As Figure 10-29 shows, although BTC and ETH maintained an overall positive correlation of about 0.597 over the sample period, their 30-day rolling correlation coefficient exhibited pronounced time variation (ranging from about -0.20 to 0.77). A collapse in correlation often carries far-reaching systemic implications. When the rates of BTC and ETH diverge, it usually means that the unified pace of macro capital has been broken and the market has entered a phase of structural divergence based on asset-specific narratives (such as expectations of an Ethereum upgrade or Bitcoin ETF inflows). Conversely, when the correlation rises sharply to a high level, it often means the market is in a state of systemic risk-on or risk-off, with all assets driven by the same macro force.

Two methodological qualifications are in order. First, a correlation collapse may partly reflect a macro-level flight-to-quality effect (BTC receiving safe-haven inflows on the "digital gold" narrative while ETH comes under pressure from DeFi deleveraging), rather than pure crypto-native narrative rotation. A formal attribution analysis should introduce macro control variables (such as the U.S. Dollar Index (DXY) and the global risk-appetite gauge VIX) to distinguish the relative contributions of macro transmission and crypto-native factors. Second, the BTC-ETH bivariate analysis can naturally be extended to a principal component analysis (PCA) of the full-sample (10+ assets) rate correlation matrix. If the variance explained by the first principal component surges from about 40% in normal periods to more than 80% in a crisis, it would confirm the hypothesis that "rates collapse into a single systemic factor in a crisis." This constitutes a natural extension for future research.

10.8.4 Comparison with traditional sentiment indicators

In both traditional financial markets and the crypto market, there are many indicators for measuring market sentiment and state. The funding rate, however, has advantages over survey- or text-based sentiment indicators along two dimensions—directness and economic consequence—the core reason being "the price paid in real money."

Most traditional sentiment indicators are indirect, inferential, or soft indicators based

on questionnaires and text analysis. The funding rate, by contrast, is a cash flow that market participants must actually pay to maintain their directional exposure. Table 10-8 systematically compares the funding rate with the main traditional sentiment indicators:

Indicator	Data basis	Update frequency	What it measures	Directness for the perpetual futures market
Funding rate	Long-short imbalance in perpetual futures	Every 1–8 hours	Leverage cost = a direct measure of risk	Extremely direct: generated by the market itself, with real economic consequences
VIX	Implied volatility of S&P 500 options	Continuous	Expected volatility over the next 30 days	Indirect: measures the traditional market, not directly corresponding to crypto
Put/call ratio	Options volume	Daily	Bearish vs. bullish sentiment	Indirect: the crypto options market is far smaller than the perpetual futures market
Fear and Greed Index	Multi-factor composite	Daily	A composite measure of fear/greed	Indirect: factor weights are subjective and not reproducible
Social-media sentiment	Text sentiment analysis	Real-time	The direction of public sentiment	Indirect: noisy and easily manipulated
Open interest	Derivatives position size	Continuous	Total leverage accumulation	Relatively direct: but contains no directional information

Table 10-8. A multidimensional comparison of the funding rate with traditional sentiment indicators (Data source: compiled by the author)

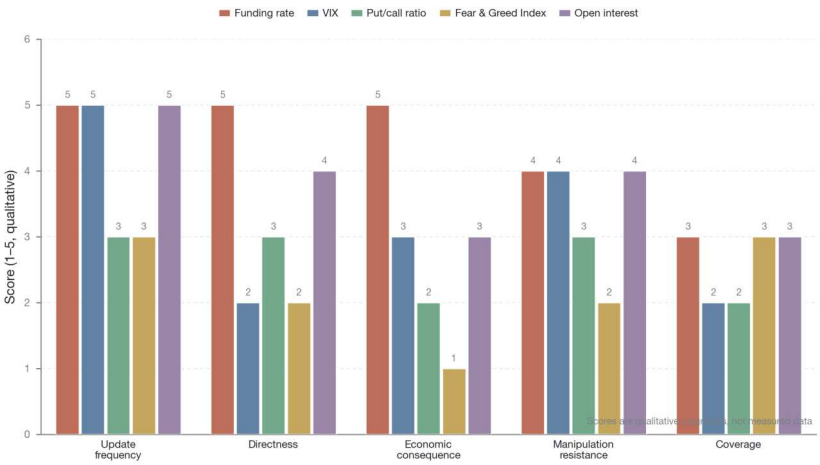


Figure 10-30. A multidimensional comparison of the funding rate with traditional sentiment indicators (subjective/qualitative 1–5 scoring, assessed by the author, not quantitative measurement; the dimensions correspond to Table 10-8 in the main text, the figure shows a selection of five representative indicators, and the cross-comparison reflects the author's judgment)

The predictive value of the rate stems from its directness: it is not a sentiment indicator derived indirectly through an intermediate model but a price signal, with real economic consequences, generated by the perpetual futures market itself every 1–8 hours. Traders actually pay or receive the rate each period, which gives the rate higher credibility than survey- or model-based indicators.

As the comparative analysis in Figure 10-30 shows, the funding rate demonstrates unique advantages along several dimensions. On directness and economic consequence, an extreme value of the Fear and Greed Index does not directly bankrupt any market participant, but an extreme funding rate does: a high rate genuinely erodes margin and forces traders to close positions, thereby directly triggering a chain price reaction. It not only measures risk; it is itself a catalyst of risk. On update frequency, most traditional indicators are daily, whereas the funding rate is updated continuously or at high frequency (such as hourly or every 8 hours), capturing instantaneous shifts in market sentiment at extremely fine granularity. On manipulation resistance, social-media sentiment or questionnaires are easily manipulated by bots or coordinated behavior, whereas to manipulate the funding rate, a manipulator must build a real, large one-sided position in the derivatives market and bear the enormous price risk and friction cost that entails. On directional clarity, unlike total open interest, which can show only the scale of leverage, the sign of the funding rate directly gives the direction of the leverage tilt.

The funding rate is not a description of market sentiment but the pricing of market sentiment. This distinction carries important methodological significance: the Fear and Greed Index or social-media sentiment analysis belongs to the category of "observation," and a researcher can obtain and interpret these signals at no cost; the funding rate belongs to the category of "pricing," and every position holder is paying for its directional view with real cash flow. This economic constraint makes the noise in the rate

signal inherently lower than that in cost-free observational indicators, because traders holding positions in the wrong direction are continuously drained by the rate until they exit.

10.8.5 Empirical evidence and limitations of predictive power

Based on the analysis above, the funding rate as a state variable exhibits, at the descriptive level, statistical associations consistent with systemic fragility, volatility bursts, and cross-asset rotation. Existing academic research also supports this judgment. Inan (2025) shows that the funding rate of perpetual futures has significant out-of-sample predictability, indicating that the rate encodes forward-looking information about the state of the Bitcoin market [34]. Presto Research (2024) finds a statistically significant but weak positive correlation between rate changes and contemporaneous price changes (R^2 of about 12.5%), but near-zero out-of-sample predictive power for the next-period price change of a single asset; that study did not specifically test extreme rates or price reversals [35].

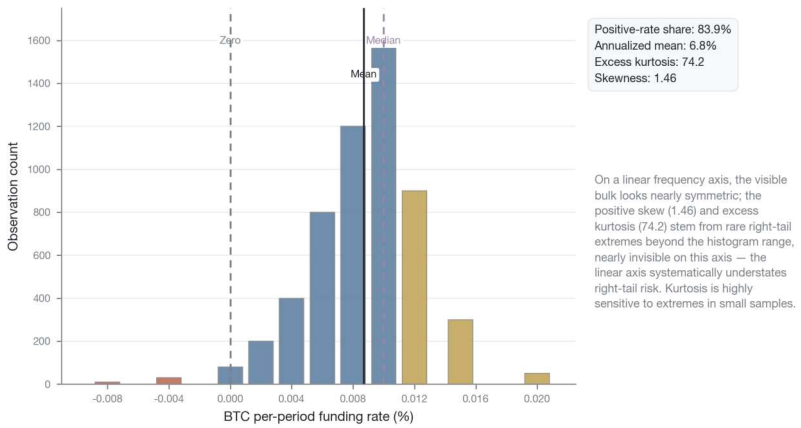


Figure 10-31. The asymmetry and long-tail characteristics of the BTC funding rate distribution (the distribution statistics—skewness, excess kurtosis, positive-rate share, annualized mean, etc.—are computed by the author from the Hyperliquid API; descriptive and sensitive to extreme values in small samples [33]; the predictability $R^2 \approx 12.5\%$ is cited from [35] and the out-of-sample predictability from [34])

Figure 10-31 shows the empirical distribution of the BTC funding rate. Its distribution exhibits extreme positive skew (skewness = 1.46) and very high kurtosis (excess kurtosis = 74.2, i.e., the deviation relative to the normal-distribution kurtosis of 3; note that this estimate is highly sensitive to extreme observations in small samples). The positive-rate share is as high as 83.9% and the annualized mean is about 6.8%, confirming the long-standing "structural long bias" of the cryptocurrency market. This extreme non-normal distribution means that traditional finance models based on the normality assumption will fail severely when handling the funding rate, and the rate's tail risk (especially the right tail) far exceeds theoretical expectations.

The empirical evidence presented in this section is descriptive rather than a formal inferential test. The conditional relationship between the rate level and future volatility shown in Figures 10-24 through 10-26, although consistent with economic intuition, has not been rigorously validated through multivariate regression controlling for other predictor variables (such as realized volatility and the rate of change of open interest) or through Granger causality tests. In addition, the Hyperliquid data [33] used in this chapter have the limitations of a short sample period (September 2025 to March 2026, about six months) and a single-exchange source; differences in user composition and liquidity characteristics across exchanges may cause heterogeneity in rate dynamics, and conclusions drawn from a single platform are not necessarily universally applicable across platforms. Elevating the descriptive findings to formal predictive regressions and out-of-sample tests is a priority direction for future research. Specifically, a standard predictive regression specification can be written as:

$$\sigma_{t+7} = c + \beta_1 |FR_t| + \beta_2 \sigma_t + \beta_3 OI_t + \varepsilon_{t+7}$$

where σ_{t+7} is the realized volatility over the next 7 days, $|FR_t|$ is the absolute value of the current rate, σ_t is the current realized volatility (controlling for the volatility-clustering effect), and OI_t is the rate of change of open interest. Because the overlapping 7-day window of σ_{t+7} introduces moving-average autocorrelation, a formal inferential test should use Newey-West heteroskedasticity- and autocorrelation-consistent standard errors, with a lag truncation parameter no less than the window length (i.e., 7 periods). If β_1 remains statistically significantly positive after controlling for σ_t and OI_t, the incremental predictive power of the rate level for future volatility can be formally confirmed. The current evidence supports this direction at the level of the conditional distribution, but rigorous statistical inference is left for subsequent research on a longer sample period and multi-exchange data.

At the methodological level, two technical issues also warrant attention. The first is **collinearity**: $|FR_t|$ and σ_t may be highly collinear, since in a high-volatility environment the rate itself tends to be more extreme, which may destabilize the estimate of β_1. Mitigations include orthogonalization or principal-component methods, or using a HAR-RV-style multi-scale benchmark model (introducing daily/weekly/monthly realized-volatility components) and testing $|FR_t|$ as an incremental predictor to gauge its marginal contribution after controlling for known volatility persistence. The second is **the characterization of the evidence**: the nature of the current evidence should be labeled more explicitly. The conditional-distribution comparisons shown in Figures 10-24 through 10-28 are descriptive analysis, and moving from descriptive findings to a formal predictive-causal claim requires completing out-of-sample tests with multivariate controls (such as rolling-window out-of-sample R^2). These two methodological improvements are the necessary steps for turning the chapter's descriptive findings into a rigorous predictive model.

Using the funding rate as a predictor, however, also has inherent limitations that must be treated with care.

On timeliness, an extreme funding rate can signal with high confidence that the system is fragile, but it cannot pinpoint the timing of a collapse. A market can persist in an irrationally high-rate state for weeks or even months (as in the first quarter of 2021

or the early-2024 bull phase), and shorting too early on the basis of a high rate may lead to severe paper losses. The rate's predictive power lies mainly in the volatility dimension: it can indicate that the probability of a volatility increase is rising, but it cannot pinpoint direction or timing. This is similar in nature to the predictive power of VIX and is an inherent limitation of the funding rate as a state variable. Structural shifts are a second confounding factor. As noted in Section 10.2, the rate contains a financing cost layer, and when the macro risk-free rate changes sharply (such as during a Federal Reserve hiking cycle) or when structural friction appears in the spot lending market, the center of the rate shifts. Without using the three-layer decomposition model to strip away these structural factors, looking only at the absolute rate level may produce false overheating or over-cooling signals. In addition, the trend of rate financialization is lowering the rate's signal-to-noise ratio as a sentiment indicator: with the rise of systematic arbitrage entities such as Ethena, extreme rates are flattened faster, and when large amounts of capital systematically short perpetual futures to earn a positive rate, their collective behavior itself becomes a force depressing the rate, forming a new reflexive loop in which more rate changes reflect only the entry and exit of arbitrage capital rather than genuine speculative sentiment. This structural impact of rate financialization and its dilution of the state variable hypothesis are developed in Section 10.8.6.

In sum, the funding rate is a powerful state variable that requires careful interpretation. It is not a simple "overbought/oversold" oscillator but a comprehensive macro indicator reflecting the market's leverage structure, liquidity condition, and arbitrage-game equilibrium. Using it correctly requires combining the three-layer decomposition model to strip away structural factors, the volatility dimension to judge uncertainty, and the cross-asset surface to locate the risk distribution. Only in this way can the rate be refined from a crude directional signal into a precise diagnostic tool for the market state.

10.8.6 Rate financialization: Ethena and the funding rate swap market

The analytical framework of the funding rate as a state variable is facing a structural challenge: the funding rate itself is evolving from a passive holding cost into a financial asset that can be independently traded, hedged, and structured. This transformation, represented by the Ethena protocol and Boros Finance's funding rate swap products, marks the arrival of the era of "rate financialization."

Ethena's USDe stablecoin is the most systemically significant instance of rate financialization. USDe's core mechanism is to hold spot ETH (or BTC) and establish an equal-size short position in the perpetual futures market, thereby systematically collecting the positive funding rate in a market-neutral manner. As of early 2026, USDe's size has reached billions of dollars, making Ethena an important structural short seller in the perpetual futures market (accounting for several percent of ETH perpetual open interest, per reports from outlets such as The Block). Ethena's scale effect has a dual impact on the rate market. On the supply side, billions of dollars in persistent short positions inject stable sell-side liquidity into a positive-rate environment, accelerating the rate's return to equilibrium in normal markets and objectively strengthening the

anchoring mechanism. On the crisis side, however, when the rate suddenly turns negative (as during the panic selling triggered by the DeepSeek event in late January 2025), Ethena's enormous short positions face reverse rate-outlay pressure. At that point, Ethena not only cannot collect income but becomes a net payer of the rate, and its forced position-reduction may further intensify market volatility.

More systematically, the risks Ethena faces can be analyzed along three dimensions. On the **economic-risk contagion path**: a persistently negative rate → USDe yield turns negative → large-scale redemptions by holders → Ethena is forced to close short positions (executing large buys in the contract market) → an upward price shock is produced → if the redemption pressure causes USDe to depeg, its role as widely used collateral in DeFi lending protocols will amplify cross-protocol contagion. On the **operational-risk** dimension: Ethena holds billions of dollars in short positions on centralized exchanges, facing exchange custody risk (the roughly $1.5 billion hack of Bybit in February 2025 is a real-world warning of this kind of risk), account-freeze risk, and execution-slippage risk when rebalancing large positions across exchanges. On the **regulatory-risk** dimension: USDe's yield mechanism may be characterized by the SEC as an unregistered investment contract (the Howey test)—users deposit funds and expect returns from an actively managed strategy run by the team, which fits the characteristics of an investment contract closely; in addition, the stablecoin classification and reserve requirements under the EU's Markets in Crypto-Assets (MiCA) framework may also constrain USDe's compliance.

Funding rate swap products such as those of Boros Finance push rate financialization from another dimension. A funding rate swap allows a trader to convert floating funding-rate exposure into a fixed yield (or vice versa), essentially creating a derivatives market with the funding rate as the underlying asset—a "derivative of a derivative." Such products enable institutional investors to express a view on rate direction or volatility purely, without holding any perpetual futures position.

The core impact of rate financialization on the state variable hypothesis is a decline in the signal-to-noise ratio. As a growing share of rate changes reflect the capital inflows and outflows of arbitrage entities such as Ethena rather than genuine shifts in speculative sentiment, the rate's signal value as a "market thermometer" is systematically diluted. Quantifying this dilution effect (for example, by decomposing rate changes into a component from systematic arbitrage capital and a component from directional speculation) is an important topic for future research. Chapter 18 will further analyze the long-term impact of Ethena's scale expansion on the market microstructure from a systemic-risk perspective.

10.9 Chapter summary

This chapter has conducted a systematic game-theoretic analysis of the core mechanism of perpetual futures—the funding rate—from micro mechanism to macro state. The funding rate is not a simple interest-payment formula but an automatic regulator arising from the interaction of multiple layers of games in the crypto derivatives market. The chapter's analysis is organized around five nested theoretical frameworks.

The three-layer decomposition model deconstructs the apparently single funding rate into a financing cost layer (a slowly varying institutional benchmark, about 10.95% annualized), a sentiment premium layer (the high-frequency directional preference of the market), and an institutional friction layer (the stable-band effect created by the clamp function; BitMEX research shows the rate is locked exactly at 0.01% for 78.19% of the time [8]). This framework shows that the rate is in fact a hybrid product of the macro interest rate, market speculative preference, and microstructural contract design, and that only by separating the three layers can the meaning of the rate signal be accurately diagnosed.

Building on this, the anchoring strength spectrum reveals the nonlinear character of the funding rate's effectiveness. Anchoring strength is a monotonically increasing function of the intensity of arbitrage activity, and arbitrage weakens systemically in extreme markets because six conditions—capital constraints, counterparty risk, liquidity drought, execution risk, liquidation risk, and cross-asset margin contagion—deteriorate simultaneously. The resulting anchoring paradox—that anchoring strength falls to its lowest precisely when the market most needs anchoring—is validated in events such as the March 12, 2020 episode when the BTC perpetual price fell hundreds of dollars below spot [5] and the October 10, 2025 episode with a visible 24-hour liquidation notional of about $19 billion [17].

The analytical framework of anchoring strength extends further into the three-layer game model, which abstracts the dynamic evolution of the funding rate into three nested games: the innermost anchoring game (the war of attrition and reflexive loop between longs and shorts over the basis), the middle convergence game (cash-and-carry arbitrageurs driving price convergence and cross-exchange arbitrageurs driving rate convergence; the annualized return ranges of cross-exchange arbitrage are given in Section 10.4.2 [22]), and the outermost strategic game (the rules attack and defense between exchanges and attackers). The balance of forces among the three layers determines the markedly different dynamics the rate exhibits in the normal, stress, and crisis market states.

When the game equilibrium is broken, the rate polarization positive feedback mechanism is set in motion. The condition for the flip from negative to positive feedback is that the directional profit exceeds the rate cost, while the disposition effect, confirmation bias, and herding jointly drive holders' refusal to close. Polarization terminates in only one of two ways: mild mean reversion or a liquidation cascade. In the PIPPIN case, shorts' rate outlay alone reached -19.77% of notional principal within 24 hours, illustrating the effect of positive feedback amplified by high-frequency settlement. Rate polarization thereby becomes a directly observable signal that the leverage cycle is approaching its critical point, which leads to the chapter's fifth framework—the funding rate as state variable hypothesis—elevating the rate from a passive output variable to an active macro state variable.

Figure 10-32 integrates the three information dimensions of the rate (level, volatility, and the cross-asset surface) with their corresponding analytical outputs into a systematic analytical framework.

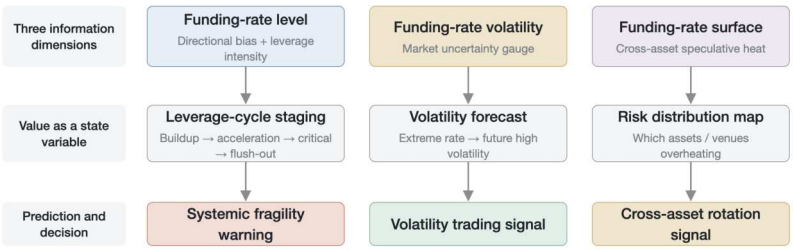


Figure 10-32. The analytical framework and application path of the funding rate as a state variable (a conceptual illustration based on the analytical framework, not empirical data)

As Figure 10-32 shows, the rate's level, volatility, and cross-asset surface together form a multidimensional analytical matrix capable of identifying the phase of the leverage cycle, predicting future volatility bursts, and mapping the risk distribution of the entire market. Compared with indirect indicators such as VIX, the funding rate can reflect the market's leverage-structure state at higher frequency and with lower latency. The spot price reflects the market's instantaneous supply and demand, while the funding rate reveals the market's leverage structure: the rate continuously prices the positions of longs and shorts through actual cash-flow transfers, and the flow of arbitrage capital delineates the effectiveness boundary of the anchoring mechanism. At the same time, the trend of rate financialization is creating a new game dimension: when the rate itself becomes a tradable asset, the game around it will grow more complex still.

The funding rate does not exist in isolation. It is closely linked to the accumulation of leverage, the triggering of the liquidation mechanism, and the depth of liquidity pools. The chapter's analysis points to a core conclusion: the boundary of the funding rate's anchoring efficacy is precisely the starting point at which leverage and liquidation risk begin to dominate the market. When the rate loses its anchoring force in a crisis state, what takes over the market is no longer the game of long-short willingness but the forced execution of the liquidation mechanism. The next chapter starts from this breaking point of rate failure and enters the extreme-risk domain of the liquidation cascade, analyzing how the forced-liquidation mechanism operates and fails when the price breaches the maintenance-margin floor. The funding rate determines the cost of leverage, while the liquidation mechanism determines the survival of leveraged positions.

References

[1] James, N. (2021). Changes to the extreme and erratic behaviour of cryptocurrencies during COVID-19. *Physica A: Statistical Mechanics and its Applications*, *565*, 125462. https://pmc.ncbi.nlm.nih.gov/articles/PMC7687370/ https://doi.org/10.1016/j.physa.2020.125581

[2] Godbole, O. (2020, March 13). Bitcoin price briefly dips to 12-month low in overnight trading. *CoinDesk*. https://www.coindesk.com/markets/2020/03/13/bitcoin-price-briefly-dips-to-12-month-low-in-overnight-trading

[3] CoinDesk. (2020, March 12). Bitcoin's crash triggers over $700M in liquidations

on BitMEX. https://www.coindesk.com/markets/2020/03/12/bitcoins-crash-triggers-over-700m-in-liquidations-on-bitmex

[4] CryptoSlate. (2020). BitMEX freezes as XBTUSD crashes 10% below Bitcoin spot price. https://cryptoslate.com/bitmex-freezes-as-xbtusd-crashes-10-below-bitcoin-spot-price/

[5] Multicoin Capital. (2020, March 17). March 12: The day crypto market structure broke (Part 1). https://multicoin.capital/2020/03/17/march-12-the-day-crypto-market-structure-broke/

[6] Alexander, C., Choi, J., Park, H., & Sohn, S. (2020). BitMEX bitcoin derivatives: Price discovery, informational efficiency, and hedging effectiveness. *Journal of Futures Markets*, *40*(1), 23–100. https://onlinelibrary.wiley.com/doi/abs/10.1002/fut.22050 https://doi.org/10.1002/fut.22050

[7] MilkRoad. (n.d.). Perpetual contracts: Definitions & FAQ for funding rates. https://milkroad.com/funding/perpetual-contracts/

[8] BitMEX Research. (2025). Funding rates structure: Floor & ceiling explained — Q3 2025 derivatives report. https://www.bitmex.com/blog/2025q3-derivatives-report [Industry report, not peer-reviewed]

[9] Palepu, A. (2020, June 23). Funding rates: Under the hood. *Medium / DerivaDEX*. https://medium.com/derivadex/funding-rates-under-the-hood-352e6be83ab

[10] Binance Support. (2025). Important updates on funding rate settlement frequency. https://www.binance.com/en/support/announcement/detail/6d707b1f7ae34f419621b4c807464ab1

[11] Hyperliquid. (n.d.). Funding. *Hyperliquid Documentation*. https://hyperliquid.gitbook.io/hyperliquid-docs/trading/funding

[12] Schmeling, M., Schrimpf, A., & Todorov, K. (2023). Crypto carry (BIS Working Paper No. 1087). Bank for International Settlements. https://www.bis.org/publ/work1087.htm

[13] Bloomberg. (2026, January 21). Wall Street pulls back from a money-spinning Bitcoin trade.

[14] Ackerer, D., Hugonnier, J., & Jermann, U. (2025). Perpetual futures pricing. *Mathematical Finance*. https://doi.org/10.1111/mafi.70018

[15] Gabaix, X., & Maggiori, M. (2015). International liquidity and exchange rate dynamics. *Quarterly Journal of Economics*, *130*(3), 1369–1420. https://doi.org/10.1093/qje/qjv016

[16] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[17] CoinShares. (2025, October 20). Billions in crypto liquidations: Inside October's $19B crash.

[18] Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *Journal of Finance*, *52*(1), 35–55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

[19] Scheinkman, J. A., & Xiong, W. (2003). Overconfidence and speculative bubbles. *Journal of Political Economy*, *111*(6), 1183–1220. https://doi.org/10.1086/378531

[20] Collins, T. (2025). The future of Bitcoin: Market maturity, strategic reserves, and the paradox of institutional accumulation. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5788582 https://doi.org/10.2139/ssrn.5788582

[21] De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, *98*(4), 703–738. https://doi.org/10.1086/261703

[22] Pendle Team. (2025, December 19). Cross-exchange funding rate arbitrage: A fixed-yield strategy through Boros. *Medium / Boros Finance*. https://medium.com/boros-fi/cross-exchange-funding-rate-arbitrage-a-fixed-yield-strategy-through-boros-c9e828b61215 [Industry report, not peer-reviewed]

[23] Baumgartner, T., & Guettler, A. (2022). Trust in a trustless network: The case of the Bitcoin flash crash on May 19, 2021. *EUROFIDAI-ESSEC Paris December Finance Meeting*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4260876 https://doi.org/10.2139/ssrn.4260876

[24] Gromb, D., & Vayanos, D. (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics*, *66*(2–3), 361–407. https://doi.org/10.1016/S0304-405X(02)00228-3

[25] CoinDesk. (2021). Bitcoin drops below $31K before rebounding; $8B in liquidations triggered.

[26] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[27] Binance Square. (2025). The reasons behind Alpaca's pre-delisting short squeeze.

[28] Skwarek, M. (2025). Why do investors behave irrationally in the cryptocurrency and emerging stock markets? *SAGE Open*, *15*(3). https://doi.org/10.1177/21582440251361212

[29] Banerjee, A. V. (1992). A simple model of herd behavior. *Quarterly Journal of Economics*, *107*(3), 797–817. https://doi.org/10.2307/2118364

[30] Delfabbro, P., King, D. L., & Williams, J. (2021). The psychology of cryptocurrency trading: Risk and protective factors. *Journal of Behavioral Addictions*, *10*(2), 201–207. https://doi.org/10.1556/2006.2021.00037

[31] Minsky, H. P. (1992). The financial instability hypothesis (Working Paper No. 74). The Jerome Levy Economics Institute of Bard College. https://www.levyinstitute.org/pubs/wp74.pdf

[32] U.S. Department of Justice. (2023, January 10). Avraham Eisenberg charged in connection with manipulation of Mango Markets decentralized cryptocurrency exchange. https://www.justice.gov/usao-sdny/pr/avraham-eisenberg-charged-connection-manipulation-mango-markets

[33] Hyperliquid. (2026). Hyperliquid API documentation: Funding rate history. https://hyperliquid.xyz/

[34] Inan, E. (2025). Predictability of funding rates (SSRN Working Paper No. 5576424).

https://doi.org/10.2139/ssrn.5576424

[35] Presto Research. (2024, August 2). Can funding rate predict price change? https://www.prestolabs.io/research/can-funding-rate-predict-price-change

The Endogenous Reflexivity of Leverage and Liquidation

On October 10, 2025, a macroeconomic report about an unexpected deadlock in U.S.–China trade negotiations—the U.S. president's announcement of a 100% tariff on Chinese imports—stirred only mild ripples in traditional financial markets. When the same news reached the crypto perpetual futures market, which trades around the clock, it unexpectedly triggered a systemic liquidation event of unprecedented scale. Within roughly 24 hours, visible market-wide forced liquidations exceeded $19 billion in notional value; measured over the full deleveraging process, the window spanned about 36 hours. The event affected roughly 1.62 million trading accounts, of which about 87% held long positions [1]. The price of Bitcoin plunged from an intraday high of about $122,500 and reached a low of roughly $104,800 on the spot/aggregate-index basis (an intraday drawdown of about 14.5%; because depth had dried up, perpetual futures printed even deeper wicks, with individual venues briefly touching about $102,000). Market-wide open interest evaporated by about $19.2 billion within the 40-minute cascade window (a peak-to-trough decline of about 25%), order-book depth on major exchanges collapsed by more than 99%, bid-ask spreads widened abruptly by several orders of magnitude, and liquidity dried up suddenly amid violent price swings (the complete microstructure data appear in Section 11.9.2, Table 11-7) [2].

This constitutes a structural paradox. In the weeks before the event, the vast majority of traders operated at leverage in the seemingly "safe" range of 5x to 10x. They set stop-losses, kept margin buffers, and believed their risk management was prudent. Yet when the market began to fall, these individual-level prudential measures not only failed to prevent a systemic crisis but intensified it: every stop-loss and liquidation that was triggered injected fresh selling pressure into the market, and together they drove a systemic collapse. This phenomenon, in which individual rationality leads to collective irrationality, reveals a structural blind spot in the microstructure of perpetual futures markets: when everyone manages risk "rationally" in the same direction, their collective behavior itself constitutes the greatest systemic risk.

The "10.10 event," along with the many similar crashes in crypto history—from the "3.12" COVID crash of March 2020 to the shock of China's mining ban in May 2021—was not a mere "black swan" or a chance accident. Such episodes are the periodic eruption of a predictable systemic risk driven by the market's own internal structure. This risk originates in an endogenous reflexivity feedback loop that forms between

leverage and automated liquidation mechanisms: under specific conditions, leverage combines with the automated execution of the liquidation engine to become a counter-intuitive risk amplifier, forming a positive-feedback mechanism that can self-reinforce and drive systemic collapse. As the systematic analysis of liquidation mechanisms in Sections 11.1 and 11.6 reveals—examining, from a mechanism-design perspective, how the liquidation engine seeks a balance between protecting system solvency and maintaining market stability—the liquidation mechanism is meant to "clear bad debt and protect the system," yet under conditions of high leverage, limited depth, and positive feedback, liquidation itself becomes the creator of systemic risk rather than its solver.

To dissect this complex dynamic process, this chapter builds a two-layer theoretical framework. At the philosophical level, it introduces George Soros's theory of reflexivity, which supplies a macro perspective for understanding the non-equilibrium nature of markets; at the mechanism level, it applies the liquidity spiral model of Brunnermeier and Pedersen, which supplies the economic tools for dissecting market micro-processes. With their extremely high leverage ceilings (up to 125x), around-the-clock trading, automated liquidation, and global access, perpetual futures have become the most extreme and purest testing ground for these two frameworks. This chapter first dissects the micro transmission chain of leverage and liquidation, clarifying the mechanical process that runs from a price move to forced liquidation; it then establishes the two-layer theoretical foundation of reflexivity. Building on this, the chapter constructs five core theoretical tools in turn: the liquidation reflexivity equation, which formalizes the liquidation cascade; the compound positive feedback that couples the funding rate with liquidation; the impossible trinity of liquidation-mechanism design; the time-compression effect unique to perpetual futures; and the information paradox of liquidation together with mechanism-induced volatility. Finally, it applies these tools jointly to a complete reflexivity dissection of the "10.10 event," testing the explanatory power of the entire analytical framework.

11.1 The transmission chain of leverage and liquidation

The core function of leverage is capital efficiency: it allows a trader to control a position far larger than their own capital. The price of this privilege, however, is a precise, stringent, code-automated risk-management system—the margin and liquidation system. This system builds a deterministic transmission chain that runs from continuous fluctuations in the market price to the forced liquidation of a trader's position. Understanding each link in this chain—how margin constrains leverage, how the mark price defines "truth," how the liquidation engine enforces its verdict, and how the margin mode determines the path of risk transmission—is a necessary precondition for understanding how systemic risk emerges from individual behavior.

11.1.1 Margin and the leverage constraint

In perpetual futures trading, leverage does not arise from nothing; it rests on the margin system. Initial margin is the minimum amount of a trader's own capital that must be committed to open a leveraged position. Its formula is straightforward:

> Initial margin = position notional value / leverage multiple

For example, a trader who wants to open a $100,000 BTC long at 10x leverage must post $10,000 as initial margin. The initial margin rate directly determines the maximum leverage a trader can use: if the initial margin rate is set at 5%, the trader can use up to 20x leverage; if it is 1%, up to 100x; and in the lowest tier at some exchanges, an initial margin rate of 0.8% implies an extreme 125x leverage.

The minimum capital level required to keep a position open is called maintenance margin, and this ratio is typically far below the initial margin. For the same $100,000 position, for instance, the maintenance margin requirement might be only 0.4% to 0.5%, or $400 to $500. The difference between initial margin and maintenance margin constitutes the trader's liquidation buffer. The size of this space directly reflects the fragility of the position: when the market price moves adversely and the account's margin balance (initial margin plus or minus unrealized profit and loss, or P&L) falls below the maintenance margin level, the liquidation process is triggered automatically. For a long position, the liquidation price can be expressed, in simplified form, as:

> Liquidation price (long) $\approx$ entry price $\times$ (1 $-$ initial margin rate + maintenance margin rate)

This formula is a first-order approximation that ignores the continuous erosion of margin by trading fees and funding payments (the latter is addressed by the dynamic formula below). It essentially represents a lower bound on the price decline required to trigger liquidation. In actual trading, because funding costs and fees accumulate, the true liquidation price is closer to the entry price than the formula's estimate.

The formula above, however, assumes that the margin balance is affected only by price changes and ignores the continuous consumption of margin by the funding rate (see Section 11.4). Once the time effect of the funding rate is included, the liquidation price is no longer a static constant but a dynamic variable that rises monotonically with holding time. For a long position, the time-dependent liquidation price can be expressed as:

$$P^{\text{liq}}(t) = P_0^{\text{liq}} + \frac{1}{Q}\int_0^t r(s)\cdot V\,ds$$

where P_0^{liq} is the static liquidation price at the moment the position is opened, $r(s)$ is the funding-rate density at time s (the rate per unit time, positive for the party paying on a long position), V is the position notional value, and Q is the position size (the number of coin-denominated contracts). Since $V = Q \cdot P$, the integrand $r(s) \cdot V$ is the cumulative funding payment per unit time (with dimensions of dollars per unit time); dividing by Q converts it into the rate at which the liquidation price rises (dollars per coin per unit time), consistent with the dimensions of the left-hand side $P^{\text{liq}}(t)$ (dollars per coin). Equivalently, this can be written as $P^{\text{liq}}(t) = P_0^{\text{liq}} + \int_0^t r(s)\,P(s)\,ds$. Contract type must be distinguished: for a linear contract (USDT-margined), $1/Q$ is simply the reciprocal of the number of coins held, and the expression can be written directly as $r(s)\,P(s)$; for an inverse contract (coin-margined), the face

value is denominated in coin terms, so a nonlinear transformation of the coin price must be layered on top of $1/Q$. The convention for $r(s)$ must also be specified: taken as the rate density per unit time, $\int r\,P\,ds$ is the cumulative price rise; if it is recorded per exchange convention as a discrete fraction per settlement period, the integral is replaced by a discrete sum over the settlement times. Moreover, this integral is a continuous approximation of discrete settlement: most exchanges settle the funding rate every 8 hours (some, such as OKX, have shortened this to every 4 hours or even every hour), and the margin deduction occurs at the settlement moment rather than continuously. This discreteness introduces a "funding-settlement-timing fragility": a single deduction at the settlement instant may push a marginally safe position into the liquidation zone, especially when settlement coincides with intensifying volatility. For this reason, some protocols (such as Hyperliquid) have moved to a continuously accruing funding rate, smoothing settlement from a discrete jump into a continuous micro-adjustment and eliminating the concentration of fragility at the settlement instant.

When the funding rate stays positive, the integral term accumulates over time and the liquidation price $P^{\text{liq}}(t)$ moves steadily toward the current market price. The core implication of this formula is that, even if the market price does not change at all, the continuous consumption by the funding rate alone can drive the liquidation price to reach the market price within a finite time, forcing the position into liquidation. This formula is demonstrated numerically in the analysis of the margin-erosion effect in Section 11.4.

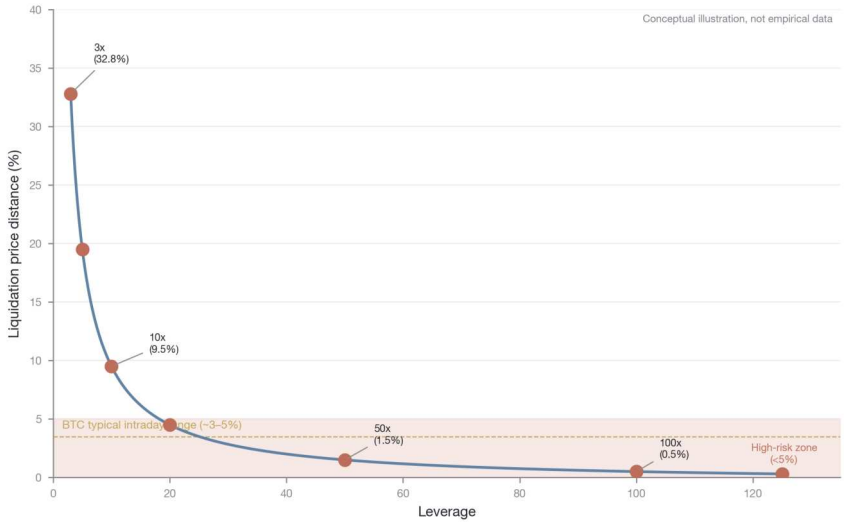


Figure 11-1. Relationship between the leverage multiple and the distance to the liquidation price (a first-order approximation based on the simplified liquidation formula, excluding funding and fees; the orange dashed band marks typical BTC intraday volatility; conceptual illustration)

The core insight is that the higher the leverage and the lower the initial margin rate, the closer the liquidation price is to the entry price, and the sensitivity of the position to price fluctuations rises nonlinearly and sharply. As shown in Figure 11-1, the orange dashed line marks the typical range of BTC intraday volatility (about 3%–5%), and at

leverage above 20x the liquidation distance already falls within this range: at 3x leverage the price must fall by about 32.8% to trigger liquidation, which is a rare and extreme event in BTC's history; at 10x leverage a decline of about 9.5% suffices; at 50x leverage a move of less than 1.5% (which does not even constitute a meaningful intraday move) is enough to force liquidation; and at 100x leverage the liquidation distance is only 0.5%, so almost any random price noise can trigger liquidation. Leverage is therefore an amplifier that transforms "normal market volatility" into "liquidation events," mapping tiny fluctuations in price space into enormous changes in margin space.

Across crypto exchanges, setting the initial margin rate is a core design decision and an ongoing competitive game. Table 11-1 compares the margin parameters for BTC perpetual futures at major exchanges as of 2025 (note: actual margin rates increase in tiers with position size; the lowest tier is shown here):

Exchange	Max leverage	Min initial margin rate	Maintenance margin rate (lowest tier)	Liquidation buffer	Type
Binance	125x	0.80%	0.40%	0.40%	CEX
OKX	125x	0.80%	0.40%	0.40%	CEX
Bybit	100x	1.00%	0.50%	0.50%	CEX
dYdX v4	20x	5.00%	3.00%	2.00%	DEX
Hyperliquid	40x	2.50%	1.25%	1.25%	DEX

Table 11-1. Comparison of margin parameters for BTC perpetual futures at major exchanges (Data source: exchanges' official documentation. CEX figures are lowest-tier/small-position parameters using dynamic tiering, with leverage falling sharply for large positions; DEX figures are BTC-market parameters. Hyperliquid lowered its ceiling from 50x to 40x on March 12, 2025, and the dYdX v4 BTC ceiling is about 20x; data collected in 2025)

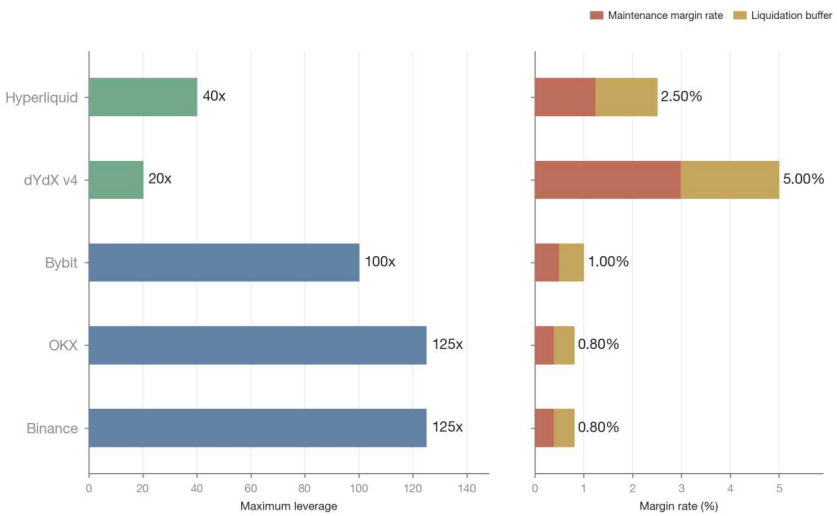


Figure 11-2. Visual comparison of margin parameters across major exchanges (Data source: exchanges' official documentation, as of 2025; CEX = blue, DEX = green; headline leverage is for the

lowest-risk tier and declines dynamically with position size; Hyperliquid lowered its ceiling from 50x to 40x on March 12, 2025, and the dYdX v4 ceiling is about 20x)

As shown in Figure 11-2, centralized exchanges (CEXs) generally offer higher leverage (100x–125x) and correspondingly very narrow liquidation buffers (only 0.4%–0.5%), whereas decentralized exchanges (DEXs) such as dYdX and Hyperliquid take a relatively conservative approach, capping BTC's maximum leverage at about 20x–40x with buffers of about 1.25%–2.0%. This difference reflects different risk appetites under different market positionings: an overly low margin requirement permits extremely high leverage and, while it can attract high-yield users and boost trading volume, markedly increases systemic fragility; an overly high margin requirement limits capital efficiency and drives users away. This game between market share and system safety often devolves into a "race to the bottom." As the risk-substitution theorem of Chapter 5 ("risk cannot be eliminated, only transferred") reveals, there is a structural trade-off between capital efficiency and system safety, and under competitive pressure safety is often the first dimension to be sacrificed (here the capital-efficiency–safety tension is a derived reparameterization of Chapter 5's service-provider impossible trinity from the perspective of leverage risk, rather than that trinity itself; see Section 11.6.5).

Most exchanges use a tiered margin system, in which the initial and maintenance margin rates rise step by step as the position grows. Binance, for example, has more than a dozen tiers for BTC perpetual futures, with position sizes ranging from $50,000 in the lowest tier to hundreds of millions of dollars in the highest, and corresponding maintenance margin rates rising from 0.40% to more than 5%. The purpose of this design is to limit the leverage available on large positions and reduce the potential impact of a single whale on the market.

The tier boundaries themselves, however, can become a counterintuitive risk trigger—the phenomenon of "cross-tier liquidation." Consider a concrete scenario: a trader holds a BTC long with a notional value of $49,000, which sits in the first tier (maintenance margin rate 0.4%), so the required maintenance margin is only $196. When a rise in the BTC price lifts the position's notional value to $51,000 and into the second tier (maintenance margin rate 0.5%), the required maintenance margin jumps to $255, an increase of 30%. If the trader's margin balance happens to be between $196 and $255, a price increase that should have produced a profit instead triggers liquidation, creating the paradox that "the higher it rises, the more dangerous it becomes." This discontinuity, caused by the jump at a tier boundary, is especially pronounced for large positions at high leverage. To mitigate it, some exchanges have begun to introduce smoothed margin curves, replacing the discrete tier steps with a continuously increasing function and thereby eliminating the jump in margin at the boundaries.

The tiered system also introduces a subtle nonlinear effect: when price fluctuations push a position's notional value across a tier boundary, the maintenance margin requirement rises in a jump, which can cause an otherwise safe position to trigger liquidation suddenly. This threshold effect is formalized in the liquidation reflexivity equation of Section 11.3.

11.1.2 The mark-price mechanism

If the liquidation system used the last traded price on a single exchange directly as its trigger condition, then in a thin market a single anomalous trade—or a malicious manipulator using a small amount of capital to "wick" the price (i.e., instantaneously create an extreme price)—could trigger the erroneous liquidation of many other positions. To guard against this risk, almost all major exchanges have introduced the mark price as the benchmark for triggering liquidation, rather than relying on their own last traded price.

The mark price is usually constructed by aggregating the real-time prices of several major spot exchanges (such as Coinbase, Kraken, and Bitstamp) and taking a volume-weighted average. Some exchanges also add a moving average of the funding rate to reflect the basis between perpetual futures and spot.

Take Binance's mark-price calculation as an example: it uses a median-of-three method, where Mark price = Median(Price_1, Price_2, Contract price), with Price_1 = Index price $\times$ (1 + latest funding rate $\times$ fraction of time to next settlement) and Price_2 = Index price + EMA(fair contract price − index price). The median function provides implicit resistance to outliers: even if one of the three components is contaminated or anomalous, the other two still anchor the mark price within a reasonable range. Different exchanges use different mark-price formulas (OKX uses a weighted-moving-average method and Bybit an exponentially weighted scheme), and this difference in calculation means that the same asset can have a slightly different mark price at the same moment on different exchanges, leading to different liquidation-trigger sensitivities. In extreme market conditions, this difference can determine whether a position is liquidated.

Behind this design lies a game-theoretic point: the cost of simultaneously manipulating multiple highly liquid, mainstream spot markets is far higher than the cost of manipulating a single derivatives market. The mark price is therefore considered a better reflection of an asset's "fair value," providing a more robust anchor for liquidation decisions.

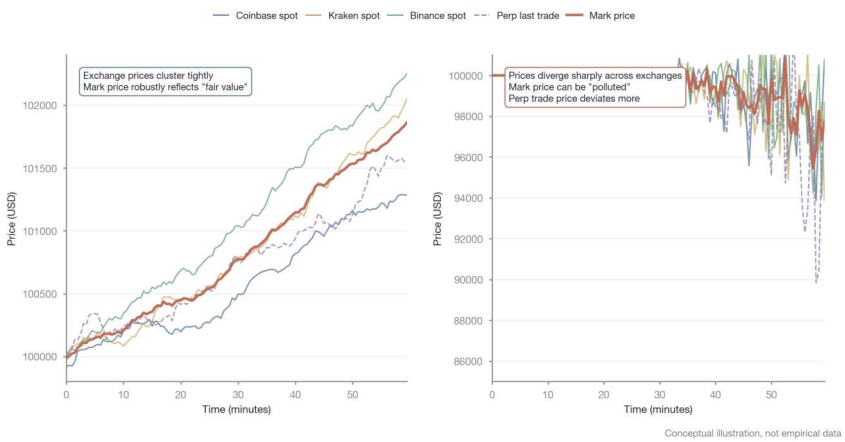


Figure 11-3. Comparison of the mark-price construction mechanism in normal and extreme states (conceptual illustration, not empirical data)

The mark price, however, is not an absolutely reliable benchmark but a derived reference price whose robustness depends on a set of assumptions that hold in normal markets but can break down in extreme conditions. As shown in Figure 11-3, the left panel shows that in a normal market the prices of the various exchanges cluster tightly and the mark price (the thick red line) robustly reflects fair value; the right panel shows that under an extreme shock, spot prices diverge markedly, the perpetual traded price (the purple dashed line) departs substantially from the mark price, and the mark price itself may be "contaminated." When these assumptions collapse in extreme conditions, the mark-price mechanism can turn from a risk buffer into a risk conduit:

On the oracle-risk front, the mark price relies on real-time price feeds from external price sources. For centralized exchanges, this usually means obtaining data via API from other exchanges; for decentralized exchanges (such as dYdX and Hyperliquid), it relies on on-chain oracles (such as Chainlink and Pyth). Either way, the price-feed system can be subject to attack, network latency, or failure. Specifically, oracle failure modes can be grouped into several categories. First, intrusion at the data-source level: an exchange's API keys are stolen, or the exchange itself reports manipulated price data, contaminating the price-feed input at its source. Second, transport-layer attacks: a man-in-the-middle attack or DNS hijacking can tamper with price data as it travels from the source exchange to the oracle node, which is especially insidious in a centralized API price-feed architecture. Third, smart-contract vulnerabilities in on-chain oracles, including reentrancy attacks, access-control flaws, or exploited upgrade mechanisms, through which an attacker can directly tamper with the price stored on-chain via a contract-level vulnerability. Fourth, economic-manipulation attacks: an attacker uses a flash loan to temporarily distort, within a single transaction, the on-chain price source on which the oracle relies (such as the spot price of a Uniswap pool); these attacks mainly target scenarios in DeFi protocols that use an instantaneous on-chain price as the oracle. Fifth, latency arbitrage: exploiting the time lag between the oracle's update frequency and actual price movements, an attacker executes trades when the oracle price has not yet updated but the market price has already moved substantially, profiting from the stale mark price. Once an erroneous mark price is adopted, catastrophic and unfair liquidations follow. As discussed in Chapter 6, maximal extractable value (MEV)—the economic gain an attacker obtains by manipulating or front-running state changes, including oracle price updates—makes this risk especially prominent in the DeFi environment.

On the data-quality front, the quality of the mark price depends on the data quality of its constituent exchanges. If one constituent exchange itself experiences a liquidity drought or a price anomaly (for example, a flash crash), its anomalous price will "contaminate" the weighted average and cause the mark price to deviate from true fair value. During the "10.10 event," when USDe traded at a severe discount on Binance, the prices of USDe-denominated pairs became sharply distorted, and the extent to which this distortion seeped into the mark-price calculation remains a focus of industry debate to this day [3].

During periods of extreme volatility, the basis between the spot market and the perpetual futures market can widen sharply. The last traded price of perpetual futures may depart substantially from the mark price because order-book depth has dried up.

At such times, although the mark price is still "correct" mathematically, there can be an enormous gap between the "fair value" it reflects and the execution price a trader can actually obtain in the perpetual futures market. This means that even when the mark price has not yet reached the liquidation line, a trader may be in de facto distress because they cannot close out at a reasonable price.

The design of the mark-price mechanism is essentially a trade-off between robustness and sensitivity. A mark price that is too robust (for example, one using a long-window moving average) may lag the true price in a rapid decline, delaying the triggering of liquidations and increasing bankruptcy risk; a mark price that is too sensitive is easily affected by short-term noise and manipulation, causing unnecessary liquidations. This trade-off has no perfect solution; it is one of the intrinsic conflicts in liquidation-mechanism design and is analyzed more systematically in the "impossible trinity" framework of Section 11.6.

11.1.3 The liquidation engine

Once the mark price confirms that an account's margin balance is below the maintenance margin requirement, the liquidation engine takes over the position. The liquidation engine is essentially an automated trading program with the highest execution priority, whose sole objective is to force-close a position on the brink of bankruptcy as quickly as possible and with as little market impact as possible, thereby preventing the system from incurring bad debt from underwater positions. The system has an absolutely deterministic character: once the conditions are met, the code executes strictly according to preset rules, regardless of market depth, the trader's wishes, or potential consequences.



Figure 11-4. The transmission chain from a price move to liquidation execution

As shown in Figure 11-4, this transmission chain presents in full the process running from the assessment of the margin balance, through mark-price verification and the choice of liquidation mode, to the liquidation order's impact on the market and, in turn, the triggering of a new batch of liquidations that forms positive feedback. The design of the liquidation engine has undergone a marked shift over time, from a pursuit of efficiency to a balance that also weighs impact, and the mainstream liquidation-execution modes today can be grouped into the following four.

Cliff-edge liquidation is the earliest and most direct mode. When the trigger condition is met, the liquidation engine dumps the entire position onto the market at once as a market order, clearing bad debt with great speed and high execution certainty but with an enormous impact on the order book: in extreme conditions where liquidity is already thin, a single large liquidation market order may "eat through" multiple price levels at once, driving the price down further and triggering more liquidations. BitMEX around 2019 was known for this mode, and the price impact it produced repeatedly sparked accusations of "liquidation hunting." In response to this destructive impact, the industry gradually evolved progressive liquidation: when liquidation is triggered, the position is not closed all at once but reduced in batches. The proportion cut each

time is calculated dynamically by the liquidation engine based on the severity of the margin shortfall and the position's tier—the larger the shortfall and the higher the tier, the larger the proportion cut in a single step. The liquidation orders are submitted as limit orders, with the target execution price set between the bankruptcy price and the mark price, rather than impacting the order book with market orders. After each partial close, the system re-assesses the remaining position: if the price recovers to a safe level, liquidation stops; if it continues to deteriorate, the cutting continues. Notably, liquidation limit orders that are not immediately filled remain on the order book as resting orders, forming a dense layer of sell orders that constitutes a resistance level to any price rebound and is gradually absorbed as sentiment deteriorates further—smoothing the liquidation impact over time while fixing it in place in price space. This approach markedly reduces the immediate impact of a single liquidation but lengthens the total liquidation window, and when the price is falling rapidly the delay can lead to larger bankruptcy losses. Binance was the first to adopt progressive liquidation comprehensively, around 2020, and OKX, Bybit, and others followed, making it the industry standard for centralized exchanges.

Another mode is the vault backstop, exemplified by Hyperliquid's HLP vault (see Chapter 9, which systematically analyzes the operating mechanism, return structure, and risk exposure of HLP as an in-protocol market-making vault). Hyperliquid's liquidation process is in fact a multi-step mechanism: when a position triggers liquidation, the liquidation engine first attempts to place the position on the order book as a limit order at a price near the bankruptcy price. Only when the order book cannot absorb the position (i.e., the position is already "underwater," with the market price worse than the bankruptcy price) do the insurance fund and then the HLP vault step in, in sequence, as successive backstops to absorb it (the full three-tier liquidation waterfall of order book $\rightarrow$ insurance fund $\rightarrow$ HLP vault is detailed in Section 9.4.3 of Chapter 9). After the JELLY manipulation event of March 2025, the protocol introduced major mechanism reforms, including a cap on the size of any single position and improved oracle-price handling logic. This tiered design channels most liquidation pressure to the public order book in normal markets and calls on the vault to absorb positions only in extreme cases, seeking a balance between minimizing market impact and preserving system solvency. In essence, however, this design creates a risk-concentration trade-off: the vault, as a single entity, absorbs liquidation pressure that should have been dispersed across the entire market, and its risk exposure rises monotonically as a one-sided market persists. Once the vault's net asset value is eroded to a critical level, it transforms from a provider of liquidity into a demander of liquidity and even becomes a new node of systemic risk (as the analysis of the JELLY manipulation event in Chapter 9 reveals, this is precisely what happens when the scale of liquidation exceeds the vault's capacity). Unlike the vault backstop, auto-deleveraging serves as the last risk buffer: when the bankruptcy losses produced by liquidation exceed the capacity of the insurance fund (the dedicated risk-reserve pool an exchange accumulates from liquidation residuals, whose capital structure, replenishment mechanism, and stress capacity are detailed in Chapter 12), the system automatically "trims" profits from the most profitable counterparty positions to cover the losses. Auto-deleveraging (ADL) ensures that the system does not incur socialized losses (in which all users share the loss) in

extreme cases, but it does so at the expense of the legitimate interests of profitable traders, provoking serious fairness disputes.

Although the design of the liquidation engine has been continually optimized, its limitations remain fully exposed in extreme conditions. The data from the "10.10 event" provide striking empirical evidence: within the 40 minutes of the cascade's eruption (20:50–21:30 UTC), the liquidation rate surged from an average of $120 million per hour over the preceding 8 hours to $10.39 billion per hour, an acceleration of 86-fold [2]. At the peak moment (21:15 UTC), $3.21 billion of positions were force-closed within a single minute, of which 93.5% were market sell orders produced by long liquidations [2].

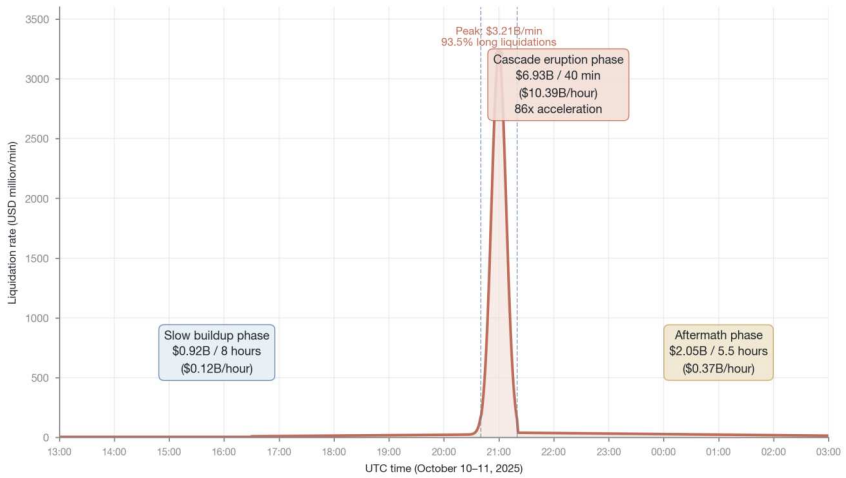


Figure 11-5. Timeline of the liquidation rate during the "10.10 event" (Data source: Amberdata [2], as of October 2025; 93.5% is on a peak-minute basis, while the 24-hour aggregate is about 87%; the roughly $9.9 billion three-phase total is Amberdata's tracked subset and is not the same basis as CoinGlass's market-wide 24-hour figure of about $19 billion—do not conflate the two)

As shown in Figure 11-5, the entire event exhibits three clearly distinguishable phases: slow accumulation, cascade eruption, and aftermath. When such a massive volume of liquidation requests floods in simultaneously, the liquidation engine itself may queue and lag because its processing capacity has reached its ceiling. This infrastructure overload means there is a discrepancy between actual execution time and theoretical trigger time, and while the price is falling rapidly, this delay often leads to larger slippage and more severe bankruptcy losses. The "3.12" event of March 2020 is a classic case of infrastructure overload amplifying risk: on March 13, BitMEX suffered a service outage of about 25 minutes (officially attributed to a DDoS attack on the platform), during which matching and liquidation nearly halted and the price subsequently rebounded, so that many positions were actually executed only at levels far below their theoretical liquidation prices before and after the outage, with bankruptcy losses far exceeding expectations [4].

11.1.4 Margin modes

The way a trader manages margin fundamentally determines the path along which risk is transmitted within the account and across the entire market. The two mainstream margin modes are isolated margin and unified margin (cross-margin). The choice between them is far from a mere tuning of technical parameters; it is an underlying institutional variable that fundamentally shapes the path of systemic-risk transmission.

Under isolated margin, the trader allocates a specific amount of margin to each position, and risk is fully isolated between positions. The maximum loss on one position is limited to the margin allocated to it and does not spill over to other funds or positions in the account. For example, if a trader allocates $5,000 of margin to a BTC long, then even if that position is fully liquidated, the funds used for ETH and SOL trading in the account are unaffected. The advantage of this mode is precise risk control and predictable losses, but its disadvantage is equally clear: low capital efficiency, because the unrealized profit of a winning position cannot be used to support other losing positions, and the trader must set aside an independent margin buffer for each position.

Under unified margin, the entire account is treated as a single liquidity pool. All available balances and the unrealized P&L of all positions serve jointly as margin, and the system decides whether to trigger liquidation based on the account's total net equity. This mode greatly improves capital efficiency: the unrealized profit of a winning position can automatically offset the margin consumption of a losing position, making long-short hedges and cross-asset strategies possible. For this reason, unified margin has been actively promoted by the major exchanges in recent years and has become the mainstream choice for professional traders.

In extreme conditions, however, unified margin exposes a fatal weakness: it creates a cross-asset, cross-position path for risk contagion.

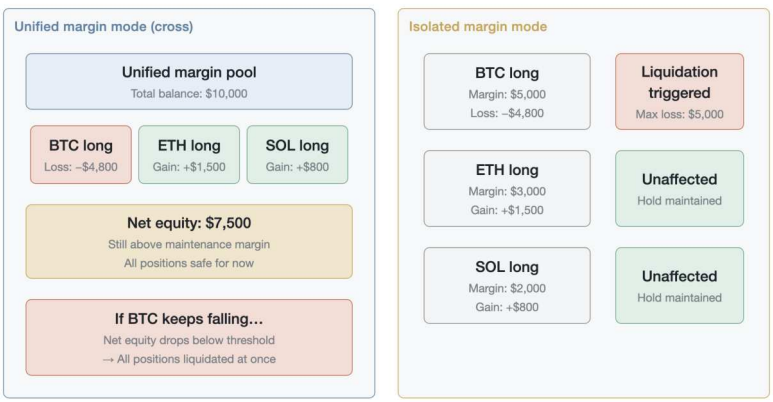


Figure 11-6. Comparison of isolated margin and unified margin under extreme conditions (conceptual illustration, not empirical data)

As shown in Figure 11-6, under the isolated mode the liquidation of the BTC position

does not affect the ETH and SOL positions, whereas under the unified mode a huge loss on BTC erodes the entire margin pool, and if net equity falls below the threshold, all positions face simultaneous liquidation. When the price of one asset in the account moves sharply and adversely, the large loss substantially reduces the net equity of the entire unified margin pool. Once total net equity falls below the total maintenance margin requirement, the system may be forced to liquidate healthy positions in the account that are themselves still profitable and entirely unrelated to the loss-making asset. This cross-asset contagion mechanism was on full display in the "10.10 event."

The most representative contagion path in the "10.10 event" came from the synthetic stablecoin USDe. Many traders used USDe as margin collateral for their unified accounts—regarded as a reasonable choice in normal markets, since it was treated as a "stablecoin-like" asset. When the collateral value of USDe plunged, however, total account net equity shrank instantly, and unified margin transmitted the discount on this single collateral asset directly into the forced liquidation of a large number of BTC and ETH longs that had nothing to do with USDe [3]. A seemingly isolated stablecoin-pricing problem thus evolved into a liquidation cascade that engulfed the entire market (the full dissection of USDe's depegging to \$0.65 on Binance during the "10.10 event"—a discount of about 35%—is given in Section 11.9.3).

Viewed from a network perspective, the unified margin pool is merely one of several contagion channels operating simultaneously. Other key contagion channels include: shared market makers withdrawing liquidity from all trading venues simultaneously once they detect adverse order flow, causing synchronized depth droughts across platforms; the cross-exchange linkage of oracle prices, which transmits a price shock on one exchange to the liquidation decisions of other exchanges via the mark-price mechanism; common collateral assets (such as USDe and USDT) being used simultaneously in the margin pools of multiple protocols, so that a discount on a single asset spreads through the collateral value chain to all protocols using that asset; and the transmission of operational risk—for example, a brief failure in one oracle's price feed for a given asset, or the temporary freeze of a cross-chain bridge, can affect all positions in a unified margin account that depend on that infrastructure. As Acemoglu, Ozdaglar, and Tahbaz-Salehi (2015) [5] argue in the theory of financial networks, network structure determines whether a shock is absorbed and buffered by the connections between nodes or amplified and propagated. When connection density exceeds a critical threshold, the network turns from a stabilizer that "disperses risk" into an amplifier that "propagates risk." In the "10.10 event," the simultaneous activation of these multiple contagion channels produced a superadditive amplification effect: the scale of liquidation and the depth of the collapse actually observed far exceeded the sum of the independent predictions for any single channel, indicating a nonlinear coupling in which the channels reinforced one another.

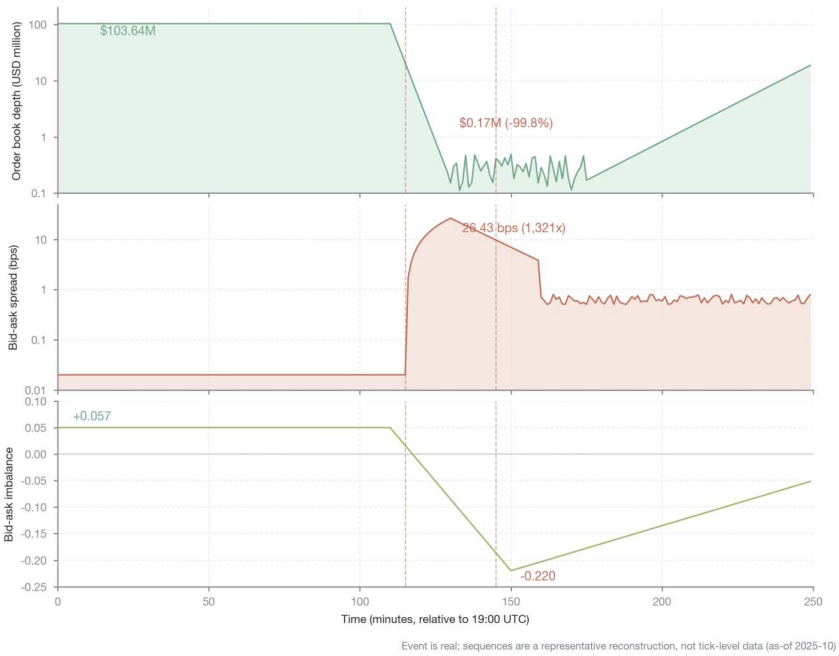


Figure 11-7. The three-dimensional collapse of liquidity during the "10.10 event" (Data source: Amberdata [2], as of October 2025; depth of about $103.6 million → $170,000 and a peak spread of 26.43 basis points are Amberdata's measured top-of-book values for BTC perpetual futures, while the three-dimensional point-by-point series is a representative reconstruction)

The essence of this procyclicality is that unified margin is a fair-weather friend: it improves capital efficiency in calm markets but turns into a dangerous liability in extreme ones. The fundamental reason is that in a true crisis the correlations among crypto assets tend to spike toward 1. As the preliminary discussion of cross-margin procyclicality in Chapter 7 (the endogenous tendency of the unified margin system to amplify risk contagion rather than provide diversification during market downturns) points out, the probability that "all positions lose money at the same time" is far higher than typical risk models assume. As shown in Figure 11-7 (the red shaded region marks the 40-minute window of the cascade's eruption), liquidity underwent a three-dimensional collapse during the "10.10 event" (see Section 11.9.2, Table 11-7 for the specific figures): order-book depth contracted by more than 99% (here depth refers to the aggregate visible limit-order depth within ±2% of the mid-price on major exchanges, with data from Amberdata, excluding iceberg orders and the dynamically refreshed liquidity of algorithmic market makers), bid-ask spreads widened by three orders of magnitude, and the buy-sell imbalance swung sharply toward overwhelming selling. This depth evaporation, spread spike, and buy-sell imbalance are the direct expression, at the order-book microstructure level, of cross-asset contagion. When all assets fall at once, all margin shrinks at once, and all liquidations trigger at once, the unified margin mode not only fails to provide a buffer but becomes a conduit for risk contagion.

The choice of margin mode therefore constitutes a deep design dilemma: the isolated mode sacrifices efficiency for safety, while the unified mode sacrifices safety for effi-

ciency. Under competitive pressure, exchanges tend to promote unified margin to attract professional traders who seek capital efficiency, but this simultaneously plants a larger procyclical risk at the system level. This dilemma receives a deeper theoretical treatment in the "impossible trinity of liquidation design" in Section 11.6.

11.1.5 Differences in cascades across market architectures

Although the underlying dynamic logic of the liquidation cascade is the same in all leveraged markets, the fundamental differences in market architecture between centralized and decentralized exchanges cause the manifestation, propagation speed, and destruction pattern of the cascade to display marked structural divergences.

In the CEX environment, the liquidation engine runs on the exchange's internal matching servers, with execution latency typically on the order of microseconds to milliseconds. CEX order-book depth is maintained continuously by professional market makers through high-frequency API connections, and liquidity supply is relatively ample in normal times. This centralized architecture, however, also means that all liquidation decisions and execution are concentrated in a single entity. When extreme conditions overload the matching engine (as in the March 2020 BitMEX event), the queuing delay in liquidation processing makes actual execution prices deviate far from trigger prices, and bankruptcy losses are sharply amplified. In addition, CEXs hold unilateral authority to adjust margin parameters and liquidation thresholds, and this centralized risk-control power can, in a crisis, act as a stabilizer (for example, by temporarily lowering the leverage ceiling) or become an additional source of uncertainty (for example, by suddenly adjusting collateral factors).

In the DEX environment, the transmission chain of liquidation is fundamentally constrained by the underlying blockchain architecture. The liquidation execution of on-chain DeFi lending protocols on Ethereum and similar chains (such as Aave and Compound) is limited by block confirmation time (typically 12 to 15 seconds), a delay that in the traditional sense constitutes a passive "speed bump." The order-book DEXs at the center of this chapter, however, are not subject to this limitation: dYdX v4 runs on a Cosmos application chain with sub-second block times, and Hyperliquid's block latency is only about 200 milliseconds. Nonetheless, the transparency of the blockchain introduces an additional risk dimension that does not exist on CEXs: the visibility of liquidation transactions in the mempool makes them a target for MEV (see the discussion of MEV and liquidation sniping in Chapter 6). A sandwich attacker can insert a front-running transaction before a liquidation transaction is confirmed, artificially widening the price slippage and thereby aggravating liquidation losses. In periods of extreme congestion, a sharp spike in gas fees further compresses the profit margin of liquidation bots, which can leave small liquidations unexecuted and let them accumulate into a larger systemic risk. The market architecture of DEX perpetual futures is not homogeneous; it comes in at least three distinct types whose liquidation dynamics differ markedly. The first type is the application-chain order-book DEX, exemplified by dYdX v4, which runs on a dedicated Cosmos application chain with block times down to the sub-second level; the execution characteristics of its liquidation engine are relatively close to those of a CEX, though it is still constrained by the blockchain's inherent

finality delay and validators' ordering power. The second type is the self-built-L1 order-book DEX, exemplified by Hyperliquid, whose block latency is about 200 milliseconds and whose execution characteristics have in practice approached CEX levels; Hyperliquid's sequencer architecture, however, is highly centralized, and its classification as a "decentralized exchange" is disputed within the industry. The third type is DEX perpetual futures based on an automated market maker or oracle quotes (such as GMX and Perpetual Protocol); these protocols genuinely lack the discrete depth structure of a traditional order book, their liquidity is distributed continuously along a price curve or an oracle price, and the price impact of a large liquidation order follows a mathematical function different from the order-book model, typically exhibiting smoother but larger total slippage. Both dYdX and Hyperliquid, the focus of this chapter, belong to the first two order-book architectures, whose cascade dynamics are structurally similar to a CEX rather than analogous to the third, automated market maker (AMM), model.

The differences between these two architectures produce distinct fragility patterns at the level of cascade dynamics. A CEX cascade tends toward a "concentrated-eruption" type, releasing a large amount of liquidation energy in an extremely short time (on the order of minutes), but it may also be abruptly cut off by the exchange's manual intervention. A DEX cascade tends toward a "diffuse-penetration" type, in which cross-protocol and cross-chain liquidations propagate gradually through oracle price updates and the delays of cross-chain bridges, with a longer time span (tens of minutes to hours) but more complex and harder-to-trace contagion paths. In the "10.10 event" of 2025, the liquidations between CEXs and on-chain DeFi were not a simple sequential process but formed a concurrent, two-way feedback loop: a price decline on CEXs triggered oracle price updates (whose delay depends on Chainlink's heartbeat interval or Pyth's push frequency), the updated oracle prices drove liquidations in DeFi lending protocols and on-chain perpetual futures, and the selling pressure from on-chain liquidations was transmitted back to the CEX spot market via arbitrageurs, further depressing CEX prices and in turn triggering a new round of oracle updates and on-chain liquidations. This CEX–DeFi two-way feedback loop ran concurrently during the event, and the oracle-update delay acted as a "clock" regulating the speed of the feedback: Chainlink's heartbeat interval and deviation threshold on Ethereum determined the lag in DeFi liquidations' response to CEX price moves, a lag that both prevented, to some degree, the instantaneous transmission of shocks and led to a temporal dispersion of liquidations, so that on-chain aftershocks persisted for hours after the main CEX shock had ended.

We have now fully dissected the micro transmission chain from a price move to forced liquidation: the margin system defines the boundaries of leverage and the measure of fragility, the mark price constructs the "truth" benchmark that triggers liquidation, the liquidation engine executes forced closure in a deterministic manner, and the margin mode determines the path along which risk is transmitted within the account and across assets. Each link in this chain is a carefully designed risk-management tool in normal markets; but when they are activated simultaneously and coupled with one another in extreme conditions, they jointly weave a precise network that converts individual risk into systemic risk. This calls for the two-layer theoretical framework of reflexivity, which provides both a philosophical and an economic lens for understanding how this

network generates self-reinforcing positive feedback.

11.2 The theoretical foundations of reflexivity

Before we can understand the endogenous reflexivity of leverage and liquidation in crypto-asset markets, it is worth tracing its theoretical roots. Most traditional financial theory rests on the efficient market hypothesis (EMH) and the rational-agent assumption, holding that prices passively reflect fundamental information and tend toward some equilibrium. In the highly leveraged and automated crypto derivatives market, however, price movements often no longer merely reflect fundamentals but become the core force driving changes in those fundamentals (such as liquidity and participants' capital), which in turn feed back on prices themselves. This dynamic process is precisely the essence that the theory of reflexivity reveals. This section starts from George Soros's philosophical epistemology and, combining it with the microstructure model of Markus Brunnermeier and Lasse Heje Pedersen, builds a two-layer theoretical framework for understanding liquidation cascades in crypto markets.

11.2.1 Soros's theory of reflexivity

George Soros first systematically advanced the theory of reflexivity in *The Alchemy of Finance* [6]. Unlike traditional economics, which assumes that supply and demand curves are independent and jointly determine price, Soros held that in social systems involving thinking participants there is a two-way feedback mechanism between participants' perceptions and the actual state of affairs. Soros decomposed this feedback mechanism into two basic functions: the cognitive function and the participating function. In the cognitive function, participants try to understand their environment, and information flows from the world to thought; in the participating function, participants act on their understanding in an attempt to change the environment, and information flows from thought to the world [7].

In the traditional equilibrium framework, these two functions are treated as independent: price merely reflects reality passively. Under the reflexivity framework, however, the two functions are intertwined and mutually causal. When the market price changes (reality changes), it affects traders' risk perception and sentiment through the cognitive function (perception changes); traders adjust their positions on the basis of the new perception (for example, by actively selling or adding margin), and this behavior is converted, through the participating function, into actual buying and selling pressure in the market (action changes reality); these trades in turn push the price further in some direction, closing the loop.

In perpetual futures markets, this reflexivity framework can be formalized as a set of coupled difference equations. Let C_t denote the market's collective cognitive state at time t (which can be operationalized as an observable variable such as a market sentiment index, the funding rate, or the long-short ratio), P_t the asset price, and $Q_{\text{liq},t}$ the liquidation order flow at time t. The cognitive function maps price changes into perception updates:

$$C_{t+1} = f(P_t, C_t) = C_t + \gamma \cdot \Delta P_t + \eta_t$$

where $\gamma > 0$ denotes the strength of price changes' influence on perception (in crypto markets, given the immediacy of information diffusion, γ is markedly higher than in traditional markets), and η_t is an exogenous information shock. The participating function maps the cognitive state into price changes, but in perpetual futures markets it must incorporate the mechanical, forced execution of the liquidation engine:

$$P_{t+1} = g(C_t, Q_{\text{liq},t}) = P_t + \lambda \cdot h(C_t) - \beta \cdot \frac{Q_{\text{liq},t}}{D_t}$$

where $h(C_t)$ denotes the influence on price of perception-driven voluntary trading, λ is the market-impact coefficient of the participating function, and $\beta \cdot Q_{\text{liq},t}/D_t$ is the mechanical price impact of the liquidation engine—a rigid participating function specific to crypto markets that does not exist in the traditional Soros framework. A note on dimensional conventions: in this equation P_{t+1} and P_t are price levels (in dollars), so β and λ here carry price dimensions (dollars). They describe the same impact mechanism as the identically named dimensionless coefficients β and λ that act on the dimensionless log return ΔP in Section 11.3.3, and they merely adopt different scalings because the dependent variable is defined differently (price level versus log return); the two should not be equated numerically. When the system is in a steady state ($\Delta P_t \approx 0$, $Q_{\text{liq},t} \approx 0$), the two equations converge to equilibrium; when a price decline triggers liquidation ($Q_{\text{liq},t} > 0$), the liquidation impact depresses the price further, the price decline in turn updates perception and triggers more liquidations, and the coupling of the two equations drives the system along a divergent path.

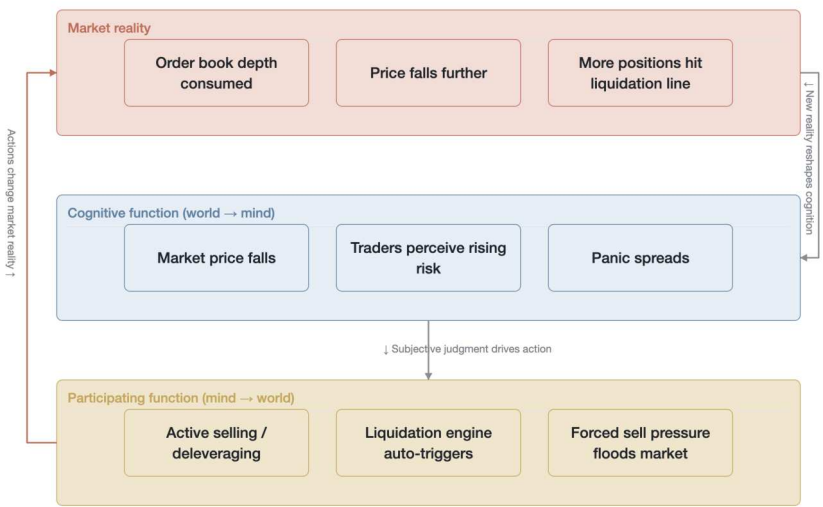


Figure 11-8. The reflexivity feedback loop of leverage and liquidation (based on the Soros framework)

As shown in Figure 11-8, in the crypto derivatives market this reflexivity is greatly amplified by the automated liquidation engine. Acting as a hard-coded "participating function," the liquidation engine forcibly converts under-collateralized positions into market orders, directly consuming order-book depth and depressing the price, shaping a harsher market reality and pushing the system away from equilibrium into a

self-reinforcing divergent state. The core difference the figure reveals is this: the "participating function" of a traditional market must be mediated by human judgment (assessing risk, deciding to cut positions, choosing timing), and these steps introduce time delays and behavioral heterogeneity that constitute a natural damping of positive feedback; in crypto markets, by contrast, the participating function is hard-coded directly into the smart-contract logic of the liquidation engine, and its execution is deterministic, instantaneous, and unconditional, with no buffer space for human intervention. This positive-feedback loop therefore has no natural attenuation mechanism, and each cycle may release greater selling pressure than the last, so that the system, absent external intervention, keeps evolving toward divergence.

11.2.2 The liquidity spiral model

If Soros's theory of reflexivity provides a macro philosophical framework, the model of Brunnermeier and Pedersen lays bare the mutually reinforcing relationship between an asset's market liquidity (how easily the asset can be converted to cash) and a trader's funding liquidity (how easily the trader can obtain funds) [8]. They show that when an external shock hits, these two forms of liquidity form a vicious circle through two mechanisms: the loss spiral and the margin spiral.

The loss spiral describes an equity effect. When an asset's price falls, leveraged traders holding a long position in it suffer a loss of equity. To maintain a given leverage ratio or meet regulatory requirements, they are forced to sell part of the asset. This selling impacts the market price and drives the price down further, triggering fresh equity losses and another round of selling. Consider a concrete transmission process: when a long trader using 10x leverage holds a \$100,000 BTC position, a 5% price decline shrinks their \$10,000 of equity to \$5,000, and their actual leverage jumps from 10x to 20x; to restore leverage to its target level, the trader must sell about \$50,000 of the position, and this forced sale itself exerts downward pressure on the market price, subjecting other leveraged traders to the same predicament.

The margin spiral describes a collateral-constraint effect. When a price decline is accompanied by rising volatility, funding providers (such as brokers and exchanges) raise margin requirements to control risk. This means the trader's existing collateral can borrow less than before. To make up the funding gap, the trader must sell more assets, which not only depresses the price further but may also trigger greater volatility, leading margin requirements to be raised again. In crypto markets this mechanism is especially aggressive: during high-volatility periods, exchanges dynamically raise maintenance margin rates or lower collateral factors, and such adjustments often affect a large number of accounts simultaneously within a very short time, creating collective forced-deleveraging pressure driven by the change in risk-control rules itself.

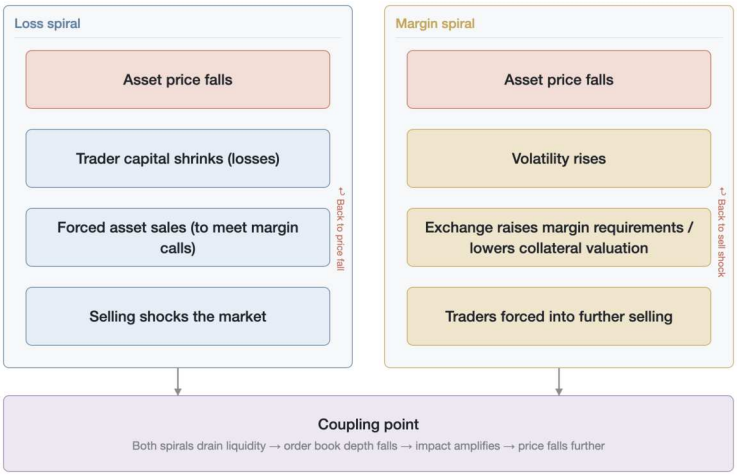


Figure 11-9. The coupling mechanism of the loss spiral and the margin spiral

As shown in Figure 11-9, these two spirals often occur simultaneously and couple with each other during a crisis: a price decline both directly cuts capital (the loss spiral) and raises margin requirements by pushing volatility higher (the margin spiral), and together they generate enormous forced-selling pressure that rapidly consumes order-book depth. The model of Brunnermeier and Pedersen proves mathematically that, under certain conditions, the margin mechanism itself is a source of market instability. The key is this: margin adjustments in traditional markets settle daily and involve manual review, so the two spirals often unfold sequentially and can be separated in time, providing a window for identification and intervention; in crypto markets, margin is computed in real time, so the two are activated synchronously on the same time scale and amplify each other. The mathematical consequence of this synchrony is that the joint effect is not additive but multiplicative—the selling pressure produced by the loss spiral lowers market liquidity, and lower liquidity makes each unit of selling in the margin spiral produce a more violent price impact, and vice versa.

11.2.3 The applicability to perpetual futures

Perpetual futures in crypto-asset markets constitute an extreme testing ground for the two-layer theory above. Compared with traditional financial markets, the crypto derivatives market pushes reflexivity to its limit through its mechanism design.

Table 11-2 compares the core differences in liquidation mechanisms between traditional finance and crypto perpetual futures:

Dimension	Traditional finance (e.g., securities margin trading, traditional futures)	Crypto perpetual futures (e.g., Binance, Bybit)
Margin assessment	Brokers/clearinghouses settle daily (T+0 or T+1), with manual review	Smart contract/liquidation engine monitors in real time (millisecond level)
Grace period	A margin call is usually issued, allowing top-up within T+1 to T+3	Zero grace period; forced closure the instant the maintenance margin line is touched
Liquidation execution	Traders close manually or brokers execute in batches, considering market impact	The engine takes over automatically, typically dumping to market via market orders or better-than-market limit orders
Lender of last resort	The central bank can inject liquidity through the discount window, breaking the spiral	No lender of last resort; reliance only on a limited insurance fund or auto-deleveraging
Spiral evolution speed	Days to weeks (e.g., the 1998 LTCM crisis, the 2008 subprime crisis)	Seconds to minutes (e.g., 3.12 in 2020, 5.19 in 2021, 10.10 in 2025)

Table 11-2. Comparison of liquidation mechanisms between traditional finance and crypto perpetual futures (Data source: compiled by the author)

Notably, the table summarizes the spiral-evolution speed of traditional finance as "days to weeks," which applies mainly to the traditional deleveraging process that relies on human decision-making and daily settlement; when traditional markets likewise introduce highly automated execution, the phenomenon of time compression is not unique to crypto. The U.S. stock market "flash crash" of May 6, 2010, is a case in point: the Dow Jones Industrial Average plunged nearly 1,000 points (about 9%) in roughly 36 minutes and then rebounded over roughly the same span, and its underlying dynamics are highly isomorphic with a crypto liquidation cascade—automated algorithmic execution replacing human judgment, liquidity providers systematically withdrawing quotes as volatility intensified, and positive feedback forming between selling pressure and the liquidity vacuum. The exchange-traded fund (ETF) flash crash of August 24, 2015, further confirmed this pattern, when the prices of hundreds of ETFs deviated from net asset value by 20% to 30% within minutes of the open. The key difference, however, lies in the interruption mechanism: the limit up-limit down (LULD) rule introduced in the United States after the 2010 flash crash, together with earlier circuit-breaker rules such as Rule 48, can halt trading and forcibly break the spiral when a price deviates beyond a threshold, and this external "time wedge" provides a window for liquidity to recover and information to be digested; the crypto perpetual futures market, to this day, lacks any equivalent market-wide circuit breaker, trading proceeds around the clock, the liquidation engine runs continuously, and once positive feedback begins there is no external force to arrest it.

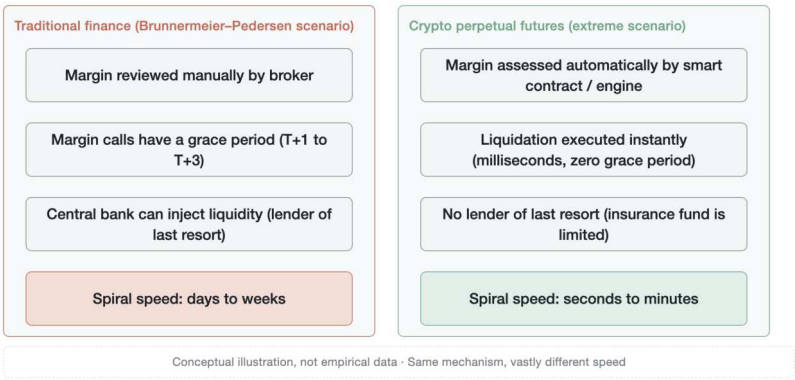


Figure 11-10. Comparison of spiral speed between traditional finance and crypto perpetual futures (conceptual illustration, not empirical data)

As Table 11-2 and Figure 11-10 show, in crypto markets the "loss spiral" and "margin spiral" of the Brunnermeier-Pedersen model are compressed to the millisecond level. Without the buffer of human intervention or the backstop of a lender of last resort, the feedback loop between price declines and liquidation selling has no damping at all. This homogenization and automation of the mechanism make the "participating function" Soros described entirely deterministic and mechanical. When the market is in a highly leveraged state, a tiny price disturbance can topple the first domino and evolve into a systemic liquidation cascade.

Historical data clearly confirm this. Figure 11-11 shows the ten largest liquidation events in crypto history, ranked by total 24-hour liquidations.

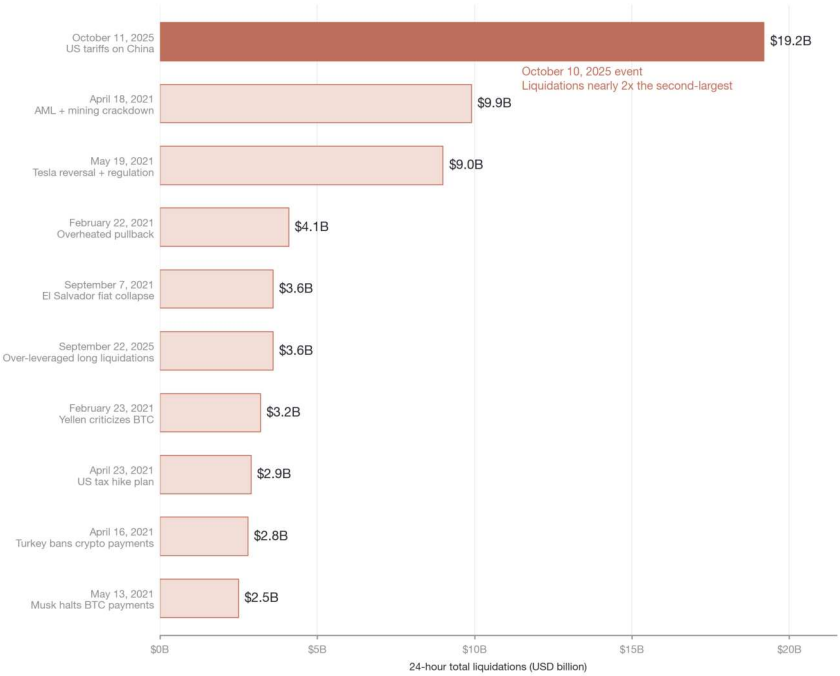


Figure 11-11. Major liquidation events in crypto history (Data source: CoinGlass [9], as of October 2025; the top 10 ranked by 24-hour liquidations; because of exchanges' API reporting caps, the CoinGlass basis historically underestimates and its absolute figures are conservative, so orders of magnitude should be emphasized)

As shown in Figure 11-11, the liquidation event of October 10–11, 2025, set a historical record, with total single-day liquidations of about $19.1 billion (on CoinGlass's exchange basis; constrained by a rate limit that pushes only one liquidation data point per second, this value is a lower bound, and the true notional scale may be higher), nearly double the second-ranked event (April 18, 2021) [9]. In this crash triggered by macro news, the microstructure signatures—Bitcoin's sharp plunge, more than 1.62 million accounts liquidated (about 87% of them long), order-book depth evaporating by more than 99% within minutes, and spreads widening by more than a thousandfold (see the chapter opening and Section 11.9.2 for data) [1]—are precisely the result of reflexivity being amplified to its extreme under the automated liquidation mechanism.

The endogenous reflexivity framework above describes the transmission mechanism of the liquidation cascade, but the system's baseline fragility is modulated to a great extent by macro liquidity conditions. The high open interest and extreme funding rates that constitute the cascade's "fuel" tend to accumulate preferentially in a risk-on macro environment—periods of a weakening dollar index, an accommodative Federal Reserve monetary policy, and narrowing global credit spreads. In such easing cycles, abundant external liquidity not only fuels the swelling of leverage demand but also temporarily sustains higher order-book depth by attracting incremental capital to make markets, thereby suppressing the triggering and propagation of cascades. Conversely, when the macro environment shifts from risk-on to risk-off (for example, a sudden spike in expectations of Fed tightening, or a geopolitical shock that triggers a global flight to

safety), trigger sensitivity rises markedly, because the contraction of external liquidity simultaneously weakens market makers' risk-bearing capacity and arbitrageurs' capital supply. This means that in a macro-easing period, even when market microstructure indicators (such as the OI/market-cap ratio and the funding-rate level) have reached levels comparable to those of historical crash events, the cascades that actually occur tend to be fewer and milder; whereas in a period of macro tightening, even a lower degree of microstructure fragility can brew a severe cascade. A complete cascade-risk assessment model therefore cannot rely on endogenous market-structure indicators alone but must be calibrated conditional on macro state variables. The subsequent analysis in this chapter focuses on the endogenous transmission mechanism, but readers should remain aware throughout of this implicit premise of macro conditionality.

11.3 The liquidation reflexivity equation

The liquidation cascade is not an accidental event. Under specific market microstructure conditions, the occurrence of a cascade is a mathematical inevitability. This section formalizes the positive-feedback process of liquidation into a concise difference equation, revealing the critical condition that drives a cascade out of control.

11.3.1 Liquidation order flow

In market microstructure theory (e.g., Kyle, 1985 [10]), price changes are driven by order flow. Ordinary order flow is random and involves information asymmetry, and its direction is two-sided. Liquidation order flow, however, is a distinctive kind of passive order flow. Its direction is fully determined: when a price decline triggers long liquidations, the liquidation engine necessarily sends sell orders to the market; when a price rise triggers short liquidations, it necessarily sends buy orders. This one-sidedness means that liquidation orders cannot be hedged internally and must be absorbed entirely by market makers or contrarian traders. On top of this, they form a liquidity devourer that is extremely insensitive to price. The liquidation engine's primary task is to close positions as quickly as possible to avoid the systemic risk (socialized losses) caused by bankruptcy. It therefore typically uses market orders or limit orders with an extremely high slippage tolerance. This execution logic means that liquidation orders consume order-book liquidity regardless of cost and constitute a pure demander of liquidity.

11.3.2 The three-step mapping

We can decompose the liquidation-cascade process into three successive mapping steps. The trigger mapping describes how a price change ΔP_t drives the margin balance of some positions below the maintenance margin level and thereby triggers liquidation. Next comes the quantity mapping, in which the triggered positions are converted into actual liquidation order flow Q_{liq}. Building on this, the impact mapping shows how the liquidation order flow Q_{liq} hits the current order-book depth D_t and produces a new price change ΔP_{t+1}.

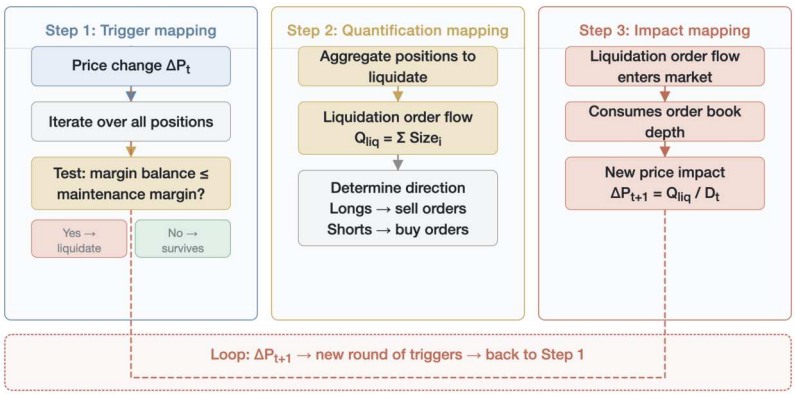


Figure 11-12. The three-step mapping from a price change to the next round's price change

As shown in Figure 11-12, if ΔP_{t+1} has the same direction as ΔP_t and is large enough to trigger new liquidations, this loop continues. The figure decomposes the loop into three mapping steps with clear economic meaning. The first step is the trigger mapping, whose core question is which positions' margin balances will fall below the maintenance margin line under a given price decline and thus be captured by the liquidation engine; the output of this step depends on the distribution density of leverage multiples and the degree of clustering of liquidation prices. The second step is the quantity mapping, which converts the triggered positions into actual market order flow, whose scale is determined jointly by the notional value of the liquidated positions and the liquidation engine's execution strategy. The third step is the impact mapping, which describes the actual effect these liquidation order flows have on the price once they enter the order book—an effect that depends not only on the absolute scale of the order flow but, more importantly, on the effective depth of the order book at the time. The loop arrow in the figure represents precisely the self-reinforcing nature of this process: the new price change ΔP_{t+1} produced in the third step constitutes the input to the first step of the next round, that is, the new trigger condition. As long as the price change output by each round is large enough to activate new liquidation positions, this three-step mapping keeps iterating.

11.3.3 The divergence condition

To formalize this process, we define the following variables. ΔP_t denotes the magnitude of the price change in round t (taken as a dimensionless log return, which for small moves is approximately equal to a percentage). α represents liquidation sensitivity, the proportion of leveraged positions that a unit price change can trigger for liquidation, which depends on the distribution density of leverage in the market. L is the total scale of leveraged positions in the market (such as open interest; note that from this section onward L refers specifically to the leverage stock/notional, distinct from the $P^{\text{liq}}(t)$ that denotes the liquidation price in Section 11.1.1, and the two must not be confused). β is the price-impact coefficient. D_t is the effective order-book depth in round t. Here D_t should be understood as the net effective depth after cross-exchange integration—that

is, the system-level depth net of the friction costs arbitrageurs incur to move liquidity across platforms—rather than the order-book depth of a single exchange. During a cascade, because arbitrageurs cannot redistribute order flow across exchanges quickly enough, the system's overall effective depth is markedly smaller than the simple sum of the individual exchanges' order-book depths. To ensure the dimensionlessness of the feedback gain, the dimensions of the variables are stipulated as follows: ΔP_t is taken as a dimensionless log return; L and D_t are both measured in dollar notional; α is defined as "the proportion of liquidatable notional triggered per unit log-price change, as a share of total notional" (with dimensions of 1/log return); and β is the impact elasticity, "the log-price change produced per unit dimensionless order-flow/depth ratio" (dimensionless). On this basis, $Q_{\text{liq}} = \alpha L \Delta P$ is measured in dollars, Q_{liq}/D_t is dimensionless, and $G = \alpha\beta L/D_t$ is a dimensionless quantity, so that comparison with the critical value of 1 is meaningful.

Based on the three-step mapping above, we can derive the liquidation volume in round t, $Q_{\text{liq},t}$, as:

$$Q_{\text{liq},t} = \alpha \cdot L \cdot \Delta P_t$$

Once this liquidation volume enters the market, the new price impact it produces, ΔP_{t+1}, is:

$$\Delta P_{t+1} = \beta \cdot \frac{Q_{\text{liq},t}}{D_t}$$

Substituting the first equation into the second, we obtain the liquidation reflexivity equation:

$$\Delta P_{t+1} = \left(\alpha \cdot \beta \cdot \frac{L}{D_t}\right) \cdot \Delta P_t$$

Letting the feedback gain be $G = \alpha \cdot \beta \cdot \frac{L}{D_t}$, the equation simplifies to:

$$\Delta P_{t+1} = G \cdot \Delta P_t$$

This concise difference equation reveals that the fate of the cascade depends entirely on the magnitude of the feedback gain G. Under the convergence condition ($G < 1$), the new price impact produced in each round is smaller than the last, the system gradually absorbs the shock, and the price stabilizes after some fluctuation. Under the divergence condition ($G > 1$), the new price impact produced in each round is larger than the last, the shock is continually amplified, and the system enters an out-of-control cascade-collapse state until leverage is fully cleared (L falls sharply) or external liquidity intervenes (D_t rises sharply).

The reflexivity equation above with a constant G, however, produces unbounded exponential divergence, which is physically unrealistic, since a price cannot fall below zero and the leverage pool L is continually depleted as liquidation proceeds. A more complete model must endogenize G: $G_t = \alpha \cdot \beta \cdot L_t/D_t$, where L_t decreases with each round of liquidation, that is, $L_{t+1} = L_t - Q_{\text{liq},t}$. This correction transforms the system from a linear divergence process into a nonlinear difference equation: in the initial phase, because L_t is still ample relative to $Q_{\text{liq},t}$, $G_t > 1$ and the system exhibits divergence; but as liquidation continually consumes the leverage stock, G_t declines round by round, eventually crossing the critical line $G_t = 1$ and decaying into the conver-

gence range. In other words, a liquidation cascade is essentially an initially divergent but conditionally self-terminating dynamic process. The recursion $L_{t+1} = L_t - Q_{\text{liq},t}$ above characterizes only the decay of the numerator (the leverage stock), whereas the denominator D_t also plunges endogenously during a cascade (see Section 11.3.4); only when the depletion rate of the leverage stock ultimately exceeds the deterioration rate of the effective depth does G_t decline monotonically, cross 1, and self-terminate. If depth collapses very fast, G_t may rise before it falls, and the system experiences a more violent overshoot before self-terminating. The total overshoot of the price depends on how many rounds of iteration the system passes through before G_t crosses 1, and this number of iterations is determined jointly by the initial leverage concentration and the depth-to-liquidation-volume ratio. The larger the initial G_0 and the more concentrated the leverage distribution within a narrow price range, the more divergent iterations the system passes through before self-terminating, and the more severe the cumulative destruction. This self-termination mechanism connects directly to the systematic analysis of cascade-termination conditions in Section 11.5.4.

In addition, it should be noted that the linear price-impact model $\beta \cdot Q/D$ in the impact mapping above is itself a first-order approximation. The empirical market microstructure literature shows that the price impact of normal order flow follows a square-root law (Bouchaud et al., 2009), that is, price impact is proportional to the square root of order-flow size. But liquidation flow is not normal order flow: it is concentrated, one-sided, and price-insensitive, and it often penetrates multiple order-book price levels at once, so its actual impact may exhibit superlinear (convex) characteristics. This means that in a liquidation-dense range, the leverage density required to reach the divergence threshold $G > 1$ may be lower than the linear model predicts, so the system's actual fragility is underestimated by the linear framework.

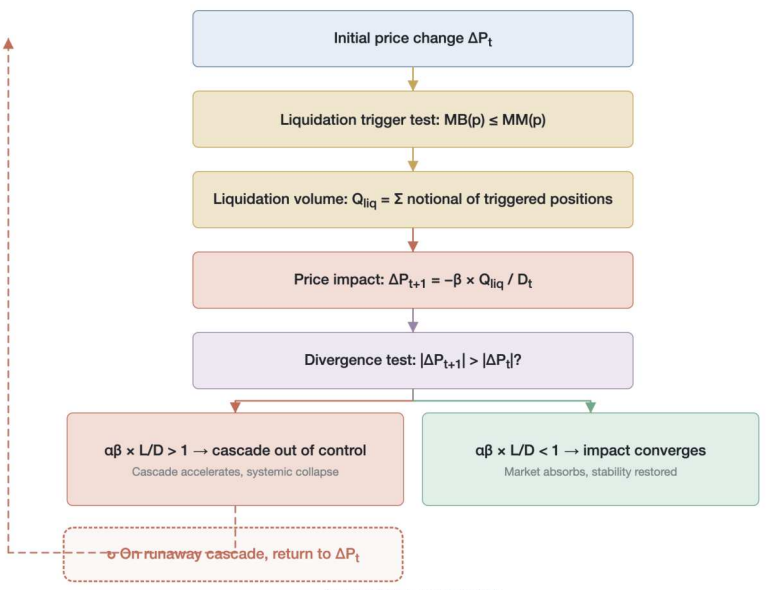


Figure 11-13. The positive-feedback loop and divergence condition of the reflexivity equation (conceptual illustration, not empirical data)

As shown in Figure 11-13, the figure's core is to reveal a key bifurcation structure: when $G < 1$, the price impact of each round of liquidation attenuates round by round, the initial disturbance is absorbed after a finite number of iterations, and the system behaves as a stable negative-feedback equilibrium; when $G > 1$, each round's shock is larger than the last, the price sequence diverges exponentially, and the system enters a self-reinforcing cascade-amplification path (irreversible by a small disturbance before leverage is cleared, but not never-terminating; see the conditional self-termination analysis in Section 11.3.3). Every component of $G = \alpha \cdot \beta \cdot L/D_t$ is a function of market conditions: in a bull market the continual accumulation of leveraged positions increases L and α together, while rising volatility prompts market makers to withdraw liquidity and causes D_t to contract sharply, so the same market can transition smoothly from the $G < 1$ steady state to the $G > 1$ divergence region as leverage accumulates and liquidity ebbs. This explains why a liquidation cascade always appears to erupt "suddenly" to an outside observer: the system may already have been operating on the edge of danger for a considerable time before crossing $G = 1$, but when G is slightly below 1 it still displays apparent stability, and fragility is almost unobservable before the threshold is breached.

11.3.4 Three sources of nonlinearity

In real crypto markets, the variables in the liquidation reflexivity equation are not linear. The equation above assumes that liquidation sensitivity α and order-book depth D_t are constants, but this simplification is severely distorted in actual markets, because the threshold-clustering effect of liquidation prices, the high homogeneity of trader behavior, and the endogenous withdrawal of liquidity by market makers in times of crisis together create multiple discontinuities and positive-feedback acceleration mechanisms. It is precisely the following three sources of nonlinearity that give liquidation cascades their characteristic sudden eruption and highly destructive nature. The research of Danielsson, Shin, and Zigrand shows that the procyclicality of leverage is itself a core driver of endogenous risk [11].

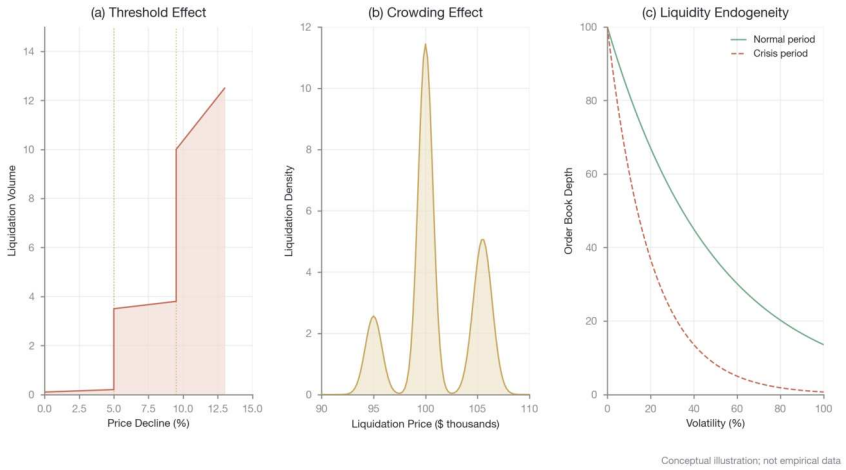


Figure 11-14. The three sources of nonlinearity in a liquidation cascade (conceptual illustration; qualitative axes; not empirical data)

As shown in Figure 11-14, the figure characterizes the three sources of nonlinearity in liquidation reflexivity. The first is the threshold effect: liquidation does not occur smoothly as the price falls; every leveraged position has a precise maintenance margin rate (such as 0.5%), the position exerts no selling pressure on the market before the price reaches the threshold (a latent state), and once the threshold is breached, the entire notional value of the position is instantly converted into a market sell order. This discontinuous jump from "zero" to "a huge amount" can drive the derivative of liquidation volume with respect to price (α) toward infinity in a specific range. The second is the crowding effect: retail traders and quantitative strategies tend to favor integer leverage multiples (such as 10x, 20x, 50x) and to open positions near key round-number price levels or technical support levels, and this has multiple psychological microfoundations—the anchoring effect drives traders to use integer leverage and round-number price levels as reference points; the coordination effect of real-time social media (the "key levels" shared on Crypto Twitter and in Telegram groups) amplifies herd behavior; the informational-cascade mechanism (Bikhchandani, Hirshleifer, and Welch, 1992) [12] leads later entrants to treat peers' opening of positions at the same price as evidence of price support; and the representativeness heuristic ("this support level held in the past, so it will hold again") further entrenches an excessive reliance on historical levels. Because position information in traditional markets is private and social amplification is far less immediate and transparent than in crypto communities, both behavioral homogeneity and the clustering density of liquidation prices are markedly lower there. This clustering forms "liquidation-dense zones," and when the price falls into one, α spikes sharply and generates a massive liquidation order flow instantly. The third is the endogeneity of liquidity: order-book depth D_t is not an exogenous constant; in normal times market makers provide ample liquidity, but when prices swing violently and liquidations occur frequently, market makers face enormous adverse-selection and inventory risk and rationally choose to widen spreads or withdraw their orders outright, so that effective depth D_t falls sharply precisely when the system most needs liquidity to absorb selling pressure, and this "procyclicality" further

amplifies the price-impact coefficient β. The multiplicative superposition of the three nonlinearities (a spike in $\alpha \times$ a plunge in D_t) allows the feedback gain G to jump from below 1 to far above 1 in an extremely short time, driving the market out of control in an instant.

This nonlinear mechanism also answers the paradox posed at the opening of this chapter—why individual prudence produces precisely systemic fragility at the aggregate level. The causal bridge has three parts. First, the stop-loss orders that individuals set out of prudence, like leverage liquidation prices, tend to cluster near round-number levels and technical support levels, and the two superimpose on the price axis, jointly raising α in the liquidation-dense zone. Second, under unified margin and correlations approaching 1 in a crisis, the effective portfolio-level leverage of a "safe" 5x–10x notional leverage is far higher than its notional value (see Section 11.7.5), so that prudent individuals contribute far more liquidation sensitivity at the system level than they realize. Third, the voluntary sell orders triggered by individual stop-losses have exactly the same direction as the forced sell orders of the liquidation engine and are released simultaneously, which in effect further amplifies α. Thus it is precisely the risk-management actions of countless individuals—microscopically "correct"—that, under the structural coupling of unified margin, liquidation-price clustering, and procyclical liquidity withdrawal, aggregate into a macroscopically high α, high liquidation concentration, and low D_t, causing systemic risk to emerge endogenously from individual prudence.

11.3.5 The cascade risk index

To quantify this divergence risk in actual trading and risk control, we can construct an operational indicator based on the reflexivity equation: the cascade risk index (CRI).

$$\mathrm{CRI} = \frac{\text{leverage density} \times \text{liquidation concentration}}{\text{available depth}} \approx \frac{(\text{OI/market cap}) \times \text{liquidation-heatmap density}}{\text{stress depth}}$$

The precise definitions of the sub-components are as follows. Leverage density is measured by the ratio of open interest (OI) to the asset's circulating market capitalization, that is, leverage density = OI/market cap. When this ratio exceeds 0.3, it indicates that the derivatives market's notional exposure is approaching one-third of the underlying asset's market cap and that the leverage level has entered a warning range. Liquidation concentration uses the liquidation heatmap to quantify the share of liquidatable notional value clustered within a specific range below the current price (usually taken as 5%) as a proportion of total open interest, that is, liquidation concentration = $\sum_i N_i \cdot \mathbf{1}[P_i^{\mathrm{liq}} \in (P \times 0.95,\ P)]/\mathrm{OI}$ (where N_i is the notional value of position i, P_i^{liq} is its liquidation price, P is the current mark price, and $\mathbf{1}[\cdot]$ is the indicator function). This indicator reflects the relative scale of the liquidation order flow that might be triggered by a 5% price decline. Available depth cannot be read from the current static snapshot of the order book alone; it should instead be assessed as the "stress depth" that market makers can maintain under a specific volatility-stress scenario—that is, the effective bid depth the order book can still provide within 2% below the price when volatility spikes to its historical 99th-percentile level.

Based on the definitions above, we can retrospectively calibrate the CRI for four major liquidation events in crypto history, providing an empirical calibration benchmark for

the indicator's warning thresholds.

Event	OI/market cap	Liquidation concentration (5%)	Stress depth ($100 million)	Estimated CRI	Actual cascade severity
March 12, 2020, "Black Thursday"	0.18	0.12	0.8	0.027	Severe (~$6 billion liquidated)
May 19, 2021, "5.19 crash"	0.25	0.15	1.2	0.031	Severe (~$8.6 billion liquidated)
August 5, 2024, "yen carry-trade unwind"	0.28	0.18	1.5	0.034	Moderate (~$10.2 billion liquidated, rapid recovery)
October 10, 2025, "10.10 event"	0.35	0.22	0.4	0.193	Extreme (~$19 billion liquidated)

Table 11-3. Retrospective CRI calibration for major historical liquidation events (Data source: author's estimates, derived from public data)

The table above reveals two key patterns. First, the estimated CRI of the "10.10 event" is an order of magnitude higher than in the three preceding events, and its core driver was not a proportionate rise in leverage density or liquidation concentration but a sharp contraction of stress depth, from $150 million in August 2024 to just $40 million—meaning the market lacked sufficient liquidity buffer to absorb selling pressure under the same liquidation shock. Second, the absolute value of the CRI must be read in light of the historical evolution of market size: as the total volume of the crypto derivatives market grows, absolute liquidation volume is markedly amplified even when the change in a relative indicator (such as OI/market cap) is limited. The CRI's warning thresholds should therefore be dynamically calibrated to market size.

As a risk indicator, however, the CRI has several limitations that warrant caution. First, on dimensional consistency: the numerator of the CRI is the product of two dimensionless ratios, while the denominator is stress depth priced in absolute dollars, which means the CRI's value changes mechanically as market volume grows and lacks an intrinsic standardization mechanism comparable across periods. Second, Table 11-3 contains only four data points spanning five years, a sample too small for robust statistical inference or for constructing warning thresholds with confidence intervals; this is essentially a retrospectively calibrated descriptive analysis, not a cross-validated predictive model. Third, the event of August 5, 2024, constitutes an anomaly worth examining closely: its estimated CRI (0.034) is higher than that of May 19, 2021 (0.031), yet its actual cascade severity was clearly lower and the market recovered markedly faster. This discrepancy suggests that the differences in the speed of external intervention and in the macro transmission path that the CRI fails to capture (the yen carry-trade unwind had a clear exogenous source and was self-limiting) may be more decisive than

microstructure fragility itself in determining a cascade's ultimate destructiveness. On balance, the CRI should be improved in the direction of a conditional indicator—that is, by introducing macro state variables (such as the dollar index and credit spreads discussed in Section 11.2.3) as modulating factors on top of the microstructure variables, so that the same CRI value maps to different risk probabilities under different macro environments.

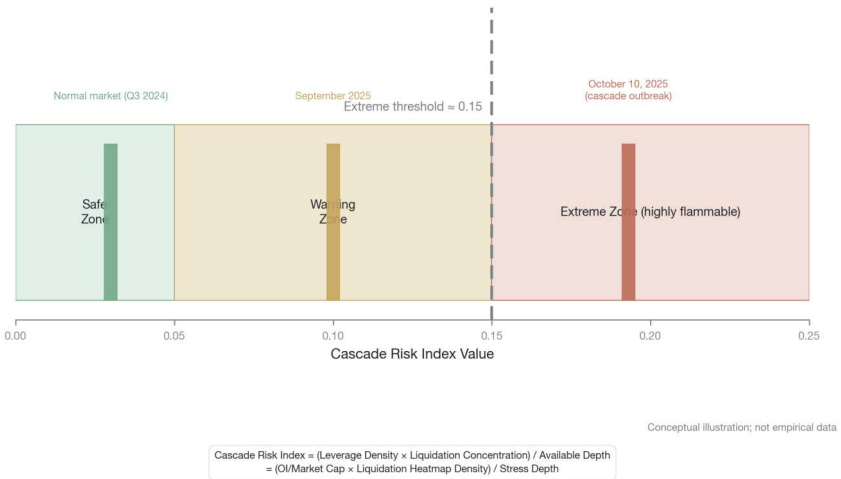


Figure 11-15. Visualization of the cascade risk index (the CRI is an exploratory indicator constructed by the author, based only on retrospective calibration of four historical events and not dimensionally standardized; the threshold bands in the figure are a conceptual illustration, not cross-validated, operational risk-control thresholds)

As shown in Figure 11-15, the CRI can serve as an exploratory, descriptive indicator for monitoring market fragility—though it must be stressed that it is a crowding proxy inspired by the feedback gain G, not yet dimensionally standardized, not the same construct as G, and not sharing the same critical value. Based on the retrospective calibration of four historical events in Table 11-3 (with estimated CRI values falling in the range of about 0.027–0.193), an empirical relative banding can be offered: in the safe range (about < 0.05) the market has enough resilience to absorb local shocks; the warning range (about 0.05–0.15) indicates that leverage is overcrowded and the market is in a subhealthy state, as when the funding rate became extreme in September 2025; and in the extreme range (about > 0.15, into which the October 10, 2025, event's calibration of about 0.193 falls), the system is highly flammable, and any minor negative catalyst can ignite a systemic cascade. Note that this banding is based only on the retrospective, descriptive calibration of four data points, that its warning thresholds and predictive power await large-sample validation, and that it should not be treated as a cross-validated, operational risk-control threshold.

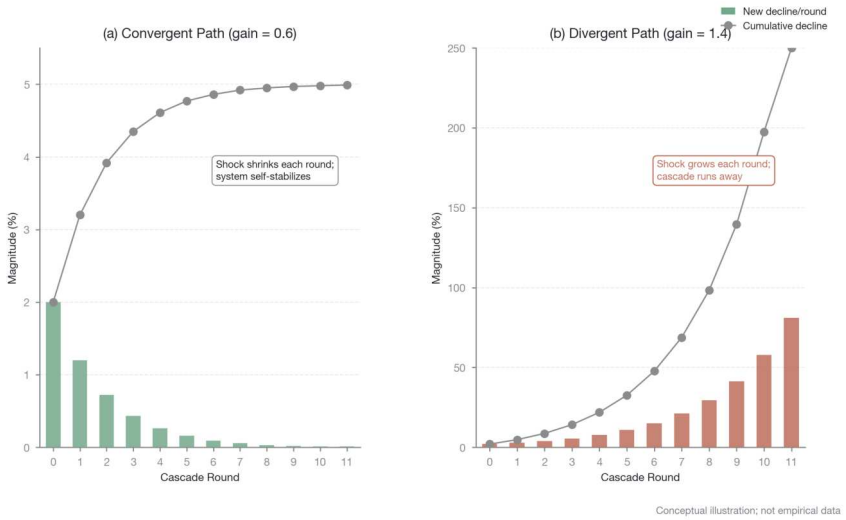


Figure 11-16. Simulation of cascade dynamics—convergent versus divergent paths (dynamic simulation; conceptual illustration; not empirical data)

As shown in Figure 11-16, the dynamic simulation reveals the starkly different evolutionary paths of the feedback gain G on either side of the critical value of 1: on the convergent path ($G = 0.6$), although the initial shock triggers liquidations, the incremental decline in each round attenuates rapidly, and the system recovers on its own after a few iterations with a limited cumulative decline; on the divergent path ($G = 1.4$), a tiny initial shock is continually amplified, the incremental decline in each round grows exponentially, and the result is ultimately a cliff-edge drop and the complete destruction of leverage. This shows that in a highly leveraged crypto market, the core of risk management lies not in predicting when the initial shock will arrive but in monitoring the system's fragility (the CRI is qualitatively positively correlated with G and can serve as an approximate cue to fragility, but $G > 1$ is the strict divergence criterion derived from the difference equation) and avoiding exposure to the market when the conditions for divergence are in place.

11.4 The coupling of the funding rate and leverage

In the market microstructure of perpetual futures, liquidation is the most direct and most violent manifestation of reflexivity, reshaping the price directly through an instantaneous "hard shock." This hard shock, however, does not occur out of nowhere; behind it lies a second reflexivity loop, subtler and slower but no less consequential: the chronic coupling of the funding rate, the basis, and leverage demand. The previous section's liquidation reflexivity equation described the instantaneous transmission of a liquidation cascade, whereas this section analyzes how system fragility is gradually accumulated before a liquidation erupts. The funding-rate mechanism does not directly create forced order flow, but by continuously affecting holding costs and leverage demand, it fundamentally alters the market's "flammability" and prepares the necessary soil for a liquidation cascade.

11.4.1 The margin-erosion effect

The funding rate was originally designed as a stabilizer that anchors the price to spot, but in a high-leverage environment it inevitably reveals its other face as a "chronic cash flow." For the paying side, the funding rate constitutes a continuous margin-consumption mechanism that steadily erodes the account's margin balance. This consumption is hard to notice in everyday trading, because it does not depend on violent swings in the market price but occurs mechanically with the passage of time. The direct consequence of this chronic erosion is that it dynamically raises the effective liquidation price, so that a position originally in the safe range imperceptibly moves close to the edge of liquidation. This process can be understood through a quantitative example. Amid extreme bull-market sentiment, the funding rate reaches 0.1% every 8 hours (about 109% simple annualized, about 198% compound annualized), a level recorded both at the peak of the 2021 bull market and on the eve of the October 2025 event [13]. A trader opens a 1 BTC long at the $100,000 price level with 10x leverage. The initial margin is $10,000, with a maintenance margin rate of 0.5% (i.e., $500).

Even if the market price stays exactly at $100,000, the position's margin balance is rapidly consumed within a short time simply because of funding payments. At this extreme rate, $300 of funding must be paid each day ($100 every 8 hours). At around day 31, the margin balance is fully depleted to the maintenance margin level, triggering liquidation. More dangerously, as margin decreases, its effective liquidation price moves up gradually from an initial level of about $90,500, eventually approaching the entry price. From the perspective of behavioral finance, this gradual erosion has a distinctive psychological stealth. The mental-accounting theory of Thaler (1999) [14] points out that traders tend to record funding-rate expenditures and position P&L in separate mental accounts, thereby underestimating the combined consumption effect of the two on the same margin pool. Prospect theory, meanwhile, predicts that small, high-frequency repeated losses are qualitatively different in psychological perception from a one-time loss of the same total amount; the funding rate deducted every 8 hours acts as a "small incision" whose cumulative effect is far below the trader's perception threshold, forming a "boiling frog" cognitive trap in which the continuous erosion of margin has pushed the position to the edge of liquidation before it is even perceived.

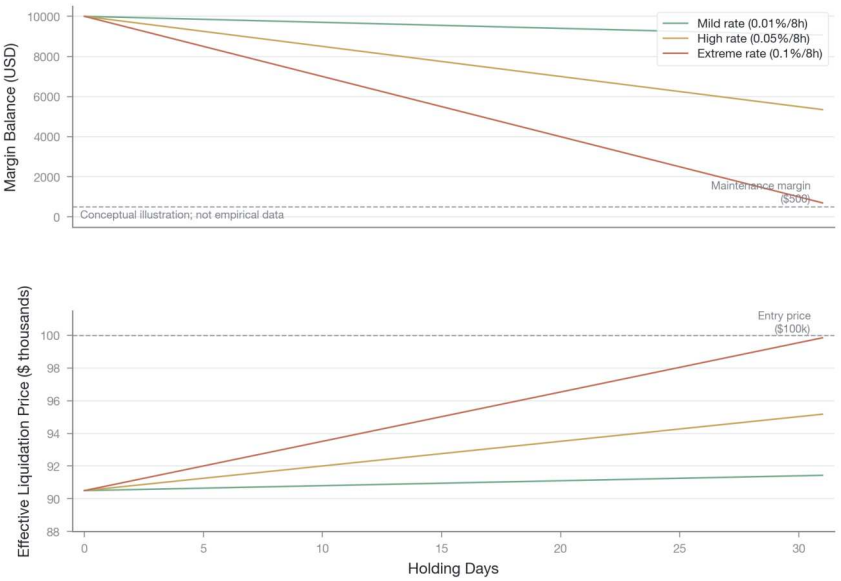


Figure 11-17. The erosion effect of the funding rate on the margin balance and the liquidation price (a numerical example based on an assumed rate; conceptual illustration; not empirical data)

The amplifying effect of the leverage multiple on this erosion process is nonlinear. Figure 11-18 shows the number of days required for positions at different leverage multiples to have their liquidation buffer fully depleted at the same funding rate (0.05% per 8 hours, i.e., about 55% simple annualized, about 73% compound annualized). A position at 100x leverage faces liquidation risk from rate erosion in only about 4 days, whereas a position at 5x leverage can withstand more than 60 days of rate consumption.

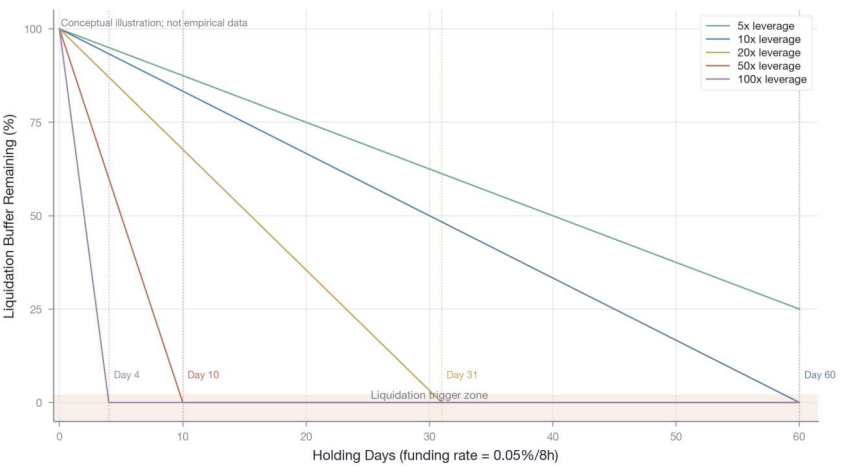


Figure 11-18. The cumulative erosion effect of the funding rate at different leverage multiples (a numerical example based on an assumed rate; conceptual illustration; not empirical data)

This constitutes a gradual process of risk accumulation. The funding rate is not only a pricing of the use of leverage but also a countdown on the account's survival time.

Moreover, the funding-rate mechanism itself contains a negative-feedback regulatory function: a high holding cost theoretically forces some leveraged longs to close voluntarily and simultaneously attracts arbitrageurs to enter through a "buy spot, short perpetual" strategy. This economic incentive is meant to push the market back toward equilibrium. This self-correcting design, however, relies on a key assumption: that arbitrageurs are both able and willing to carry out basis arbitrage. When this assumption is broken under extreme market conditions, the funding-rate mechanism degenerates from a "stabilizer" into an "accelerator."

11.4.2 The failure of basis arbitrage

In the normal market state, there is a healthy negative-feedback loop between the funding rate and basis arbitrage. When bullish sentiment runs high and drives the perpetual price above the spot price, the funding rate turns positive and keeps climbing. This attracts arbitrageurs: they short perpetual futures and go long spot, thereby depressing the perpetual price, narrowing the basis, and eventually returning the funding rate to a normal level. This loop is the core of the perpetual price's anchoring mechanism. During the transition from a normal period to a stress period, however, this negative-feedback mechanism can face three serious failure conditions that render the "stabilizer" ineffective.

Funding constraints are the primary factor in arbitrage failure. In an environment of extreme panic or credit tightening, arbitrageurs' funding costs rise sharply, and they may even face an inability to obtain funding at all. As the liquidity spiral model of Brunnermeier and Pedersen points out [8], when "funding liquidity" dries up, arbitrage is not only unprofitable but arbitrageurs themselves may even be forced to join the selling as they deleverage. At such times, far from acting as a force that stabilizes the market, they become drivers of greater volatility. During the "Black Thursday" event of March 12, 2020, even though the perpetual basis briefly fell to an extreme level below -30%, arbitrageurs were unable to enter because the funding market had frozen completely [4]. The crypto market, moreover, has no central-bank lender of last resort of the kind the traditional financial system provides. Once funding constraints tighten, no external mitigation mechanism exists, and the binding force of a liquidity drought is correspondingly far stronger than in traditional markets.

Inventory risk is the second constraint. When market volatility is extremely high, the cost of managing the price-fluctuation risk of a spot position becomes extremely expensive. Even if the basis between perpetual futures and spot appears to offer ample profit, arbitrageurs may choose to stay on the sidelines because they cannot effectively hedge the risk of a plunge on the spot leg. This arbitrage opportunity—present in theory but unobtainable in practice—allows the basis to remain anomalous for a long time. Shleifer and Vishny (1997) call this the "limits of arbitrage" [15]: even when there is an obvious pricing deviation, risk aversion and capital constraints prevent arbitrageurs from correcting it. At the same time, the cost of delta hedging also rises sharply, because in extreme conditions the implied volatility of the options market spikes in tandem, making the cost of hedging tail risk through options prohibitive.

Infrastructure-level constraints are especially prominent in the decentralized-exchange

environment. Extreme conditions are often accompanied by network congestion and spiking gas fees, which sharply compress the profit margin of arbitrage. At the same time, MEV activities such as sandwich attacks and transaction-ordering manipulation increase execution risk. The delays and security hazards of cross-chain bridges also limit cross-platform arbitrage efficiency. During the "yen carry-trade unwind" of August 5, 2024, API latency at multiple exchanges exceeded several seconds, making cross-platform arbitrage strategies nearly impossible to execute [16]. On the Ethereum network in particular, gas fees can spike 10- to 50-fold during extreme events, directly eroding the arbitrage profit margin and turning an otherwise profitable basis trade into a net loss after execution costs.

Beyond these three failure conditions, there are two further failure mechanisms that are equally critical in practice but less systematically analyzed. The first is funding-rate-reversal risk. Spot-perpetual basis arbitrage (long spot, short perpetual futures) earns a profit in normal markets by collecting a positive funding rate, but when a market crash triggers large-scale long liquidations, the funding rate can flip sharply from positive to deeply negative. At that point, the short position turns from a rate collector into a rate payer, and the arbitrageur faces a double penalty: the mark-to-market loss on the perpetual short (a temporary paper loss from price fluctuation, or margin pressure from an extreme widening of the basis) compounded by the continuous consumption of a negative funding rate. This double pressure may force the arbitrageur to close the short position that had been playing a stabilizing role, and closing a short is a buy operation, which reduces the hedging force against selling pressure on the perpetual side; at the same time, to rebalance delta, the arbitrageur may simultaneously sell the long position on the spot leg, thereby transmitting a shock originally confined to the perpetual futures market directly into the spot market. This mechanism reveals the direct link between Section 11.4.2 and Section 11.4.3: the failure of arbitrage means not only the passive absence of a stabilizing force but potentially the active transformation of a stabilizing force into a destabilizing one. The second is cross-exchange capital fragmentation. Executing basis arbitrage requires capital to be pre-distributed across multiple exchanges so that hedged positions can be established simultaneously on different platforms. During extreme events, however, exchanges may freeze or delay withdrawals, blockchain confirmation times may stretch from a normal few minutes to 10 to 60 minutes, and fiat deposit-and-withdrawal channels may close in tandem. Even if an arbitrageur has ample capital in aggregate, they cannot redeploy funds across platforms within a cascade window on the order of 40 minutes. This geographic and infrastructure-level capital fragmentation means that the system's effective arbitrage capital is far smaller than the nominal available capital, and the actual ceiling on arbitrage-absorption capacity is determined by the weakest link rather than by total capital.

When these constraints all take effect at once, the basis can widen sharply into deeply negative (or extremely positive) territory while arbitrageurs' absorption capacity rapidly decays. At this point, a high funding rate can no longer effectively suppress overheated leverage demand and instead becomes the critical factor that drives high-leverage positions into liquidation (Figure 11-19 contrasts two states: the effective operation of negative feedback in a normal period and the runaway of positive

feedback in a stress period).

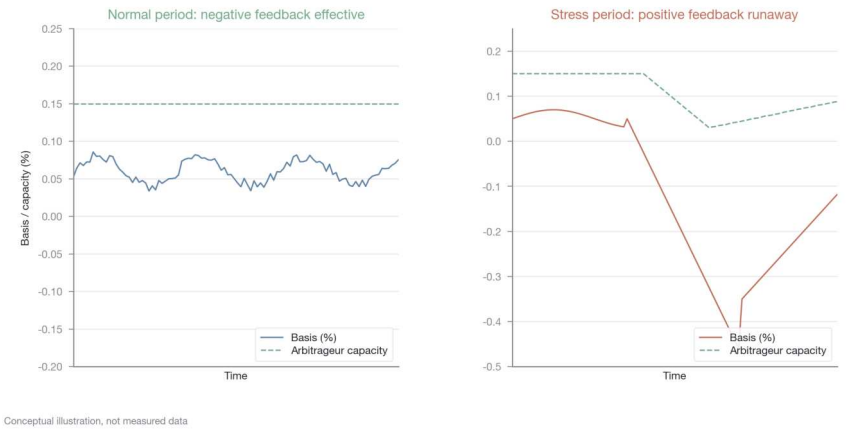


Figure 11-19. Comparison of the basis-arbitrage mechanism in normal and stress periods (conceptual illustration, not empirical data)

Figure 11-20 uses a flowchart to contrast how the arbitrage mechanism operates in normal versus stress periods.

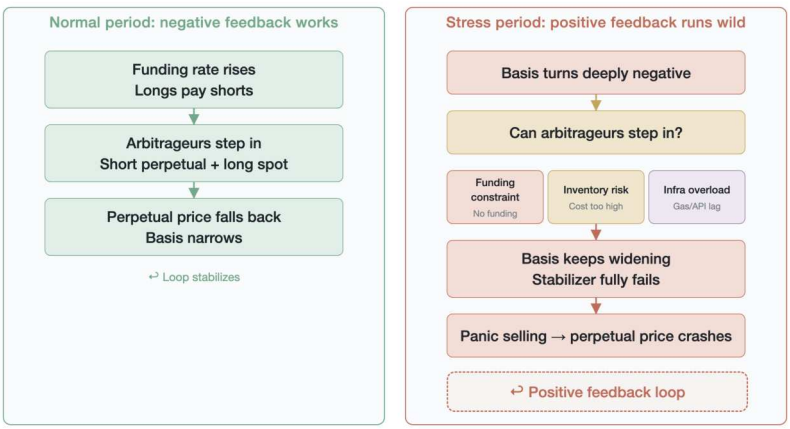


Figure 11-20. The process mechanism of basis-arbitrage failure

The core insight the flowchart conveys is that under normal market conditions arbitrageurs play the role of a negative-feedback stabilizer: when the basis deviates from a reasonable range, arbitrageurs' entry trades push it back toward equilibrium, closing the feedback loop and maintaining the anchoring relationship between the perpetual price and the spot price. Under stress conditions, however, three constraints activate simultaneously—funding constraints (arbitrageurs cannot obtain enough margin or borrowed funds), inventory risk (holding a spot hedge position faces violent mark-to-market losses), and infrastructure overload (exchange delays, API failures, on-chain congestion)—forcing arbitrageurs to withdraw or preventing them from entering at all.

When the executor of negative feedback is absent, the basis deviation is no longer corrected but instead keeps widening through panic contagion and a liquidity-siphoning effect, and the entire feedback loop turns from negative to positive. This shift in mechanism provides the direct precondition for the compound rate-liquidation positive feedback discussed in the next section: it is precisely because of arbitrageurs' systemic failure during a stress period that an extreme deviation in the funding rate can keep accumulating without being corrected by market forces.

11.4.3 The compound rate-liquidation positive feedback

The chronic erosion of the funding rate and the instantaneous shock of the liquidation mechanism do not operate in isolation; they are tightly coupled through their shared core variable, margin, forming a compound positive-feedback chain that weaves indirect pressure together with direct shock. This loop constitutes the most dangerous systemic-dynamics feature of perpetual futures markets.

This loop usually begins with a rate-extremization phase: in a strong one-sided trend, the funding rate stays at an extreme level. This high rate is deducted from the position holder every 8 hours (or, at some exchanges, every hour), continuously consuming their margin balance. This process systematically narrows the liquidation distance of all positions in the same direction, quietly "accumulating fragility" for the coming liquidation cascade. Because the system has been continuously consumed by the rate and has become extremely fragile, a tiny price reversal that would not otherwise be enough to spark a crisis (for example, a normal 2% pullback) is now sufficient to touch the liquidation line of positions that the rate has pushed to the edge, triggering the first batch of forced closures. The selling pressure from the first batch of liquidations drives the price down further and triggers more liquidations, and the liquidation reflexivity equation begins to diverge. As large-scale liquidation is executed, the market's position imbalance changes drastically, causing the funding rate to flip sharply—for example, from an extreme positive value to an extreme negative one in an instant. This sharp flip is not only a mechanical reaction of the market structure but also sends a strong signal to the entire market that "the trend has changed." This prompts the remaining longs to close voluntarily in a panic, joining forces with the liquidation engine's forced selling. In the end, high leverage is thoroughly cleared out, open interest plunges, and the system resets violently.

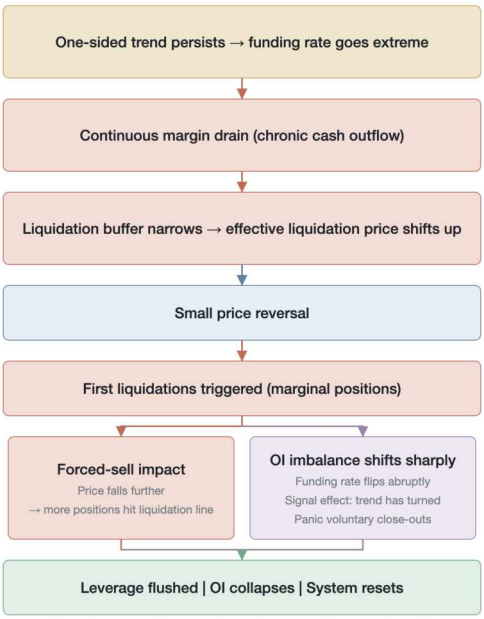


Figure 11-21. The compound rate-liquidation positive-feedback loop

In this nested loop, the "extreme termination path" is far more common than the "mild path." The mild path requires the price to stop moving in one direction before liquidations are triggered on a large scale, which is an extremely narrow condition in a system that has long been eroded by a high rate. By contrast, rate extremization itself accumulates potential energy for the liquidation cascade, and once the price reverses, this potential energy is released all at once. As the "10.10 event" of October 2025 demonstrated, the fundamental cause of a systemic collapse is often not the macro signal that triggers it but the extreme rate consumption of the preceding weeks, which is precisely what degrades the system from a robust metastable state into an extremely fragile, flammable one [17].

A counterfactual exercise helps quantify the contribution of rate erosion to the scale of the cascade. On the eve of the "10.10 event," the funding rate on BTC perpetual futures was about 30% annualized, which means the margin of a 10x-leverage long was eroded by about 1.2 percentage points over two weeks, compressing the liquidation distance of many positions from more than 5% to less than 3%. Had the rate been at a mild 10% annualized (the normal state for most of 2024), the margin erosion over the same holding period would have been only about 0.4 percentage points, and the vast majority of positions eroded to the edge would still have retained an ample liquidation buffer. On this basis, in the counterfactual 10%-rate scenario, the scale of positions liquidated in the initial trigger phase would have been reduced by about 40% to 50%. This range estimate is sensitive to the assumptions about the shape of the leverage distribution and the holding period; if leverage concentration were higher or the average

holding time longer, the actual reduction could deviate from this range. Because of the nonlinear amplification of the liquidation reflexivity equation (each round's liquidation volume determines the next round's trigger scale), the total liquidation scale of the cascade might have shrunk by about 35% to 45%, that is, to about \$10.5 billion to \$12.5 billion. This counterfactual exercise is a partial-equilibrium analysis: a lower funding rate would itself change the equilibrium open interest and market makers' willingness to supply liquidity, placing the entire market microstructure in a different initial state. Moreover, the mapping from initial trigger scale to total cascade scale is highly nonlinear, and its amplification factor depends on the specific shape of the liquidation density function (that is, whether the distribution of liquidated positions on the price axis is a concentrated "fat tail" or a dispersed "thin layer"), a distributional feature that directly determines the cascade's acceleration and duration. This estimate shows that the long-run accumulation effect of the funding rate has an influence on the cascade's ultimate destructiveness of the same order of magnitude as the trigger shock itself, and that rate governance should be regarded as a core element of systemic-risk management rather than merely an auxiliary tool for anchoring the market price.

The funding rate and the liquidation mechanism—one slow and one fast, one hidden and one visible—together forge the endogenous systemic risk of perpetual futures markets.

11.5 The four phases of a liquidation cascade

A liquidation cascade is not a crash achieved in one step but a dynamic process with a clear evolutionary path. It typically passes through four logically progressive phases: trigger, propagation, amplification, and termination. In each phase a different mechanism dominates the market's direction, and identifying the key signals of these phase transitions not only provides observable clues for understanding extreme events but also offers a theoretical basis for risk warning and intervention. This section integrates the liquidation reflexivity equation (Section 11.3) and the rate-coupling effect (Section 11.4) discussed earlier into a complete dynamic narrative, using the "10.10 event" of 2025 as the primary reference case.

Table 11-4 summarizes the key decision indicators and thresholds for the four phases, providing an operational reference framework for identifying in real time which phase a cascade is in.

Phase	CRI range	LIAR range	Feedback gain G	OI change	Depth change	Dominant mechanism
Trigger	0.8–1.2	> 1.0	< 1.0	$< -5\%$	$< -30\%$	Exogenous shock + marginal liquidations
Propagation	1.2–2.0	0.5–1.0	≈ 1.0	-5% to -15%	-30% to -70%	Liquidation reflexivity loop activates
Amplification	> 2.0	< 0.5	1.0	-15% to -35%	$> -90\%$	Multi-mechanism resonance; depth goes to zero
Termination	Falls back to < 1.0	Recovers to > 1.0	< 1.0	Stabilizes	Slow recovery	Leverage cleared + external buying

Table 11-4. CRI/LIAR decision thresholds and key indicators for the four phases of a liquidation cascade (Note: LIAR, the liquidation-impact absorption rate, $\approx$ available depth / liquidation volume per unit time, which is approximately inversely proportional to the feedback gain $G = \alpha\beta L/D$; the two are based on different measures—LIAR uses instantaneous visible depth and G uses cross-exchange effective depth—so they are placed side by side as independent diagnostic columns. The CRI column in this table is an illustrative standardized scale, on a different basis from the unstandardized raw estimates in Table 11-3, so the empirical banding of Section 11.3.5 cannot be applied to it directly. The indicators in this table are contemporaneous phase diagnostics, not forward-looking warnings.) (Data source: constructed by the author)

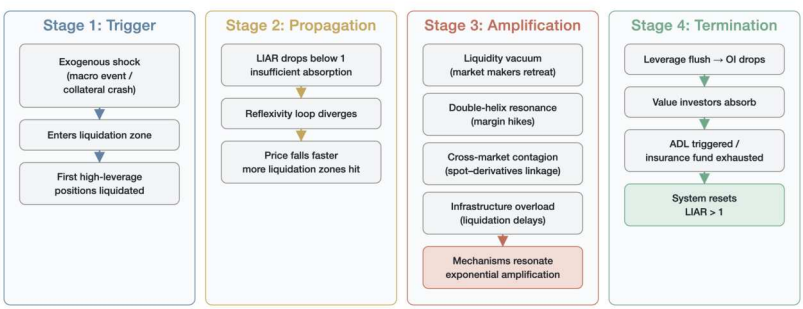


Figure 11-22. Overview of the four-phase liquidation-cascade process

11.5.1 The trigger phase

The starting point of a liquidation cascade is usually an exogenous market shock. This shock need not be large in itself; its key role is to break the system's prior metastable state and push the market price precisely toward the first significant liquidation-dense zone. Here the relationship between the trigger factor and the underlying fragility must be clearly distinguished: the trigger factor is a necessary but by no means sufficient condition, and what really determines whether a shock is absorbed or amplified is the degree of structural fragility the system has already accumulated before the shock arrives—including leverage density, the concentration of liquidation prices, and the thickness of the liquidity buffer.

The sources of trigger factors are highly diverse. A trigger may be a piece of macroeconomic news (such as the unexpected report about the trade dispute in the "10.10 event"), a violent move in the price of a related asset (such as a plunge in an important collateral asset or a brief stablecoin depeg), or even a malicious attack on a weak point in market liquidity (such as "oracle manipulation"). Table 11-5 summarizes the trigger factors and key characteristics of several major liquidation-cascade events in recent years.

Event	Trigger factor	24-hour liquidations	BTC max drawdown	Main cascade duration
March 12, 2020, "Black Thursday"	COVID-19 panic + global sell-off of risk assets	~$6 billion	~39%	~48 hours
May 19, 2021, "5.19 crash"	China's regulatory policy + Musk's remarks	~$8.6 billion	~30%	~12 hours
June 13, 2022, "Luna aftershock"	Chain liquidations of Three Arrows Capital/Celsius	~$4.8 billion	~17%	~24 hours
August 5, 2024, "yen carry-trade unwind"	Bank of Japan rate hike + global unwinding of carry trades	~$10.2 billion	~15%	~8 hours
October 10, 2025, "10.10 event"	Unexpected escalation of the trade dispute	~$19 billion	~14.5% (spot; perpetual wicks of about 16%–17%)	~1.5 hours

Table 11-5. Comparison of major liquidation-cascade events in recent years (Data source: CoinGlass [9], CoinGecko [3], Amberdata [2]). Note: the 24-hour total liquidations for each event use different sources and statistical bases (single-exchange versus market-wide, single-day versus multi-day, and notional versus actual liquidations differ substantially), and the 2020/2021 figures are aggregate estimates with high range uncertainty, so cross-event comparisons should emphasize orders of magnitude rather than precise values

A thought-provoking trend can be observed in the table above: as the crypto derivatives market has grown, the visible forced-liquidation notional of liquidation cascades has kept climbing while the duration of the cascades has shortened markedly. The "10.10

event" of 2025 set a new record for single-day liquidations in only about 1.5 hours (see the chapter opening), indicating that both the leverage density and the liquidation efficiency of the market are rising sharply and that the reaction window left to traders and risk-control systems is narrowing.

In the trigger phase, the market's reaction is usually local. Only the traders with the highest leverage and the most extreme risk appetite are liquidated. At this point, overall market liquidity is still ample, and the forced-closure orders produced by the liquidation engine can be absorbed smoothly by the market. The price decline is relatively linear, and the system has not yet entered an out-of-control state. Whether the trigger phase evolves into the more serious propagation phase depends on one key dynamic balance: whether the selling pressure from the first wave of liquidations exceeds the market's normal absorption capacity. If market depth is sufficient to take on this selling, the shock is dissolved in this phase; if the selling pressure is too great, the system inevitably slides into the next phase. In terms of the LIAR indicator introduced in this chapter, the trigger phase corresponds to the range in which LIAR is still greater than 1—that is, market depth can still take on the liquidation selling pressure. What distinguishes an event that dissolves on its own in the trigger phase from one that ultimately evolves into a systemic cascade is whether the first round of liquidations pushes LIAR below the critical value. Historical data show that in the "yen carry-trade unwind" of August 2024 LIAR briefly fell below 1 in the trigger phase and then quickly recovered, whereas in the "10.10 event" of 2025 LIAR, once it fell below 1, kept accelerating downward and never turned back.

11.5.2 The propagation phase

To quantitatively assess the critical point of phase transition, we introduce a core decision variable: the liquidation-impact absorption rate (LIAR). The design advantage of LIAR is that it integrates both the demand side (liquidation volume per unit time) and the supply side (available order-book depth) of the liquidity equation, making it a more comprehensive measure of system stress than a bare open-interest/market-cap ratio or a simple depth indicator.

> LIAR = available order-book depth within a specific price range / liquidation volume per unit time

When LIAR > 1, the market has enough liquidity to absorb the current liquidation shock. Although the price may fall, its rate of decline is relatively controllable, and the liquidation cascade tends to converge. When LIAR < 1, however, liquidation volume has overwhelmed the market's immediate absorption capacity. At this critical point, the price begins to accelerate downward, the divergence condition of the liquidation reflexivity equation described in Section 11.3 is satisfied, and the positive-feedback loop is formally activated. The system state undergoes a jump into the highly destructive amplification phase. LIAR is essentially a real-time phase-diagnostic indicator, not a forward-looking warning signal. By the time LIAR falls below 1, the cascade is already under way, and the collapse of order-book depth is itself an endogenous product of the liquidation process rather than an exogenous condition independent of it. LIAR's

denominator (available order-book depth) contracts sharply after liquidation begins as market makers withdraw their orders, which means the indicator reflects the current system state rather than the future trajectory. A truly forward-looking improved version should use estimated depth under a stress scenario (for example, assuming that X% of visible resting orders are withdrawn under the shock of adverse order flow, which is typical market-maker behavior when facing toxic liquidation flow) and compute the ratio of remaining depth to expected liquidation volume under that condition, so as to identify the system's fragility before a cascade actually begins.

The essence of the propagation phase is its contagiousness: the price decline caused by the first wave of liquidations does not stay in place but pushes the market price toward the next liquidation zone—potentially larger and more densely distributed—like the first link of a chain reaction, setting off a second and a third wave of liquidations. In this phase, the reflexivity loop between leverage and liquidation begins to replace fundamental information as the core force dominating the price's trajectory. The market's price movements are no longer driven by fundamental information but dominated by the leverage-clearing process. Figure 11-23 uses four panels to show, respectively, the dynamic evolution of price, open interest, per-minute liquidation volume, and the LIAR indicator across the full trigger-propagation-amplification-termination process (based on the data pattern of the 10.10 event of 2025).

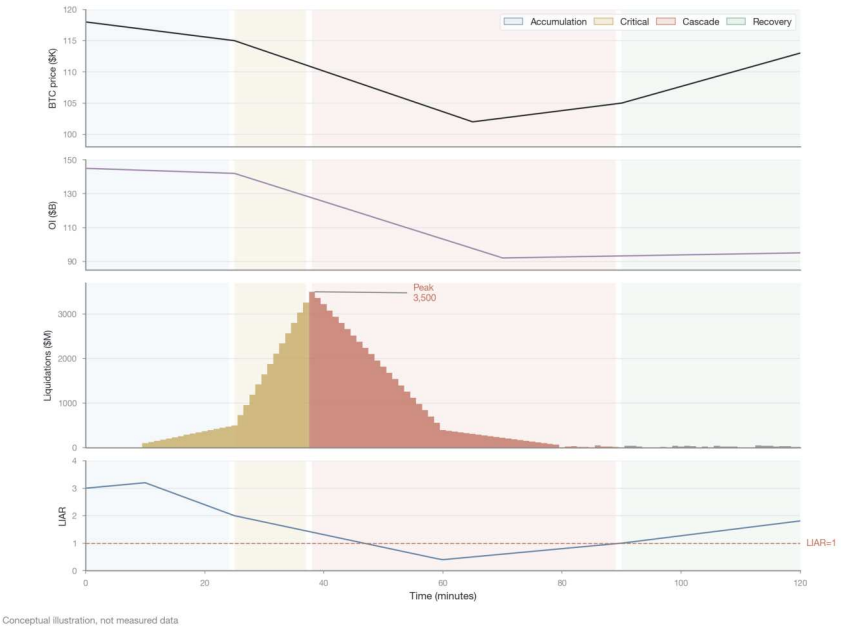


Figure 11-23. A four-phase dynamic model of the liquidation cascade (conceptual illustration: a stylized dynamic model by the author based on the data pattern of the "10.10 event" of 2025, not tick-by-tick empirical data)

11.5.3 The amplification phase

When LIAR stays below 1, the system enters the most destructive phase of the entire cascade: the amplification phase. Now a single loss spiral evolves into a resonant superposition of multiple amplification mechanisms, sharply magnifying what was a controllable risk into a systemic crisis, with multiple amplification mechanisms acting simultaneously and superimposing in resonance.

As liquidation orders continue to consume the buy orders on the order book mechanically, market makers, facing a sharply expanding risk exposure and extremely high uncertainty, choose to withdraw their quotes (i.e., "procyclical liquidity supply"). This causes order-book depth to fall sharply and bid-ask spreads to widen markedly. Subsequent liquidation orders face enormous slippage on execution, with actual fill prices far below the mark price. This slippage itself constitutes an additional huge loss that further erodes traders' margin and may even lead to bankruptcy (i.e., losses exceeding all posted margin). During the "Black Thursday" crash of March 2020, BitMEX's BTC perpetual futures briefly wicked to about $3,600 on March 13 (an instantaneous low at a single venue), a discount of about 12% to the spot price—an extreme manifestation of the liquidity vacuum [4].

The violent price fluctuations simultaneously trigger the "margin spiral" of the Brunnermeier and Pedersen model [8]. To control risk, an exchange's automated risk model may dynamically raise maintenance margin requirements or lower the collateral factor on collateral. This suddenly subjects positions that had been safe to margin-call pressure, forcing traders to close voluntarily. This active selling resonates with the liquidation engine's passive sell orders, jointly intensifying the market's selling pressure. Under a unified margin system this effect is especially pronounced: when the volatility metric rises sharply, the system automatically raises margin requirements for all instruments, subjecting a large number of previously safe positions to margin-call pressure at once.

Under the unified margin system, a plunge in one asset weakens the credit base of the entire account, forcing the liquidation of positions in otherwise unrelated assets. At the same time, arbitrageurs' behavior rapidly transmits the selling pressure of one exchange to the whole market, forming cross-exchange contagion. In addition, the liquidation of perpetual futures depresses the futures price and turns the basis negative, prompting arbitrageurs to sell spot (closing the spot long in their arbitrage position) and thereby transmitting the crisis from the derivatives market to the spot market, forming a cross-market reflexivity loop. This mechanism explains why, during a liquidation cascade, the spot market's decline is often highly synchronized with the derivatives market and can even exhibit the anomaly of spot leading the fall. Determining the causal direction between spot and perpetual futures during a cascade, however, requires rigorous lead-lag analysis methods (such as the Hasbrouck information-share model or the Granger causality test). In the 10.10 event, the causality was most likely two-way: the liquidation selling pressure of perpetual futures and the panic selling of the spot market reinforced each other, forming a closed positive-feedback loop, which makes clear causal attribution methodologically very difficult in the absence of high-frequency data decomposition.

The carrying capacity of infrastructure is far from a secondary technical constraint in

cascade dynamics; it is one of the primary determinants of the cascade's trajectory. During an active cascade, which infrastructure component fails first (the API gateway, the liquidation queue, or the matching engine) determines the cascade's actual evolutionary path. In BitMEX's service outage of March 2020 (March 13), matching and liquidation stalled for about 25 minutes (officially attributed to a DDoS attack), producing an enormous gap between actual execution prices and theoretical trigger prices and thereby sharply amplifying bankruptcy losses [4]. Some exchanges are reported to have deliberately slowed their matching engines in extreme conditions, using this as a de facto circuit breaker. Although this practice may objectively have had a protective effect in curbing the cascade, it raised serious questions of market fairness, because it constituted an implicit intervention that gave some participants (for example, institutions able to trade through over-the-counter channels) an asymmetric advantage over retail traders who depend on the exchange engine. Infrastructure capacity is therefore not merely an engineering problem but a key variable determining whether a cascade is contained within a controllable range or evolves into a catastrophic collapse.

In extreme conditions, an exchange's trading engine may lag, API responses may slow, and the front-end interface may fail to load. Traders cannot add margin, adjust positions, or set stop-losses in time. More seriously, the liquidation engine itself may become congested as requests surge, delaying liquidation execution. During the delay, the price has often already fallen to a lower level, further amplifying the actual loss.

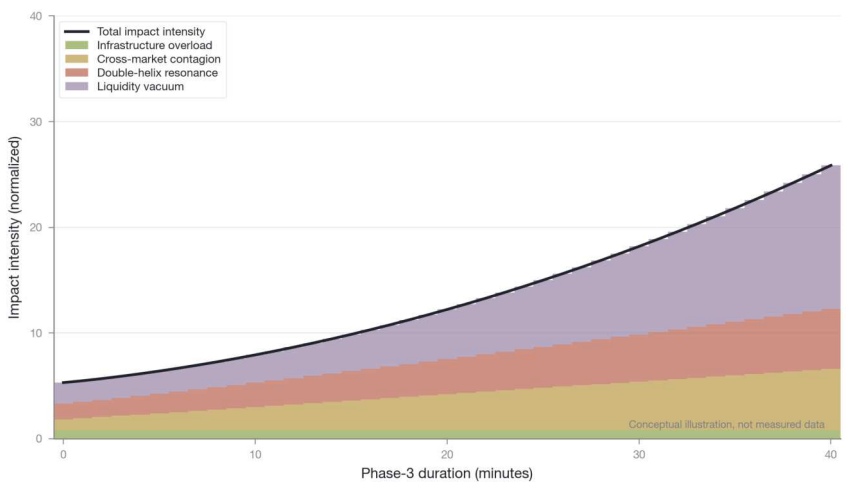


Figure 11-24. The resonant superposition of four mechanisms in the amplification phase (conceptual illustration, not empirical data)

To clearly delineate the risk level the system is in, we can divide the interaction state of the perpetual futures and spot markets into three categories, as shown in Table 11-6. This classification framework helps to assess in real time whether the market is sliding from a normal state toward a dangerous one.

Market state	Basis characteristics	Arbitrage activity	Feedback type	Order-book depth behavior	LIAR indicator
Normal state	Narrow fluctuation ($<$ 0.1%)	Active, full absorption	Negative feedback, self-correcting	Ample, normal level	> 2.0
Stress state	Significant widening (0.1%–0.5%)	Under pressure, partial withdrawal	Positive feedback beginning to appear	Declining, down 30%–60%	0.5–1.0
Extreme state	Violent deviation ($>$ 0.5%)	Withdrawal, near standstill	Positive feedback fully out of control	Evaporation, more than 90% gone	< 0.2

Table 11-6. Three interaction states of the perpetual futures and spot markets (Data source: compiled by the author)

11.5.4 The termination phase

A liquidation cascade does not continue indefinitely. Once the market has passed through this violent phase, it enters the termination phase. A cascade, however, usually ends not because the market has found a new fundamental equilibrium but because the leverage available to be liquidated has already been exhausted. This is a violent but effective forced-deleveraging mechanism.

The termination of a cascade is usually the joint work of several mechanisms. Among them, leverage clearing is the most natural way to terminate. Once the vast majority of high-risk leveraged positions have been force-closed, subsequent price declines can no longer trigger enough new liquidation volume. At this point, selling pressure drops abruptly, the LIAR indicator recovers back above 1, and the loss spiral stops operating. During this process, market-wide open interest falls sharply. In the "10.10 event" of 2025, according to Amberdata, open interest in crypto perpetual futures fell from a peak of about \$146.6 billion to a trough of about \$109.9 billion (a peak-to-trough decline of about -25.03%; this basis is Amberdata's aggregate perpetual OI across the venues it covers, accessed on June 23, 2026), marking a large-scale clearing of leverage [2].

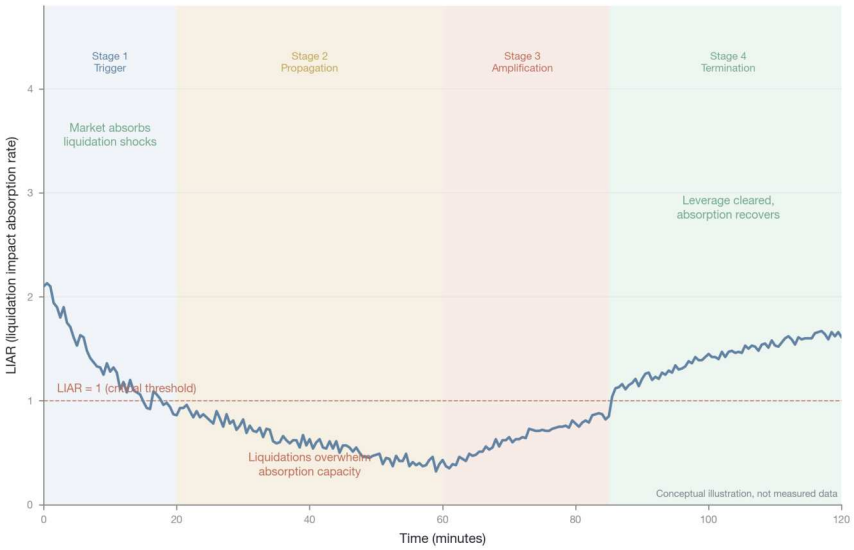


Figure 11-25. The dynamic evolution of LIAR across the four phases of a liquidation cascade (LIAR is an indicator constructed by the author; conceptual illustration, not empirical data)

As the figure shows, LIAR falling below 1 marks the activation of the reflexivity loop (the propagation phase), remaining below 0.2 marks the extreme state of the amplification phase, and recovering back above 1 marks the termination of the cascade.

When the price falls, driven by irrational liquidation selling, far below the fair value recognized by long-term investors, the entry of external buying also helps the market stabilize. These value investors provide deep absorbing liquidity that effectively soaks up the remaining liquidation selling pressure. In on-chain markets, this phenomenon manifests as the on-chain behavior of "whale addresses" buying heavily at the bottom of a crash. On-chain data analysis can track the fund flows of these large addresses in real time, and when a cohort of addresses holding more than 1,000 BTC shows accelerating net inflows during a rapid price decline, it often foreshadows the formation of bottom-absorbing force. This contrarian buying by "smart money" provides a natural price floor for panic selling.

The termination of a cascade also has a behavioral dimension that goes beyond mechanical leverage clearing. Panic selling follows a specific temporal pattern (an initial freeze response, then concentrated selling, and finally the exhaustion of the capacity for emotional selling), a process that runs independently of the mechanical liquidation cycle. Defining all whale-address buying at the bottom of a cascade as the value investing of "smart money" warrants careful critical scrutiny: some large-address activity at a cascade's bottom involves strategic price manipulation (buying to trigger a short squeeze), the release of social-media signals (publicly announcing purchases to influence market sentiment), or the building of reputational capital, rather than a deep judgment about the asset's fundamental value. The actual bottoming process often passes through a vacuum period in which no effective two-sided market exists—neither sufficient buy-side liquidity nor sustained sell-side pressure—and the market gradually finds a temporary equilibrium amid extremely shrunken trading volume.

Moreover, the narrative of orderly value buying at the bottom of a cascade contradicts the actual execution environment in significant ways. During an active cascade, an exchange's API interface frequently becomes unresponsive or imposes strict rate limits, making order submission extremely difficult or even impossible. Deposit and fund-transfer mechanisms may be suspended, preventing new capital from entering the market. Extreme bid-ask spreads (peaking at 26.43 basis points in the 10.10 event) mean that any buy order submitted as a market order will suffer enormous slippage losses. Most critically, professional market makers choose to withdraw rather than enter during a cascade, because the adverse-selection problem of liquidation flow is extremely severe, and any buy order resting on the order book is systematically eaten by the continually declining forced sell orders. A true bottom often forms in a vacuum period in which neither side provides meaningful liquidity, rather than through an orderly handoff from forced sellers to voluntary buyers.

An exchange's manual intervention can likewise arrest a cascade. In extreme cases, an exchange may take various intervention measures, such as temporarily raising maintenance margin requirements to suppress the impulse to open new leverage, or, in some highly contested cases, triggering a brief circuit breaker (suspending trading) to stem the spread of panic. Some exchanges also temporarily lower the maximum leverage multiple or impose trading restrictions on specific contracts.

After bankruptcy losses exhaust the insurance fund, auto-deleveraging is activated as the final resort, force-closing the positions of profitable parties to cover the bankruptcy losses and preserve system solvency (its ranking mechanism and fairness disputes are detailed in Section 11.6.3). The large-scale triggering of ADL usually marks the end of the most dangerous and chaotic phase of a liquidation cascade.

Viewed from the standpoint of system dynamics, the entire four-phase process reveals a clear regularity: the destructiveness of a liquidation cascade depends not primarily on the size of the trigger shock but on the degree of fragility of the system before the shock arrives. A system long eroded by extreme funding rates, with very high leverage density and a highly concentrated liquidation-dense zone, can erupt into a devastating cascade even in the face of a tiny exogenous disturbance. Conversely, a system with moderate leverage, a dispersed liquidation distribution, and ample liquidity can smoothly absorb even a large exogenous shock. This is precisely the core proposition this chapter has repeatedly stressed: the reflexivity of leverage and liquidation is the most important source of endogenous risk in the crypto derivatives market.

11.6 The impossible trinity of liquidation design

In on-chain derivatives and crypto-asset perpetual futures markets, the liquidation mechanism is the last line of defense for maintaining system solvency. Unlike traditional financial markets, which handle risk through a clearinghouse's default waterfall and manual intervention, liquidation in crypto markets relies entirely on preset algorithms and smart contracts. This algorithmic character means that the design of the liquidation mechanism faces severe trade-offs. Research shows that a strict mathematical boundary exists within automated deleveraging and liquidation mechanisms: the

"impossible trinity of liquidation design" [18].

11.6.1 The three design objectives

In designing a liquidation engine, any exchange or DeFi protocol tries to achieve the following three core objectives simultaneously:

The optimal-execution theory of Almgren and Chriss shows that executing a large order involves a fundamental trade-off between market impact and timing risk [19]. The design of a liquidation engine faces essentially the same dilemma, but under harsher constraints, because liquidation must be completed in an extremely short time and cannot choose its timing. Against this backdrop, the conflict among the three design objectives becomes especially sharp. Certainty means the liquidation mechanism must be able to ensure that the system remains solvent under any extreme market condition—that is, when a trader's account equity falls below the maintenance margin requirement, the system must be absolutely certain of being able to close the position and prevent bad debt (bankruptcy); this is the cornerstone of an exchange's survival, and a lack of certainty leads to systemic insolvency. Impact minimization means the negative effect of liquidation on market price and liquidity should be as small as possible, because large-scale forced closures release enormous selling (or buying) pressure into the market, drive the price further from fundamentals, and can thereby trigger the liquidation of more accounts in a "positive-feedback loop"; impact minimization therefore requires the liquidation engine to digest risk positions smoothly. Fairness means the liquidation process should be fair to all market participants and, in particular, must not deprive innocent traders (such as profitable ones) of their legitimate gains without cause; in an ideal state, risk should be borne by the party that took it on (the liquidated party) or absorbed by a system-established insurance fund, rather than arbitrarily shifted to other participants in the system.

11.6.2 The conflict among objectives

There is a deep structural conflict among these three objectives that makes it impossible to achieve all of them perfectly in a single mechanism at the same time. Recent theoretical research on auto-deleveraging mechanisms has even given a formal proof of an impossibility theorem from the perspectives of game theory and mechanism design [18].

The conflict between certainty and impact minimization is the most intuitive. To guarantee 100% certainty, the system often needs to sell all positions at market immediately when liquidation is triggered. This cost-agnostic selling, however, inevitably causes an enormous market impact. Conversely, if the system chooses to process positions slowly and in batches to minimize impact, then in extreme conditions of violent price fluctuation this delay may cause a position to go bankrupt before it is fully closed, thereby undermining certainty.

At the same time, the conflict between certainty and fairness is especially prominent in extreme conditions. When routine liquidation cannot be executed and the insurance fund is exhausted, the system's only option to preserve its own certainty (avoiding insolvency) is to activate auto-deleveraging, forcibly closing profitable traders' positions

to cover the losing side's bad debt. Although this preserves the system's certainty, it severely harms the fairness owed to profitable traders, forcing them to bear a risk that was never theirs.

Further analysis shows that the conflict between impact minimization and fairness manifests in the process of risk transfer. If, to avoid slamming large positions directly onto the market (impact minimization), the system chooses to transfer these positions to specific liquidity providers or market-making vaults, it is in effect concentrating unknowable tail risk onto that group of participants. When the market keeps deteriorating, these backstops may face a devastating blow, which is essentially also an unfair allocation of risk to a specific group.

Notably, this impossible trinity is not a wholly new discovery of crypto markets. The central counterparty (CCP) clearinghouses of traditional financial markets face the same fundamental tension in their mechanism design among certainty, impact minimization, and fairness (Duffie and Zhu, 2011 [20]; Pirrong, 2011 [21]). A CCP absorbs losses layer by layer through a default-waterfall mechanism and, under the constraints of a legal framework, achieves relatively clear rules for risk allocation (the layered structure of the central counterparty and the default waterfall is discussed in Chapter 1). The crypto derivatives market, however, faces harsher constraints on three dimensions. First, there is no legal-level loss-mutualization framework: a traditional CCP's loss-sharing mechanism relies on legally binding agreements among clearing members and on regulatory authorization, whereas the relationship between a crypto exchange and its users is established only through terms of service, lacking an equivalent legal basis for multilateral risk mutualization. Second, there is no central bank as a lender of last resort providing liquidity support: when a traditional CCP faces systemic pressure, the central bank can inject liquidity through emergency lending facilities to prevent cascading defaults, whereas the crypto market has no equivalent external liquidity-injection mechanism. Third, the time scale of margin calculation is compressed from a traditional CCP's intraday or overnight settlement to real-time, position-by-position updates, so that the entire process from detecting a shortfall to executing liquidation is compressed from hours to milliseconds, largely eliminating any possibility of manual intervention or mediation.



Figure 11-26. The impossible trinity of liquidation design—the positioning of the four liquidation modes (conceptual illustration, not empirical data)

11.6.3 The four liquidation modes

Faced with the impossible trinity, different exchanges and protocols make different trade-offs according to their own positioning and technical architecture, giving rise to four typical liquidation modes:

Cliff-edge liquidation is common at early crypto exchanges and some DeFi protocols (the mechanism is detailed in Section 11.1.3). It strikes a balance between certainty and fairness: it strictly follows the principle that "whoever defaults is liquidated" and executes decisively, but at the cost of abandoning impact minimization entirely, so it readily triggers a deep price cliff and a liquidation cascade when liquidity is scarce. Progressive liquidation is adopted by most modern centralized exchanges, exemplified by Binance and OKX (the mechanism is detailed in Section 11.1.3) [22]. It reduces positions by a fixed proportion or in a stepped fashion, mitigating market impact to some degree and being relatively fairer to the liquidated party, but it also sacrifices some certainty (the price may keep deteriorating and drive the remaining position into bankruptcy). The vault-backstop mode is a new mode that has emerged in recent years at decentralized perpetual futures exchanges such as Hyperliquid (the mechanism is detailed in Section 11.1.3) [23]. It transfers liquidated positions to a market-making vault composed of liquidity providers (such as the HLP vault) to be digested gradually, achieving impact minimization to a great degree but severely sacrificing fairness, because vault investors are forced, without their knowledge, to bear the tail risk of high-leverage traders' high-risk assets. Auto-deleveraging is the extreme mechanism that exists as the last line of defense (the mechanism is detailed in Section 11.1.3): when bankruptcy has already occurred and cannot be covered by routine means, the system forcibly closes the positions of the profitable side, ranked by "most profitable, highest leverage," to hedge the bad debt, pushing certainty to the extreme at the cost of shattering fairness completely. In the extreme conditions of October 10, 2025, research

shows that Hyperliquid triggered the first cross-margin auto-deleveraging in its operating history, of which about $653 million was overshoot beyond the actual bankruptcy shortfall (about $23 million), corresponding to an additional loss of about $45 million to $52 million for profitable traders [18] (this quantitative conclusion comes from Chitra's preprint and awaits cross-validation against on-chain data and peer review). The judgment of "overuse," however, requires further precision. Key questions include: at the moment ADL was triggered, did the insurance fund still have sufficient solvency to cover the bankruptcy losses? Did the ADL ranking algorithm (usually a weighted ranking based on profit percentage and effective leverage) have a bias in actual execution? Were the threshold parameters that trigger ADL adequately calibrated and backtested in advance? The answers to these questions will determine whether the event's nature is an inherent flaw in the system design or a technical error in parameter calibration. From a legal standpoint, in traditional financial markets forcibly closing a position that is correct in its directional judgment and in a profitable state would immediately provoke legal challenges; the losses to the profitable side caused by the overshoot above constitute quantifiable, concrete damages, and affected traders would in principle have the right to claim compensation. The adequacy and legal validity of the risk disclosure about the ADL mechanism in the exchange's terms of service will be the core factor determining the ultimate direction of this dispute.

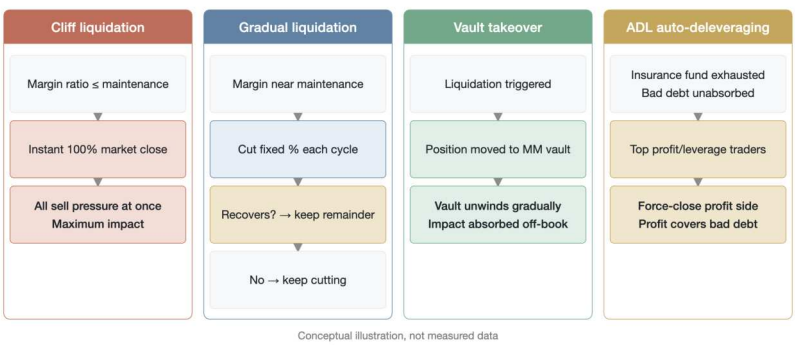


Figure 11-27. Comparison of the execution processes of the four liquidation modes

11.6.4 Strategic liquidation

The inherent design flaws of the liquidation mechanism often give rise to an attack behavior known as "strategic liquidation." Attackers exploit the mechanical nature of the liquidation engine, deliberately manipulating the price to trigger others' liquidations and profiting from it. The feasibility and manifestation of this attack differ across liquidation modes.

Under cliff-edge liquidation, strategic liquidation is most feasible: an attacker need only commit a relatively small amount of capital to push the price to a key support/resistance level, triggering large-scale liquidation orders that then self-reinforce and drive the price far from fair value, so that the attacker can buy at the low and reap a rich profit. By contrast, under progressive liquidation, because positions are released in batches,

an attacker must apply continuous pressure to sustain the liquidation cascade, which markedly raises the cost of the attack and lowers the expected profit potential. The common feature of these two attack modes is that the attacker triggers the mechanical response of the liquidation mechanism through direct price manipulation, differing only in the force required and the duration.

Under the vault-backstop mechanism, the attack vector of strategic liquidation shifts from price manipulation to risk offloading, evolving into "attacking the vault." For example, in the Hyperliquid whale incident of March 12, 2025, a whale trader deliberately built a huge high-risk position and triggered liquidation by withdrawing margin, causing the HLP vault to bear a loss of about $4 million and setting off a subsequent panic withdrawal [24]. The JELLY incident on the 26th of the same month was another, more complex kind of attack: the attacker used the external spot market and oracle transmission to force the HLP vault to take on the short risk of a memecoin, with a peak unrealized loss of about $12 million to $13.5 million, which ultimately turned into a small surplus or a limited loss after validators intervened. Both incidents exploited the structural exposure of the vault to absorbed risk, but their attack paths and profit-and-loss bases differed. Under the ADL mechanism, strategic liquidation manifests as a more insidious distortion of participant behavior: fearing being force-closed by the system, profitable traders may close early or deliberately lower their leverage to avoid the ADL ranking, and this preventive "ADL-avoidance" behavior drains liquidity from the market and instead hastens the market's collapse. Attacks under the vault-backstop and ADL modes both essentially exploit structural loopholes in the risk-redistribution logic of the liquidation mechanism, achieving the attack's aim by distorting the system's incentives; the JELLY-style path additionally layers on external price manipulation and oracle transmission.

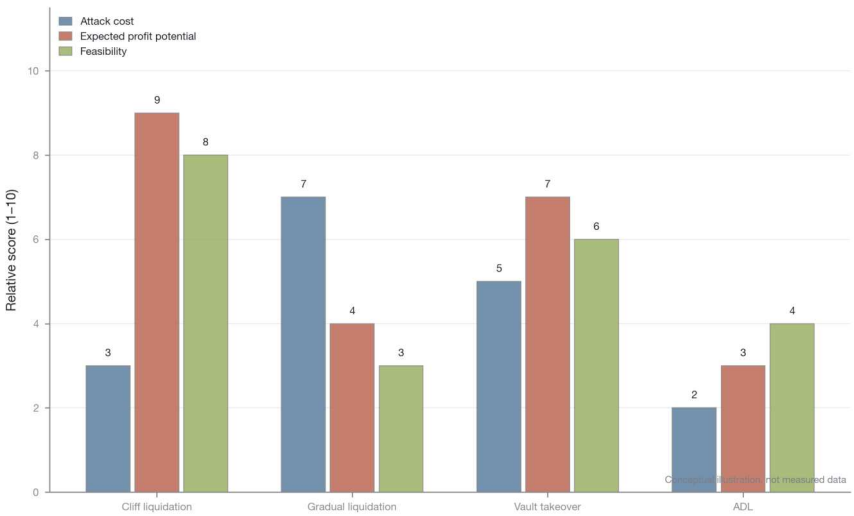


Figure 11-28. Comparison of the cost, profit, and feasibility of strategic liquidation across the four modes (conceptual illustration; qualitative relative values; not empirical data)

The comparison in Figure 11-28 shows that as the liquidation mode evolves from cliff-edge to vault-backstop, the direct feasibility of strategic liquidation declines, but the

attack vector shifts from price manipulation to offloading risk onto the vault. Under the auto-deleveraging mode, the form of strategic liquidation further evolves into a preventive distortion of participant behavior: the profitable side exits the market early to avoid being force-closed, which instead accelerates the contraction of liquidity. This finding shows that the choice of liquidation mode affects not only the system's immediate stability but also the long-run behavioral incentives of participants.

A game-theoretic analysis of strategic liquidation must incorporate the modulating role of legal deterrence as an external constraint. In traditional financial markets, deliberately manipulating the price to trigger others' liquidations falls within the scope of several legal frameworks: U.S. SEC Rule 10b-5 classifies it as a form of securities fraud, the EU Market Abuse Regulation (MAR) expressly lists market manipulation as a criminal offense, and the CFTC has already engaged in enforcement in the crypto space—for example, Avraham Eisenberg's strategic manipulation of Mango Markets in 2022 led to his arrest in December of that year and his conviction by a jury in April 2024 (although this criminal conviction was overturned by the Southern District of New York in May 2025, and only the civil suits by the CFTC and the SEC are currently proceeding—this reversal itself confirms the weakening of legal deterrence by anonymity and cross-jurisdictional arbitrage discussed below). These legal precedents show that legal deterrence markedly alters the game-theoretic equilibrium of strategic liquidation: in a pure game-theoretic model, a rational attacker will choose to attack as long as the expected payoff exceeds the execution cost; legal deterrence, by introducing an expected penalty cost (including fines, imprisonment, and reputational loss), raises the effective cost of the attack and thereby shrinks the parameter space of profitable attacks. This deterrent effect, however, is markedly weakened in crypto markets by anonymity and cross-jurisdictional arbitrage.

In the DEX environment, strategic liquidation faces a distinctive information-structure problem. Because all position data are recorded on a public blockchain, an attacker can precisely compute the liquidation price and maintenance margin level of every account and thus precisely trigger a target liquidation with minimal capital outlay. This "glass box" of complete information transparency contrasts sharply with the privacy of position information at a centralized exchange. On a CEX, an attacker can only roughly infer the location of a liquidation-dense zone through the liquidation heatmap or public aggregate data, whereas the on-chain visibility of a DEX removes this information barrier, making the planning and execution of strategic liquidation more precise and lower-cost.

11.6.5 The nested double trinity

The impossible trinity of liquidation design does not exist in isolation; it is in fact nested within the design trade-offs of the service-provider layer. Chapter 5's service-provider impossible trinity takes "performance, transparency, and decentralization" as its three vertices; for the purpose of characterizing the cross-layer transmission of liquidation and leverage, this section reparameterizes it, from the perspective of systemic risk, into the derived outer triple of "decentralization, capital efficiency, and security" (capital efficiency being expressed as the pursuit of high leverage and high performance, and

security as the guarantee of solvency and verifiability) to constrain the inner liquidation trinity. There is a strong cross-layer tension between these two trinities.

The pursuit of capital efficiency inevitably leads platforms to offer higher leverage. High leverage means traders' safety cushion is extremely thin, which directly makes liquidation events more frequent and larger in scale. This forces the liquidation mechanism to tilt toward certainty in its design, thereby intensifying the conflict within the liquidation trinity. In other words, there is an almost unavoidable causal chain between capital efficiency and system fragility: for every order-of-magnitude increase in leverage, the density of the liquidation-dense zone grows superlinearly, and the squeeze the certainty constraint exerts on fairness and impact minimization intensifies accordingly.

At the same time, the pursuit of decentralization means the system lacks a centralized arbiter with an unlimited balance sheet (such as a traditional clearinghouse) to serve as a backstop. Without the possibility of external capital injection, the system can only resolve bankruptcy through internal rules (such as ADL), which inevitably touches the bottom line of fairness.

These factors point jointly to the conclusion that the pursuit of security requires the system to build a thick insurance fund or market-making vault to safeguard solvency. This inclines the system toward modes such as the vault backstop to achieve impact minimization. As noted above, however, this mode in fact concentrates and transfers risk, and once tail risk exceeds the vault's capacity, the system's overall security instead collapses in an instant. This nested relationship shows that risk management in the crypto derivatives market is an extremely complex systems-engineering problem, in which optimizing any single dimension can trigger unpredictable chain reactions in others.

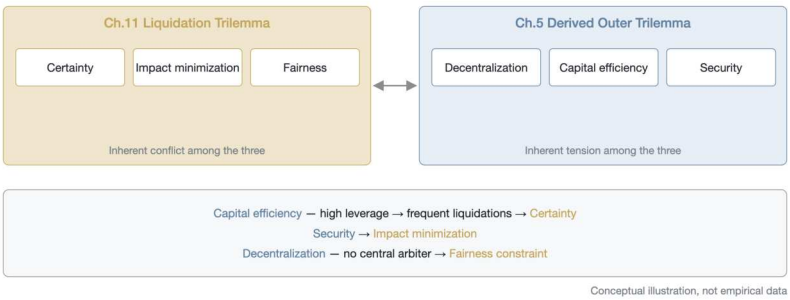


Figure 11-29. The nesting relationship between the impossible trinity of liquidation design and the service-provider impossible trinity (conceptual illustration; the outer trinity's decentralization/capital efficiency/security is a derived reparameterization of Chapter 5's service-provider impossible trinity from the perspective of systemic risk, not Chapter 5's original three vertices of performance/transparency/decentralization)

As shown in Figure 11-29, the parameter choices of the outer trinity (decentralization, capital efficiency, security) directly constrain the feasible region of the inner trinity (certainty, impact minimization, fairness): when the outer trinity tilts toward capital efficiency, the density of high-leverage positions rises, the inner trinity is forced to

tilt toward certainty, and the space for fairness and impact minimization narrows in tandem. This nesting means that risk management in the crypto derivatives market is not a single-dimensional parameter optimization but a systems-engineering problem under multi-layer constraints.

The actual binding boundary of this impossible trinity is being reshaped by external regulatory forces, forming a dynamic process that might be called "regulatory reflexivity." The EU Markets in Crypto-Assets Regulation (MiCA) established the first comprehensive crypto-asset framework, but crypto derivatives themselves are excluded from MiCA and fall under the Markets in Financial Instruments Directive (MiFID II); the European Commission's 2026 review of MiCA is only beginning to examine crypto leveraged contracts and derivatives exposure (as of mid-2026 this remains a regulatory gap rather than an established leverage cap). The UK Financial Conduct Authority (FCA), meanwhile, expressly banned the sale of crypto derivatives to retail investors through PS20/10 (effective January 6, 2021), and Japan and Hong Kong have each set leverage caps. Against this trend, the 125x leverage offered by some exchanges relies to a great extent on regulatory arbitrage—that is, operating in jurisdictions that regulation has not yet reached or where enforcement is weak. Regulatory intervention itself, however, also faces a deep reflexivity paradox: if regulators choose to forcibly lower the maximum leverage or require additional margin during a period of high open interest, this consumer-protection measure may itself become a cascade trigger, as a large stock of positions opened at high leverage is forced to close under the new margin requirements, and the resulting concentrated selling pressure superimposes on the mechanical response of the liquidation engine, an effect dynamically isomorphic with a cascade triggered by an exogenous price shock. There is therefore an inherent conflict in the time dimension between the goal of consumer protection and the goal of preventing systemic risk: in the long run, lowering the leverage cap undoubtedly reduces system fragility; but in the choice of implementation timing, if it falls during a period of high leverage stock, the protective measure may itself create the very systemic crisis it seeks to prevent.

11.7 The time-compression effect

In traditional financial markets, deleveraging is usually a slow process measured in days, weeks, or even months. In the crypto-asset derivatives market, and especially the perpetual futures market, however, this process is extremely compressed to the level of minutes or even seconds. This "time-compression effect" not only fundamentally changes the dynamics of market microstructure but also renders the traditional risk-management paradigm ineffective in the face of crypto markets.

11.7.1 A cross-market comparison

By reviewing the major deleveraging events in financial history, we can clearly see this dramatic change in time scale.

In traditional finance, the 1998 collapse of Long-Term Capital Management (LTCM) took about three months to substantially complete its deleveraging and asset-

liquidation process, with the Federal Reserve's intervention [25]. The financial tsunami triggered by the 2008 bankruptcy of Lehman Brothers took months to years to unwind its complex derivatives positions [26]. Even the high-leverage blowups of recent years, such as the 2021 collapse of Archegos Capital, took about three days to force-close its main positions [27]. In the 2022 UK pension liability-driven investment (LDI) crisis, facing a surge in UK gilt yields, pension funds took about two weeks to top up margin and sell assets [28].

In traditional markets, a standard margin call usually gives a trader T+1 to T+5 business days to replenish funds or close positions on their own. This time-buffer mechanism plays the role of a system stabilizer with a decisive influence on system stability. Take the 1998 LTCM crisis: within a single weekend, the Federal Reserve organized a consortium of 14 banks and, after a full three days of intensive negotiation, ultimately arranged $3.6 billion in rescue financing. This crisis-response process, centered on manual coordination, was extremely slow by the standards of crypto markets, yet it gave the parties ample room for information exchange and negotiation of interests, allowing a coordinated systemic rescue to be formed. This response mode, which relies on human intermediation, however, has no structural basis in a permissionless market: there is no central coordinator to convene the parties, no enforceable negotiation framework, and no legal force to bind the parties to honor rescue commitments.

By contrast, in the crypto derivatives market, during the "3.12" crash of 2020 the market's deleveraging was mostly concentrated within 24 hours. By the extreme liquidation event of October 10, 2025, according to market-data analytics firms, nearly $20 billion in leveraged positions were liquidated within a single day, as described at the opening of this chapter (see Section 11.9.2 for the complete microstructure data), while the most central destruction was concentrated within a 40-minute window, and the peak liquidation rate accelerated 86-fold above the normal level [2]. In this same crisis, according to Chitra's preprint, Hyperliquid's auto-deleveraging mechanism force-closed about $2.1 billion of positions within about 12 minutes (this was the first cross-margin ADL trigger in its operating history) [18].

From the months and weeks of traditional finance to the minutes and tens of seconds of crypto markets, the time scale of deleveraging has been compressed by roughly 10^4 to 10^5 times (this is a mid-range estimate based on a traditional typical of T+3 to a few weeks and a crypto typical on the order of minutes; taking the extreme endpoints of months $\rightarrow$ seconds yields about 10^6 times, and days $\rightarrow$ minutes about 10^3 times).

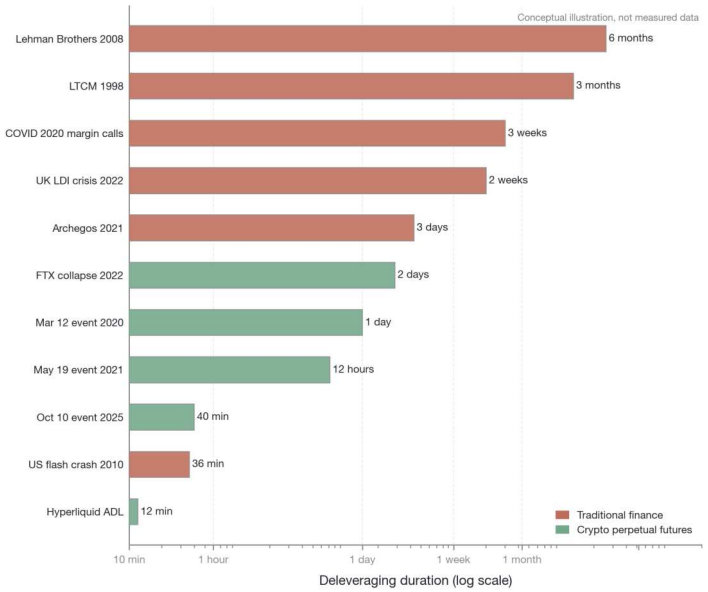


Figure 11-30. Comparison of the deleveraging timeline between traditional finance and perpetual futures (logarithmic scale; deleveraging durations are an illustrative comparison of typical orders of magnitude, not precise measurements)

11.7.2 Four driving forces

This extreme time compression is no accident but is jointly determined by the underlying architecture and rule design of the crypto derivatives market. Specifically, four core driving forces produce this phenomenon:

Around-the-clock operation with no buffer mechanism means the crypto market trades continuously, 24/7/365, with no close, no weekend break, and no official price limits or circuit breakers. There is also a fundamental difference between the price-anchoring mechanisms of perpetual futures and traditional futures: traditional futures have a hard, discrete anchor (delivery at expiry forces futures and spot to converge), whereas perpetual futures rely only on the funding rate as a soft, continuous anchor that can fail or even reverse under extreme stress (such as the funding-rate-reversal risk analyzed in Section 11.4), allowing a basis deviation to persist indefinitely in the absence of a hard delivery constraint, so that panic and liquidation selling pressure in a crisis cannot be cooled by a market close. Automated liquidation hard-codes the liquidation logic into the matching engine or an on-chain smart contract, and once the oracle price feed reaches the maintenance margin threshold, it triggers closure unconditionally within milliseconds, with none of the traditional market's room for manual notification, negotiation, and grace. The contagiousness of cross-margin arises because a unified account allows the paper profit of one asset to collateralize the loss of another, raising capital efficiency in calm times but, in extreme conditions, making portfolio risk determined by the highest-risk constituent asset—as in the October 2025 event, when the plunge of long-tail assets such as WLFI dragged down the account's overall margin ratio and triggered the forced sale of core assets such as Bitcoin and Ethereum [27], greatly accel-

erating the liquidation process. The transparent order book and the liquidity vacuum arise because liquidation lines can be precisely anticipated by professional market makers: foreseeing the arrival of large-scale liquidations, rational market makers instantly withdraw their buy orders to avoid "toxic orders," so that on October 10 the liquidity of key trading pairs evaporated by more than 99% within minutes and spreads widened by three orders of magnitude (see Section 11.9.2, Table 11-7) [2], and a small number of liquidation orders could cut through the remaining depth and trigger deeper liquidations.

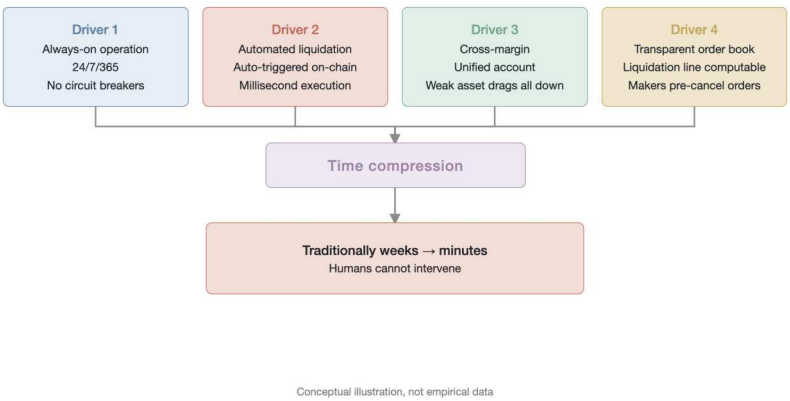


Figure 11-31. The four driving forces of time compression and their transmission mechanisms

11.7.3 The mismatch between decision-making and execution

The time-compression effect brings a serious consequence: a fundamental mismatch between the latency of human decision-making and the speed of machine execution.

In the traditional risk-control framework, after receiving an alert of abnormal market movement, a risk-management team must go through a series of steps—data verification, risk assessment, decision-making, fund allocation, and order issuance. Even for a top-tier, well-trained institution, this process takes at least tens of minutes to hours. Even an algorithmic trading firm that has deployed pre-programmed risk-control thresholds usually needs 30 to 60 seconds in extreme conditions to judge whether an abnormal move is a data error, a local exchange anomaly, or a genuine market event—and this delay already exceeds the entire duration of the most destructive phase of the 10.10 cascade crisis. We roughly estimate that a human risk-control team's ceiling for effective response and fund deployment in an emergency is about $50 million per hour.

As the October 10, 2025, event demonstrated, however, during the eruption of the liquidation cascade the system's equivalent liquidation rate reached $10.39 billion per hour; and within the peak 60 seconds it reached the extreme equivalent rate of about $192.6 billion per hour (this is the equivalent rate obtained by linearly extrapolating the peak-minute liquidation volume to an hourly basis; it lasted only about one minute and is not a sustainable rate). This means the speed of machine liquidation was hundreds or even thousands of times faster than the fastest human response.

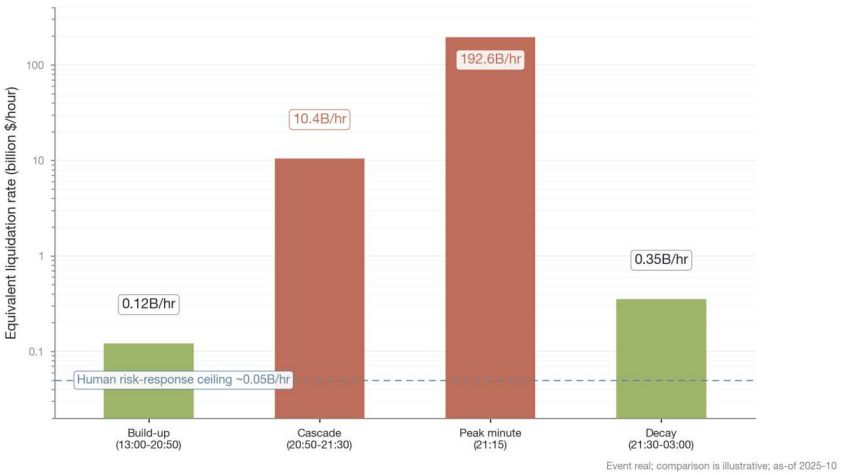


Figure 11-32. Comparison of the liquidation rate and the decision-response speed in the "10.10 event" (Data source: Amberdata [2], as of October 2025; the liquidation rate has multiple bases that should not be conflated: about $3.21 billion per minute at the peak minute, about $10.39 billion per hour averaged over the 40-minute cascade, and about $192.6 billion per hour from the linear extrapolation of the peak minute, which lasted only about 1 minute and is not a sustainable rate; the human response latency is an illustrative representative value)

11.7.4 Implications for risk control and regulation

The time-compression effect is not only a technical feature of market microstructure; it poses far-reaching challenges to the risk-control methodology and future regulatory path of the crypto derivatives market.

The definition of "effective leverage" must be reconstructed. Traditional risk-control models often calculate the safety cushion on the basis of static margin rates and historical volatility. In an environment of time compression and instantaneous liquidity evaporation, however, nominal leverage is meaningless. The real risk depends not only on directional exposure but, more importantly, on the liquidity depth of the collateral basket under extreme conditions. As research points out, "the market trades at the margin, not at the mean" [27]. Risk-control models must introduce dynamic, nonlinear liquidity-penalty factors to simulate whether a position is still safe in a scenario where order-book depth has shrunk by more than 90%.

At the same time, guarding against single points of failure directly determines the system's survivability. When infrastructure becomes unreliable in extreme conditions, concentrating all positions, collateral, and oracle price feeds at a single exchange (even a decentralized one) becomes a fatal single point of failure. For example, when an exchange's internal stablecoin price depegs because of a local liquidity drought, a liquidation engine based on that local price will mercilessly liquidate a cross-market hedge portfolio that would be entirely healthy from a global perspective [27]. Multi-centered oracle price feeds, redundant allocation of funds across exchanges, and automated programmatic hedging therefore become must-haves for the survival of professional institutions.

From the standpoint of systemic risk and regulation, the time-compression effect requires us to rethink the applicability of circuit breakers in crypto markets. Although the pure spirit of decentralization rejects any human intervention in trading, when the execution speed of the liquidation engine far exceeds the market's capacity for self-repair, some form of algorithm-level circuit breaker or volatility-based dynamic liquidation-delay mechanism may be a necessary compromise to prevent a "flash crash" from evolving into a systemic crisis. The actual effect of circuit breakers in traditional markets, however, is far from uncontested. The research of Goldstein and Kavajecz (2004) [29] reveals the "magnet effect" of circuit breakers: when the price approaches the circuit-breaker threshold, traders expect trading to be suspended and therefore accelerate their selling to avoid being locked into an unfavorable position, and this behavior instead pushes the price toward the circuit-breaker line faster, making the circuit breaker itself an amplifier of volatility rather than a damper. The theoretical root of this "magnet effect" can be traced to Subrahmanyam's (1994) [30] early analysis of how a circuit-breaker threshold distorts traders' intertemporal order placement. In the crypto market's fragmented exchange environment, this problem is further complicated: if a single exchange unilaterally suspends trading, its immediate effect is to drive order flow to competing platforms that are still trading, rather than genuinely arresting the cascade, and the selling pressure does not disappear because one exchange has suspended trading but is merely released in concentration on other platforms. A market-wide coordinated suspension could in theory solve this problem, but it is in deep conflict with the decentralized nature and permissionless operating principle of crypto markets. A more feasible alternative is therefore a dynamic liquidation-throttling mechanism: rather than suspending trading itself, it lowers the execution rate of the liquidation engine when cascade conditions are detected (for example, limiting the volume of liquidation orders processed per unit time to a certain proportion of the available order-book depth), thereby slowing the acceleration of the positive-feedback loop while keeping the market open. This design provides a time window for liquidity to recover and market makers to re-enter, without interrupting the price-discovery function.

11.7.5 The structural protections of traditional margin markets

The extremity of the time-compression effect in the crypto derivatives market becomes even sharper when contrasted with the structural protection mechanisms of traditional margin markets. Through decades of institutional evolution, traditional financial markets have built multi-layered braking mechanisms that jointly constitute a structural buffer preventing the deleveraging process from running out of control.

On liquidity depth, the order-book depth of traditional futures markets is usually tens or even hundreds of times that of crypto markets. Take the S&P 500 index futures of the Chicago Mercantile Exchange (CME): their visible order-book depth in normal times can reach the level of billions of dollars, and even in the extreme conditions of March 2020 they maintained a minimum depth of hundreds of millions of dollars. A liquidity pool of this magnitude means that the price impact produced by one or a few large liquidation orders on execution is effectively diluted, and a self-reinforcing cascade feedback is hard to form. A cross-market comparison of absolute depth should be standardized by open interest or average daily volume to yield a meaningful relative-

liquidity measure. Although the absolute depth of the crypto perpetual futures market is far lower than that of S&P 500 futures, its depth ratio relative to OI may reveal a more precise measure of fragility. An especially analytically valuable comparison case is CME Bitcoin futures: they trade the same underlying asset as crypto perpetual futures but run on traditional market infrastructure, with well-developed circuit breakers, regulated market-maker obligations, margin grace periods, and a clearinghouse's multi-layer default waterfall. Comparing the price behavior of CME Bitcoin futures with that of crypto perpetual futures in the same macro-shock event therefore makes it possible to isolate the net contribution of market microstructure differences to cascade dynamics while effectively controlling for the underlying asset's risk factors.

Circuit breakers are the most direct means by which traditional markets arrest time compression. When a price falls by a preset threshold within a short time (such as 5%, 7%, or 13%), the exchange forcibly suspends trading for 5 to 15 minutes. This pause gives market participants a time window to reassess information, add margin, and adjust strategies, effectively interrupting the positive-feedback loop between panic selling and mechanical liquidation.

In traditional markets, the prime broker acts as a credit-intermediary layer between traders and the clearinghouse. When a client faces margin-call pressure, the prime broker can provide short-term financing grace based on the client's overall credit standing and long-term relationship, allowing the client to adjust positions in an orderly way over hours or even days rather than being forced to close immediately in the worst market conditions. This elastic layer, constituted by interpersonal relationships and credit judgment, is entirely absent from the purely algorithm-driven crypto market.

In addition, the asset diversification of a traditional-market portfolio is far greater than in crypto markets. The correlations among equities, bonds, commodities, and foreign exchange are low in normal times, and even when correlations rise in a crisis, they are unlikely to reach close to 1 simultaneously. By contrast, the correlations of the vast majority of crypto assets with Bitcoin show an empirical tendency to converge rapidly toward 1.0 during a crisis. Although the precise correlation values vary by event, time window, and asset class, this high directional synchrony recurs across successive major cascade events. Correlation, however, measures consistency in the direction of price movements, not symmetry in their magnitude: in the 10.10 event, BTC's maximum drawdown was 14.5%, while the drawdown depth of some DeFi tokens reached more than 70%. The apparent contradiction between high correlation and starkly different drawdown magnitudes can be reconciled by the following mechanism: all assets fell in the same direction (hence the high correlation), but the less liquid assets bore a disproportionate price impact because of their shallower order-book depth, higher leverage concentration, and the faster withdrawal of market makers. This means that under the cross-margin system there is almost no effective risk diversification: the entire portfolio behaves, in direction, like a single high-leverage position in one asset, while in magnitude it is dominated by the least liquid constituent asset.

11.8 The information paradox and mechanism-induced volatility

A liquidation event carries both information and noise in the order book, and market makers cannot separate the two; the systemic overreaction that results constitutes an efficiency constraint peculiar to perpetual futures markets. This section first defines the information paradox of liquidation, then analyzes the adverse-selection dilemma of market makers, and finally proposes the concept of mechanism-induced volatility and its differences across modes.

11.8.1 The information paradox of liquidation

In the market microstructure of perpetual futures, a liquidation event constitutes a distinctive cognitive puzzle, which we call the information paradox of liquidation. This paradox is not only an extension of microstructure theory but also a core key to understanding the extreme volatility of crypto-asset markets. When we observe a liquidation order in the order book, it in fact carries two starkly different attributes at once: an information attribute and a noise attribute. The research of Bouchaud, Farmer, and Lillo shows that order flow's impact on price has a long-memory character, and the influence of large one-sided order flow accumulates over time [31]. The information attribute originates in the liquidated trader's misjudgment, but this statement requires a more precise classification. Liquidated positions can in fact be distinguished, by their information content, into three starkly different cases. The first is a genuine fundamental misjudgment: the trader was indeed wrong about the direction of the asset's value, the price movement reflects a negative update to fundamentals, and in this case the liquidation order carries valid information about the asset's true value, on which market makers should adjust their expectation of fair value. The second is a correct fundamental judgment but excessive leverage: the trader's medium-to-long-term judgment about price direction may be entirely correct, but because they used too high a leverage multiple, even a normal 2% market fluctuation is enough to trigger their maintenance margin line; this liquidation is pure leverage noise and carries no fundamental information. The third is an unrelated liquidation caused by cross-margin contagion: under the unified margin system, a trader's loss on some unrelated asset erodes the account's overall margin balance, causing forced closure of a position in another instrument on which the trader's directional judgment was correct and whose fundamentals have not deteriorated. This third case constitutes pure structural noise; the liquidation is entirely unrelated to any fundamental information about the asset and is merely a by-product of the margin system's design. In the actual operation of the crypto derivatives market, the latter two cases (leverage noise and structural noise) very likely dominate in number, which means the signal-to-noise ratio in liquidation order flow is far lower than in the informed-trading-driven order flow of traditional markets. The core difficulty is that these two attributes manifest identically in the order book. Whether it is a desperate sale caused by deteriorating fundamentals or a mechanical close triggered by excessive leverage, both appear in the matching engine only as ruthless market sell orders. At centralized exchanges, liquidation orders are in fact technically identifiable: Binance's forceOrder WebSocket stream, for example,

lets market participants distinguish liquidation orders from voluntary orders in real time. The real paradox market makers face, therefore, is not the inability to identify a liquidation order itself but the inability, even knowing that an order is a liquidation, to decompose it into signal and noise: which of the three cases above does it belong to? Does it reflect valid fundamental information, is it the mechanical consequence of excessive leverage, or is it a structural by-product of cross-asset contagion? This deep information asymmetry places market makers in a dilemma.

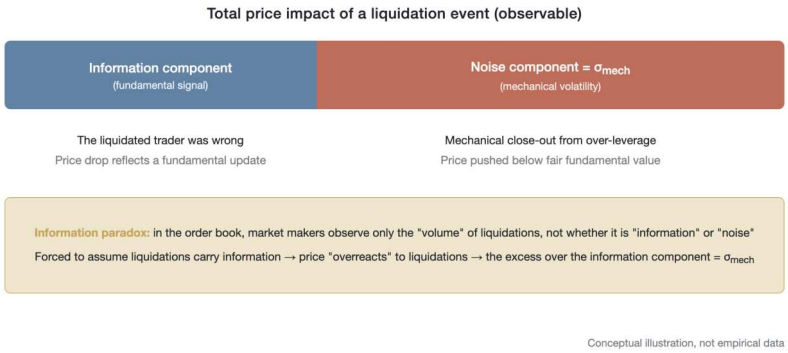


Figure 11-33. Decomposition of the information and noise superimposed in a liquidation order (conceptual illustration, not empirical data)

11.8.2 The adverse-selection dilemma of market makers

Under the microstructure framework proposed by Kyle (1985) [10], when facing informed traders, market makers can infer the information content in order flow from its characteristics and adjust their quotes accordingly. In the Kyle model, market makers gradually adjust the price to the level that reflects the asset's expected value by observing the size and timing characteristics of orders, and the premise of this inference process is that there is a strategic information motive behind the orders. Liquidation orders, however, carry no such strategic information; they are mechanically triggered by the exchange's risk engine at preset price levels and lack the tentative or strategic character of an informed trader, so the standard information-inference framework loses its basis for application. Faced with this massive volume of orders that cannot be distinguished as information or noise, market makers are forced to adopt an extremely conservative strategy. To protect themselves from potential adverse-selection risk, market makers must assume that at least some of these liquidation orders contain unknown negative information.

This forced conservative assumption leads to a systemic consequence: the overreaction of prices to liquidation events. To avoid risk, market makers widen bid-ask spreads substantially and withdraw liquidity, so that liquidation orders, in cutting through the order book, cause a decline far greater than their information content warrants. This rational self-protection by market makers manifests at the macro level as an irrational market plunge. When the dust settles and we look back afterward, we find that the

portion of the price decline beyond the information component was purely manufactured by the liquidation mechanism itself, and we define this as mechanism-induced volatility.

11.8.3 Mechanism-induced volatility

Mechanism-induced volatility is price volatility produced by the liquidation mechanism that reflects no change in fundamental information; it is the core indicator of the degree to which the liquidation system interferes with the market's price-discovery efficiency. This concept is logically similar to the excess volatility Shiller (1981) [32] identified in the stock market, but with one key distinction: the excess volatility Shiller described originates in investors' behavioral biases during valuation, whereas mechanism-induced volatility in the crypto derivatives market is produced by the structural interaction between leverage and automated liquidation, which makes it a predictable and, in principle, measurable component of total price variance. The sources of mechanism-induced volatility can be further decomposed into two components that are conceptually distinguishable but closely intertwined in practice: (a) mechanical-liquidation overshoot, the nonlinear price impact produced by the liquidation engine's forced sell orders when liquidity is thin, a purely algorithm-driven, information-free price deviation; and (b) behavioral-panic overshoot, the panic selling that traders who have not been liquidated undertake voluntarily out of fear after observing a sharp price drop and large-scale liquidations. Take the 10.10 event: at the peak moment 93.5% of sell orders were market orders produced by forced liquidation, which means about 6.5% of the selling pressure came from voluntary panic selling. Although 6.5% seems small, when order-book depth has already collapsed to below 1% of its normal level, even this small portion of voluntary selling produces a significant additional price impact. In the strict sense, "mechanism-induced volatility" should cover only component (a), the price deviation directly manufactured by the liquidation engine; component (b) belongs to the behavioral chain reaction and is essentially a second-order response to mechanical liquidation. As a preliminary measure, we can compare price volatility during liquidation-intensive periods with that during non-liquidation periods and, after controlling for other factors such as macro news, attribute the additional volatility difference to mechanism-induced volatility.

The magnitude of mechanism-induced volatility is closely related to the liquidation mode an exchange adopts. Because cliff-edge liquidation dumps the entire position onto the market at market price in an instant, it injects the most noise and produces the most violent mechanism-induced volatility. Progressive liquidation, by reducing positions in batches, injects the noise dispersed over a longer time window, so its mechanism-induced volatility is at a moderate level. Although the auto-deleveraging mechanism is extremely unfriendly to the profitable counterparty, its precise, targeted closing has limited impact on the public market price, so its mechanism-induced volatility is small. And vault-backstop liquidation internalizes the risk position so that the noise does not enter the public order book at all, producing the least mechanism-induced volatility.

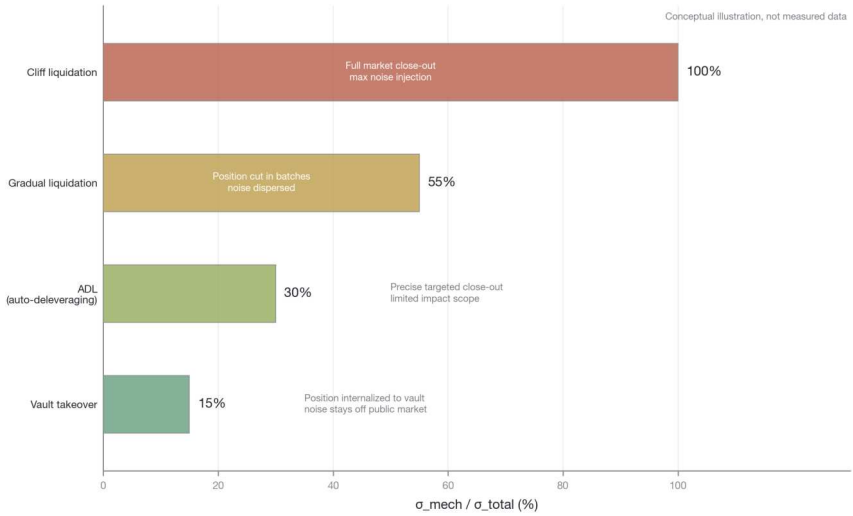


Figure 11-34. The relative magnitude of mechanism-induced volatility under different liquidation modes (conceptual illustration; qualitative relative values; not empirical data)

The information paradox and mechanism-induced volatility have important implications for the price-discovery efficiency and market quality of perpetual futures markets. It shows that during liquidation-intensive periods, the price of perpetual futures systematically deviates from the asset's fundamental value. Mechanism-induced volatility in fact constitutes a hard ceiling on the informational efficiency of perpetual futures. However precise the oracle, as long as high leverage and forced-liquidation mechanisms exist, this endogenous mechanism-induced volatility cannot be eliminated. When assessing the quality of crypto-asset markets, therefore, we must incorporate mechanism-induced volatility into our analytical framework as an efficiency constraint peculiar to perpetual futures. It reminds us that the price seen in extreme conditions often reflects the fragility of the market structure more than the asset's true intrinsic value.

11.9 A reflexivity dissection of the "10.10 event"

The "10.10 event" is a complete empirical test of this chapter's theoretical framework. This section reconstructs, in turn, the system's pre-crisis fragility, the micro process of the trigger and cascade, the cross-asset contagion path, and the empirical validation of the four-phase model and mechanism-induced volatility.

11.9.1 The system's pre-crisis fragility

The crypto-asset market crash that erupted on October 10, 2025—the "10.10 event"—provides a perfect empirical test scenario for all the theoretical tools discussed in this chapter. It was the largest single-day liquidation event in the history of crypto derivatives, with total 24-hour liquidations estimated by institutions at more than three times those of "Black Thursday" in March 2020 [17] [3] (note: sources differ substantially in how they measure the total liquidations of the March 2020 event; the complete microstructure data appear in Section 11.9.2). Many market commentators simply

blamed it on the shock of macro geopolitical news, but a close dissection of the event's microstructure reveals a more fundamental truth. The macro news was merely the exogenous catalyst that triggered the cascade; the fragile leverage structure inside the crypto-asset market, the cross-contagion of the unified margin system, and the mechanical execution of the liquidation engine were the fundamental causes that amplified a controllable price pullback into a systemic implosion.

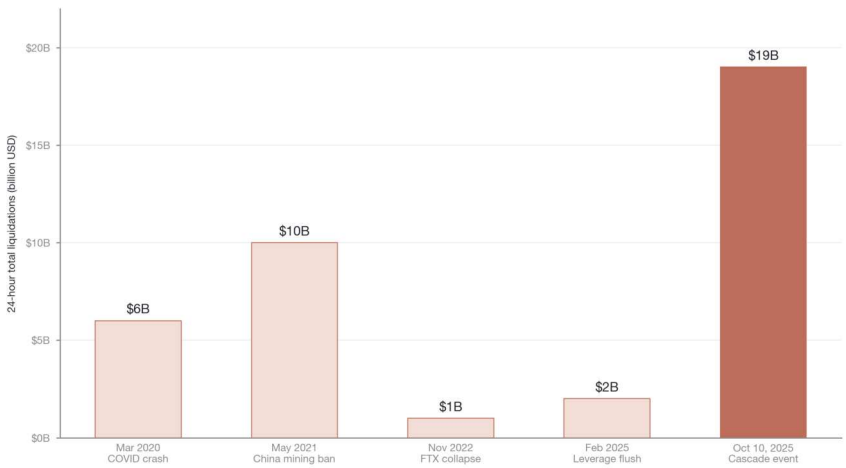


Figure 11-35. Comparison of the scale of historical liquidation events (Data source: CoinGlass [9], as of October 2025; the values are representative rounded milestones—China's mining ban of May 2021 was about $8 billion to $10 billion, with CoinGlass's canonical figure about $8.6 billion; because of API reporting caps, the CoinGlass basis historically underestimates and its absolute figures are conservative, so orders-of-magnitude comparison should be emphasized)

On the eve of the crisis, the crypto-asset market was already in a highly taut metastable state. Total open interest in perpetual futures at major venues had reached a record of about $146.6 billion (on Amberdata's basis, not a full market-wide basis) and displayed an extreme long-dominated character [2]. Because of the market's preceding sustained rise, one-sided crowding was extremely high, the ratio of open interest to market cap had climbed to a historical high, and positions using high leverage above 25x accounted for a significant share. More dangerously, the funding rate had already completed the process of “accumulating fragility.” Before October 6, according to FTI Consulting's post-crash analysis, the funding rate on Bitcoin and Ethereum perpetual futures (driven mainly by the Ethereum market) rose from about 10% annualized to nearly 30% (corresponding to about 0.027% per 8 hours; the figure does not specify whether it is a market-wide weighted or a single-exchange basis) [27]. This extreme positive rate meant that longs had to keep paying shorts a high holding cost, which imperceptibly kept consuming longs' maintenance margin and systematically pulled their liquidation prices upward, making the entire market extremely sensitive to downside risk. The liquidation heatmap showed that a massive number of liquidation trigger points had clustered in the narrow band 5% to 10% below the price at the time, and the cascade risk index had entered a dangerous divergence range.

The macro event that triggered this cascade (the announcement of a 100% tariff on Chi-

nese imports) did not act on the crypto-asset market alone; it simultaneously struck traditional risk assets such as equities, foreign exchange, and commodities. During the six-hour window in which Bitcoin fell 6.8% before the cascade erupted, macro hedge funds rebalanced their portfolios by treating crypto-asset exposure as part of a broader risk-reduction strategy, and this cross-market outflow very likely contributed significantly to the initial selling pressure. Analyzing the co-movement of S&P 500 index futures, the USD/CNY exchange rate, and the gold price over the same time window would help decompose the initial trigger factor into a crypto-market-endogenous component and a cross-market capital-flow component, though such an analysis is beyond the scope of this chapter.

11.9.2 The trigger and the cascade

The crisis was triggered at 14:57 UTC on October 10, when U.S. President Trump announced on social media that a 100% tariff would be imposed on Chinese imports [3]. This macro catalyst immediately set off a global sell-off in risk assets. The price of Bitcoin began to slide slowly from a high of $121,500 and fell to around $114,000 over the following six hours, a decline of about 6.8% [2]. Although painful, the decline in this phase was still within the range of normal price discovery, and the market showed no sign of systemic collapse. This six-hour decline, however, gradually eroded the margin buffer of high-leverage positions and pushed the price toward the liquidation-dense zone.

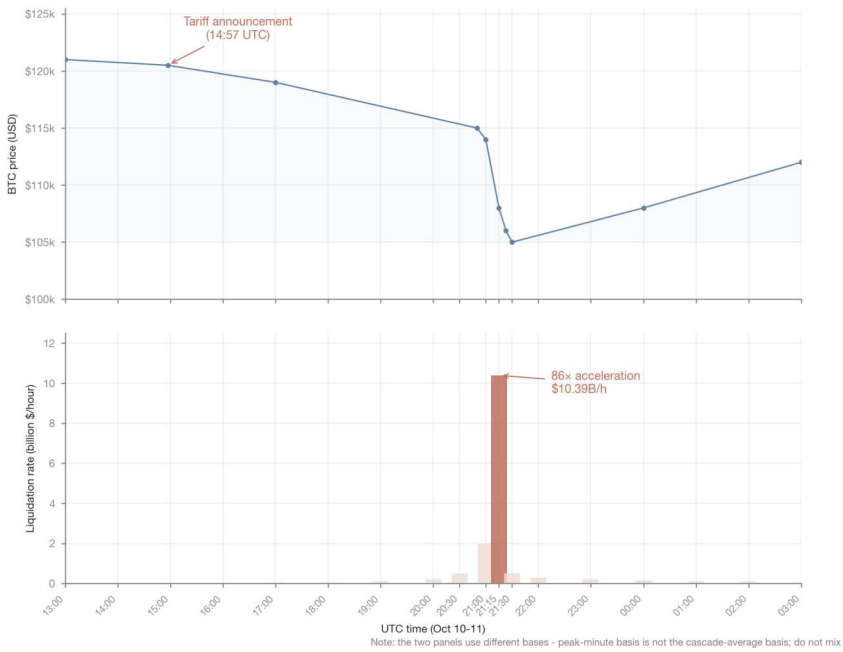


Figure 11-36. Timeline of price and the liquidation rate in the "10.10 event" (Data source: Amberdata [2], as of October 2025)

The cascade crisis began at 20:50 UTC, when the price hit a key liquidity-vacuum zone and the liquidation cascade was formally triggered. Over the following 40 minutes,

the market experienced the fastest mechanical deleveraging in the history of crypto derivatives. The liquidation rate surged instantly from a mild $120 million per hour before the cascade to $10.39 billion per hour, an acceleration of a full 86-fold [2]. In the peak minute at 21:15, as much as $3.21 billion of positions were liquidated within 60 seconds, of which 93.5% were forced closures of longs [2]. This purely algorithmic execution speed completely deprived human traders of any time window to intervene or add margin.

Table 11-7 summarizes the changes in four key microstructure indicators before and after the cascade, showing the comprehensive deterioration of the market infrastructure within 40 minutes.

Indicator	Before cascade (8-hour average)	During cascade (40-minute average)	Magnitude of change
Liquidation rate	$120 million/hour	$10.39 billion/hour	Accelerated 86x
Order-book depth	$103.6 million	$170,000	Contracted 99.8%
Bid-ask spread	0.02 bps	26.43 bps (peak)	Widened 1,321x
Open interest	$146.6 billion	$109.9 billion	Evaporated 25%

Table 11-7. Comparison of key microstructure indicators before and after the cascade in the “10.10 event” (Data source: Amberdata [2])

In this 40-minute liquidity black hole, the microstructure data reveal the complete collapse of the market infrastructure. According to Amberdata, the visible order-book depth of major exchanges fell off a cliff from $103.64 million to a mere $170,000, and more than 99.8% of liquidity contracted sharply precisely when it was most needed [2]. The pre-cascade depth figure of $103.64 million here reflects visible resting limit orders. In practice, the effective liquidity of the crypto market also includes substantial hidden depth (iceberg orders and algorithmic market makers quoting with high-frequency placements and cancellations), and under normal market conditions this hidden liquidity is typically 3 to 5 times the visible depth. During a cascade, this hidden liquidity disappears even faster than visible orders, because algorithmic market makers pause their quoting algorithms within milliseconds once they detect adverse-selection flow. Therefore, from total liquidity including hidden and refresh components down to the residual $170,000, the collapse in actual effective depth was very likely close to 99.95%, which makes the cascade dynamics more extreme than the visible order-book data alone would present. At the same time, the bid-ask spread of Bitcoin perpetual futures widened explosively from an extremely tight 0.02 basis points to 26.43 basis points, an increase of 1,321-fold [2]. This deterioration of the microstructure meant that every new liquidation order, in cutting through the order book, produced far greater price slippage than usual, thereby triggering more liquidations at lower prices and forming a classic positive-feedback spiral.

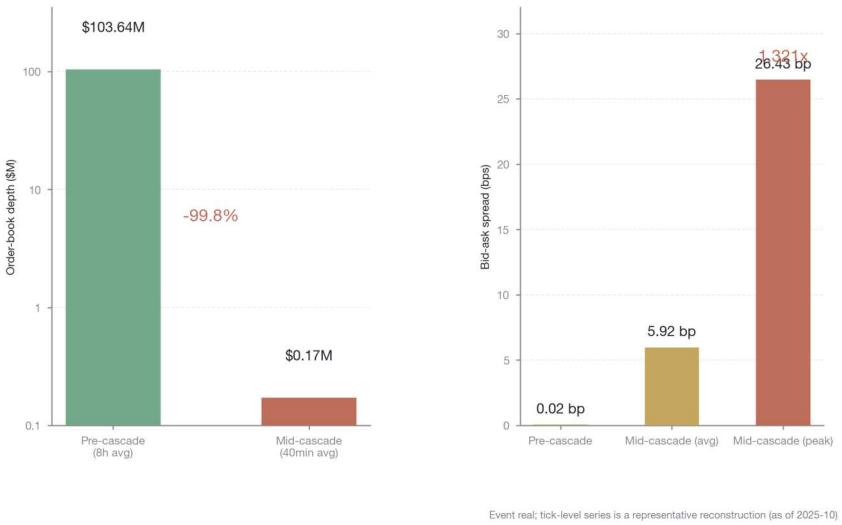


Figure 11-37. Order-book depth and the bid-ask spread in the "10.10 event" (Data source: Amberdata [2], as of October 2025; the endpoints—depth of about $103.6 million → $170,000 and spread of 0.02 → 26.43 basis points—are Amberdata's measured top-of-book values for BTC perpetual futures, while the point-by-point series is a representative reconstruction)

11.9.3 Cross-asset contagion and collateral collapse

The destructiveness of this crisis was reflected not only in the plunge of a single asset but, even more, in its ultra-rapid cross-market, cross-asset contagion. The unified margin system played the role of the key amplification mechanism here. When the fall in Bitcoin's price triggered liquidations, an exchange's risk engine not only closed Bitcoin positions but also indiscriminately sold off the high-risk altcoins in the account to make up the margin shortfall. As a result, although Bitcoin ultimately fell only about 14.5%, Ethereum fell 12.2%, while the maximum drawdown of DeFi blue-chip assets such as UNI and AAVE approached 70% [2]. This asymmetric cross-asset shock carries important structural information. The DeFi tokens' decline far exceeding Bitcoin's was very likely the result of several superimposed factors: compared with Bitcoin perpetual futures, altcoin perpetual futures have markedly thinner order-book depth, which means an equal amount of liquidation selling pressure produces a proportionally amplified price impact; leverage concentration on altcoin perpetual positions is higher; DeFi tokens are disproportionately used as collateral in unified margin accounts, thereby triggering collateral-driven forced selling; and market makers withdraw liquidity from illiquid pairs faster and more completely. A formal quantitative decomposition of these contributing factors would substantially advance our understanding of the structure of cross-asset contagion.

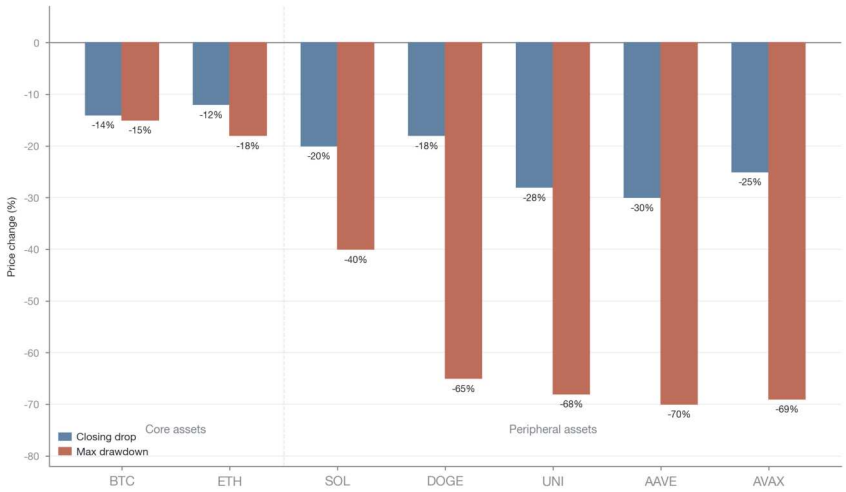


Figure 11-38. Comparison of price declines across multiple assets (as of October 2025; “closing decline” and “maximum drawdown/intraday low” are two different bases that should not be read interchangeably; SOL’s closing decline is a representative rounded figure, with a verified value of about −24% and an intraday low that at one point exceeded −40%; the specific declines for each asset follow the main text of Section 11.9.3)

Collateral collapse was the most striking link in this cross-market contagion. As a high-yield synthetic dollar stablecoin, USDe was used by many traders as margin for contract trading at the time [3]. When the liquidity crisis erupted, USDe encountered severe selling pressure on Binance, and its price briefly depegged and plunged to $0.65, a discount of as much as 35% to its $1 peg [3]. This was mainly a local liquidity event at the single venue of Binance: over the same period, USDe deviated by less than 100 basis points on Curve, fell only to about $0.92 on Bybit, and decentralized venues largely maintained the peg on hundreds of millions of dollars of liquidity; the Ethena protocol maintained about $66 million in excess collateral throughout, its minting and redemption functions operated normally, and it processed more than $2 billion in redemptions within 24 hours, showing that the 35% discount did not reflect the true solvency risk of its delta-neutral strategy but was a pricing distortion in a single low-liquidity trading pair [3]. Because the exchange’s margin system directly used the local internal oracle price feed, the plunge in USDe left countless previously safe accounts instantly insolvent, setting off a fierce secondary liquidation wave. Liquid staking tokens such as BNSOL and WBETH met a similar fate, and this double blow—collateral value and asset price plunging at the same time—constituted two mutually nested positive-feedback loops.

The collapse of USDe exposed the structural risk of using a yield-bearing stablecoin as margin collateral. Unlike traditional stablecoins (such as USDT and USDC) backed by fiat reserves or short-term Treasuries, USDe maintains its peg through a delta-neutral strategy (holding a long position in spot crypto assets—mainly BTC and ETH and their liquid staking derivatives—and shorting an equivalent position in the corresponding perpetual futures, plus some stablecoin and cash reserves). This means that USDe’s stability itself depends on the normal functioning of the perpetual futures market, which is precisely the link that collapsed first in the crisis. When the 10.10 event erupted,

the short hedge positions underlying USDe faced huge funding-rate payments (as the rate turned sharply negative), and its redemption mechanism could not be executed in time when market liquidity dried up, producing a discount of as much as 35% between the secondary-market price and the theoretical net asset value. The exchange failed to lower USDe's margin collateral factor in time during the event, meaning that USDe valued 1:1 was still counted toward margin at face value, and the system responded only passively after the price had collapsed; here a distinction should be drawn between "no haircut applied at all" (treating it as equivalent to USDT) and "a haircut that existed but was insufficient," two cases with fundamentally different implications for the direction of risk-management recommendations. This lag reveals a fundamental flaw in the current margin system's assessment of collateral risk: adjustments to the collateral factor are discrete and lagging, whereas changes in collateral's market value are continuous and instantaneous. As noted, USDe's depeg of as much as 35% was mainly a pricing distortion in a single low-liquidity Binance trading pair rather than Ethena's true solvency risk; if so, then the exchange's internal pricing oracle in fact constituted a single point of failure—using a single low-liquidity trading pair as the collateral-valuation mark for the entire unified margin system—and its failure to adjust the collateral factor in real time when the depeg occurred also raises serious questions about the exchange's duty of prudent care and the legal recourse of harmed users. For any yield-bearing asset ultimately backed by a derivatives strategy, its suitability as margin collateral needs to be re-examined on the basis of liquidity and redemption capacity under extreme stress scenarios, rather than relying only on historical volatility under normal market conditions.

At the deepest point of the crisis, the procyclical effect of the auto-deleveraging mechanism dealt a further blow to the market. When the risk reserve was exhausted and not enough liquidity could be found in the market to take on the liquidated long positions, the exchange activated auto-deleveraging. This mechanism forcibly closed short positions that were correct in their directional judgment and profitable. This not only inflicted serious losses on the professional market makers running delta-neutral strategies but, more alarmingly, eliminated the largest potential buying force in the market—because these shorts would otherwise have bought to close as the price fell to low levels, providing precious bottom liquidity to the market. The intervention of auto-deleveraging further weakened the market's remaining liquidity, causing the price to stall further at the bottom.

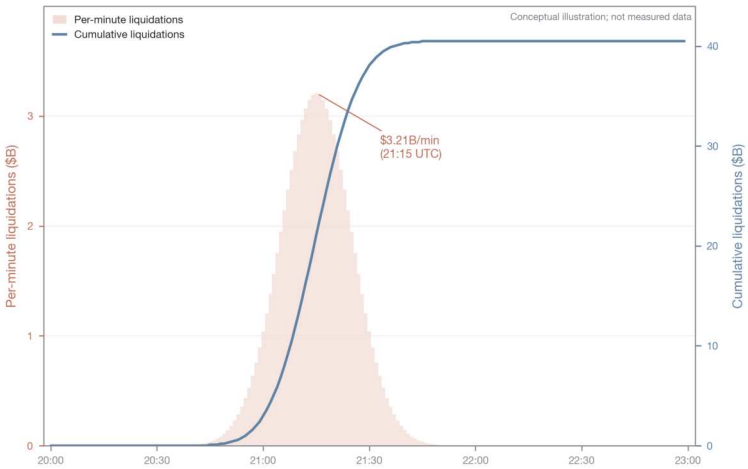


Figure 11-39. The dynamic relationship between the liquidation rate and cumulative liquidation volume (an illustration of the liquidation-cascade model, not tick-by-tick empirical data from the “10.10 event,” and distinct from Figure 11-36)

Figure 11-39 overlays the liquidation rate with cumulative liquidation volume, revealing two key dynamic features of the cascade process: the peak of the liquidation rate occurs at the inflection point of the cumulative-liquidation-volume curve, corresponding to the system’s jump from the propagation phase to the amplification phase; and the sharp fall-back in the liquidation rate is not due to external intervention but to the fact that the liquidatable high-leverage positions have largely been cleared and the system’s own “fuel” is exhausted. This asymmetric pulse shape (slow accumulation, instantaneous eruption, rapid decay) constitutes the typical temporal signature of a crypto-market liquidation cascade.

11.9.4 Empirical validation of the four-phase model

Using the liquidation reflexivity equation and the four-phase model proposed in this chapter, we can systematically reconstruct the entire process of the “10.10 event.” In the trigger phase, macro news caused an initial decline of 6.8%. Entering the propagation phase, the price hit the liquidation-dense zone, the reflexivity engine activated, and the liquidation-impact absorption rate fell below 1. In the amplification phase, the time-compression effect appeared, 70% of the destructiveness was released within a mere 40 minutes [2], and multiple resonances drove depth to zero. Finally, in the termination phase, as open interest plunged 25% and liquidatable leveraged positions decreased sharply, the cascade finally stopped.

The price trajectory after the crisis also provides strong empirical support for the mechanism-induced volatility discussed earlier. After the cascade slammed Bitcoin’s price down to a low of $104,800, as the mechanical selling pressure exhausted itself, the price rebounded quickly over the following day and stabilized near $113,000. This price difference of nearly $8,000 constitutes a direct measure of the overshoot the liquida-

tion mechanism injected into the market. As analyzed in Section 11.8.3, this overshoot comprises two components at once: mechanical-liquidation overshoot (the nonlinear price impact of the liquidation engine's forced sell orders in the liquidity vacuum, corresponding to about 93.5% of the sell orders at the peak moment being forced closures) and behavioral-panic overshoot (the voluntary panic selling of traders who had not been liquidated after they observed the cascade, corresponding to about 6.5% of the voluntary selling pressure, whose marginal price impact was equally significant under the already collapsed depth). Attributing the entire $8,000 price difference wholesale to "pure mechanism-induced noise" is reasonable in order of magnitude, but strictly speaking it requires further distinguishing the respective contributions of these two components.

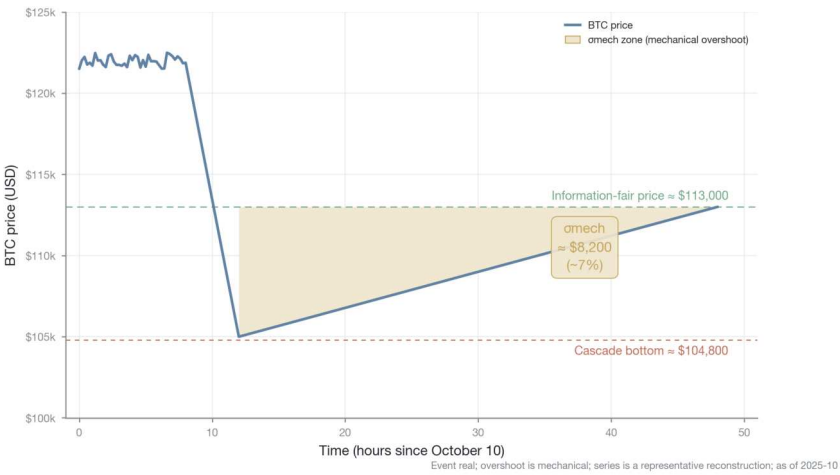


Figure 11-40. Empirical validation of mechanism-induced volatility—price overshoot and rebound (Data source: Amberdata [2], CoinGecko [3], as of October 2025; the endpoints of about $104,800/$113,000 are empirical, while the mechanism-induced overshoot $\sigma \approx$ $8,200, about 7%, is the author's model decomposition and a rough upper bound)

The rapid rebound path of the price after the cascade's termination in Figure 11-40 provides an approximate measure of mechanism-induced volatility: the price difference from the low of $104,800 to the stable range of $113,000 accounts for nearly half of the total decline during the cascade (about 46% based on the intraday high of about $122,500, or about 49% based on the trigger-moment high of about $121,500, the ratio depending on which high is used). Equating this ratio directly with the "share of mechanism-induced volatility," however, requires caution about several methodological limitations. First, using the post-crisis recovery price as a proxy for "fundamental fair value" is itself biased, because the recovery-period price may be affected by short-covering buying pressure, the narrowing of spreads after market makers re-enter, and post-crisis new information (such as signals of a policy response), deviating from the true valuation of the moment just before the crisis. Second, this ratio should be regarded as a rough upper bound on mechanism-induced volatility rather than a precise measure, because the overshoot contains both a mechanical-liquidation-overshoot component and a behavioral-panic-overshoot component (as in Section 11.8.3, with about 93.5% forced

liquidation and 6.5% voluntary panic selling at the peak, the latter still producing a non-negligible price impact under the already collapsed depth). A more robust alternative measure would compare the difference in decline between perpetual futures and spot over the same window (perpetual futures should systematically overshoot because of the liquidation mechanism), or compare the difference in decline between a crypto asset and a traditional asset with similar macro sensitivity but no cascade-liquidation mechanism. Rough as it is, this ratio still provides a valuable order-of-magnitude benchmark for assessing the market-efficiency loss of different liquidation modes.

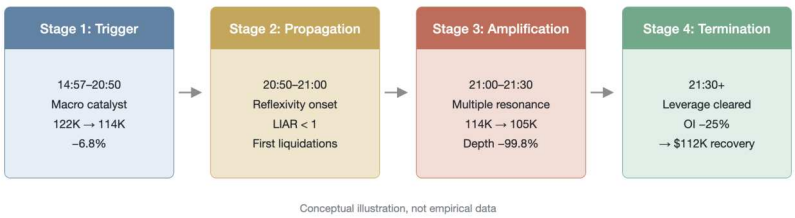


Figure 11-41. Mapping of the four-phase model onto the "10.10 event" (event data as of October 2025; the phase boundaries in the figure are drawn by the model)

As shown in Figure 11-41, superimposing this chapter's four-phase model onto the actual data of the "10.10 event" validates the consistency between theory and empirical observation: the six-hour slow decline of the trigger phase (14:57–20:50 UTC), LIAR falling below 1 in the propagation phase (20:50–21:05), depth going to zero and spreads widening more than a thousandfold in the amplification phase (21:05–21:30; see Table 11-7), and the liquidation rate falling back after OI plunged 25% in the termination phase all closely match the model's phase divisions and transition conditions. This shows that the liquidation reflexivity equation has some explanatory power, while the cascade risk index, as an exploratory proxy for fragility, still requires large-sample testing to establish its warning value.

11.10 Chapter summary

Through a systematic analysis of the liquidation mechanism of perpetual futures, this chapter has revealed the core structural paradox in the microstructure of crypto-asset markets: the liquidation mechanism was designed to eliminate bad debt and maintain system solvency, but under extreme market conditions it turns into the main amplification mechanism of systemic risk. This transformation stems not from a single code flaw but is rooted in the endogenous interaction of leverage, liquidity, and automated execution. This chapter's analytical framework centers on the endogenous microstructure while also revealing the modulating role of macro liquidity conditions on baseline fragility, the amplifying effect of behavioral-finance mechanisms in clustering and panic, and the external constraint that the legal and regulatory dimension imposes on mechanism design.

The five theoretical tools this chapter constructs deconstruct this dynamic process along

different dimensions. The liquidation reflexivity equation formalizes the two-way causal relationship between liquidation and price, showing that when liquidation selling pressure exceeds the order book's absorption capacity, the price decline is no longer driven by fundamentals but dominated by the liquidation mechanism itself. The impossible trinity of liquidation design further indicates that in a decentralized or highly fragmented market, capital efficiency, liquidation smoothness, and system safety cannot be achieved simultaneously, and existing exchanges generally pursue capital efficiency at the expense of liquidation smoothness, creating the structural conditions for cascade collapse. The time-compression effect explains why the deleveraging process in crypto markets is completed on a time scale of minutes: the millisecond execution speed of smart contracts and matching engines deprives traditional risk-management measures of their response window. The theory of the information paradox and mechanism-induced volatility quantifies the degree to which liquidation events distort price-discovery efficiency: unable to distinguish informational selling from mechanical closing in the order book, market makers are forced to adopt overly defensive strategies and, at the macro level, manufacture a price overshoot beyond the change in fundamentals. The compound positive-feedback model of the funding rate and liquidation reveals how extreme funding rates systematically accumulate fragility before a liquidation erupts by continuously consuming longs' maintenance margin.

The "10.10 event" provides a complete empirical validation of this analytical framework. The macro shock was amplified step by step through the weaknesses of the microstructure, the unified margin system turned in the crisis from a capital-efficiency tool into a conduit for cross-asset contagion, and the forced intervention of auto-deleveraging further weakened the market's remaining capacity to supply liquidity.

These findings have practical significance at different levels. Traders' risk management needs to go beyond monitoring individual positions and incorporate systemic liquidity conditions and leverage crowding into its assessment framework. In a highly reflexive market, an individual's leverage not only reflects their own risk appetite but also constitutes a micro-contributing factor to system fragility. Even if an individual adopts a conservative hedging strategy, they may still suffer collateral liquidation under cross-margin and auto-deleveraging mechanisms. Market-infrastructure construction, for its part, needs to place systemic-risk management at the core of product design; specifically, multi-node liquidity-weighted oracles, progressive liquidation curves, and elastic risk-reserve mechanisms are key paths to reducing cascade risk. At the regulatory level, static capital-adequacy requirements or leverage caps are no longer sufficient to cope with dynamic risk, and the regulatory focus needs to shift toward systemic-risk monitoring based on real-time data; an early-warning system that captures order-book depth changes, cross-market funding-rate anomalies, and liquidation-concentration indicators in real time is a precondition for intervening before a crisis takes shape.

The intrinsic paradox of the liquidation mechanism essentially reflects the fundamental tension between efficiency and stability in crypto-asset markets. Having understood how liquidation turns from a risk-management tool into a risk-amplification mechanism, we arrive at the next core question: when the liquidation engine has exhausted market liquidity and auto-deleveraging can no longer cover the margin shortfall, where is the system's final line of defense? This leads to the subject of the next chapter: the insur-

ance fund. The capital structure of the insurance fund, its replenishment mechanism, and its capacity to bear stress under extreme conditions will determine the systemic resilience of the crypto-asset derivatives market.

References

[1] Ali, Z. (2025). Anatomy of the Oct 10–11, 2025 crypto liquidation cascade: Macroeconomic triggers, market microstructure, and systemic risk lessons. *SSRN Electronic Journal*.

[2] Marshall, M. (2025). How $3.21B vanished in 60 seconds: October 2025 crypto crash explained through 7 charts. *Amberdata Research Blog*. https://blog.amberdata.io/how-3.21b-vanished-in-60-seconds-october-2025-crypto-crash-explained-through-7-charts (archived at web.archive.org)

[3] CoinGecko. (2026). *What is October 10th? Crypto's 10/10 mass market liquidation event*. https://www.coingecko.com/learn/october-10-crypto-crash-explained (archived at web.archive.org)

[4] BitMEX Research. (2020). *The 12 March 2020 Bitcoin crash*. https://blog.bitmex.com/the-12-march-2020-bitcoin-crash/ (archived at web.archive.org)

[5] Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review*, *105*(2), 564–608. https://doi.org/10.1257/aer.20130456

[6] Soros, G. (1987). *The alchemy of finance*. Simon and Schuster.

[7] Soros, G. (2009). *General theory of reflexivity* [Lecture series]. Central European University.

[8] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[9] CoinGlass. (2025). *Cryptocurrency liquidation history data*. https://www.coinglass.com/LiquidationData (archived at web.archive.org)

[10] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

[11] Danielsson, J., Shin, H. S., & Zigrand, J.-P. (2012). Procyclical leverage and endogenous risk. *Working Paper*, London School of Economics.

[12] Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026. https://doi.org/10.1086/261849

[13] Binance. (n.d.). *Introduction to Binance Futures funding rates*. https://www.binance.com/en/support/faq/detail/360033525031 (archived at web.archive.org)

[14] Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*(3), 183–206. https://doi.org/10.1002/(sici)1099-0771(199909)12:3<183::aid-bdm318>3.0.co;2-f

[15] Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35–55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

[16] Coinalyze. (n.d.). *Aggregated cryptocurrency futures market data*. https://coinalyze.net/ (archived at web.archive.org)

[17] CoinShares. (2025). *Billions in crypto liquidations: Inside October's $19B crash*. https://coinshares.com/insights/knowledge/billions-in-liquidations-what-happened/ (archived at web.archive.org)

[18] Chitra, T. (2025). Autodeleveraging: Impossibilities and optimization. *arXiv preprint arXiv:2512.01112v2*. https://doi.org/10.48550/arXiv.2512.01112

[19] Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, *3*(2), 5–39. https://doi.org/10.21314/jor.2001.041

[20] Duffie, D., & Zhu, H. (2011). Does a central clearing counterparty reduce counterparty risk? *Review of Asset Pricing Studies*, *1*(1), 74–95.

[21] Pirrong, C. (2011). The economics of central clearing: Theory and practice. *ISDA Discussion Paper Series*, No. 1.

[22] Binance. (2025). *How does the liquidation process work on Binance*. https://www.binance.com/en/support/faq (archived at web.archive.org)

[23] Hyperliquid. (2025). *Hyperliquid activates cross-margin auto-deleveraging*. https://hyperliquid.xyz (archived at web.archive.org)

[24] WisdomTree Prime. (2025). *The great whale slap: How a whale offloaded $4M in losses to Hyperliquid's HLP vault*.

[25] Shirreff, D. (1998). Lessons from the collapse of hedge fund, Long-Term Capital Management.

[26] Investopedia. (2024). *Lehman Brothers collapse explained: Causes and legacy*. https://www.investopedia.com

[27] Kubali, V. (2025). Crypto crash Oct 2025: Leverage meets liquidity. *FTI Consulting*. https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity (archived at web.archive.org)

[28] Bank of England. (2023). *An anatomy of the 2022 gilt market crisis*.

[29] Goldstein, M. A., & Kavajecz, K. A. (2004). Trading strategies during circuit breakers and extreme market movements. *Journal of Financial Markets*, *7*(3), 301–333. https://doi.org/10.1016/j.finmar.2003.11.003

[30] Subrahmanyam, A. (1994). Circuit breakers and market volatility: A theoretical perspective. *Journal of Finance*, *49*(1), 237–254. https://doi.org/10.1111/j.1540-6261.1994.tb04427.x

[31] Bouchaud, J.-P., Farmer, J. D., & Lillo, F. (2009). How markets slowly digest changes in supply and demand. In T. Hens & K. R. Schenk-Hoppé (Eds.), *Handbook of financial markets: Dynamics and evolution* (pp. 57–160). North-Holland. https://doi.org/10.1016/b978-012374258-2.50006-3

[32] Shiller, R. J. (1981). Do stock prices move too much to be justified by subsequent

changes in dividends? *American Economic Review*, *71*(3), 421–436.

The Political Economy of Insurance Funds

On October 10, 2025, the cryptocurrency market experienced an unprecedented mass liquidation event. Within just 24 hours, visible liquidation notional exceeded $19 billion (roughly $19.1–19.4 billion on CoinGlass's basis; because some platforms' liquidation feeds are rate-limited to a single record per second, this visible liquidation notional is likely systematically underestimated, and CoinGlass estimates that the true scale of deleveraging may have reached $30–40 billion), affecting roughly 1.62 million traders and erasing about $350 billion in total cryptocurrency market capitalization. This catastrophe, later dubbed the "10/10" event, laid bare for every market participant a problem long buried in the mechanics of perpetual futures: when a chain of forced liquidations finally breaks through a trader's margin floor and drives account equity below zero (a *shortfall*), who ultimately bears that loss, and how?

In the previous chapter, we analyzed in detail how risk evolves from the failure of an individual trader—through the chain reaction of a liquidation cascade—into systemic risk that sweeps across the entire market. If the previous chapter answered the question of how risk spreads, this chapter takes up that logic to examine a more fundamental proposition: how the system absorbs and allocates that risk. The deficit created by a shortfall does not vanish on its own; it must be filled. If such a deficit is allowed to widen, profitable traders can no longer be certain that their gains will be honored in full, and the market's foundation—the credible management of counterparty risk—loses its essential function. In a market where counterparty credit cannot be guaranteed, market makers will refuse to provide liquidity, institutional investors will exit, and the market as a whole will ultimately shrink amid a collapse of trust.

How losses are borne is, at bottom, a question of risk socialization. Because a shortfall loss cannot be covered by the defaulter, it must be transferred to other participants in the system. In crypto derivatives markets, risk socialization proceeds along two distinct paths: ex ante pooling (establishing an insurance fund) and ex post allocation (auto-deleveraging or socialized loss). The two are not opposing alternatives but tightly linked levels of a single risk-socialization mechanism, together forming a tiered risk-resolution system (its penalty sources, buffer-pool mechanics, and allocation mechanisms are elaborated in Section 12.1.2).

Within this system, an insurance fund is far more than a simple pool of capital; it is an institutional arrangement charged with the tensions of political economy. Where

does the money come from? Is the fund large enough? Who has the authority to decide when it is used? Are the rules for its use transparent? And does the very existence of an insurance fund create new moral hazard? This chapter argues that the insurance fund is the node where technical design and political governance intersect most deeply in the perpetual-futures system. Its technical dimension—size, adequacy ratio, and trigger rules—can be optimized with actuarial models; but its governance dimension—who controls the fund, who decides on its use, whether information is disclosed, and whether users are told the true risks—involves deeper questions of power, trust, and the distribution of interests. These questions cannot be solved by code alone; they can only be constrained by sound institutions. At the same time, the insurance fund and the auto-deleveraging mechanism together constitute a dual institutional arrangement for risk socialization, and this combination of ex ante pooling and ex post allocation ultimately determines the settlement credibility and rule predictability of the perpetual-futures market. The chapter pursues three questions in turn: What is the dual mechanism of risk socialization, and how does it guarantee settlement credibility? Why must this system inevitably involve political economy, and how are its design parameters determined amid the interplay of competing interests? And how can we build a systematic analytical framework to evaluate the risk-mutualization mechanisms of different platforms?

12.1 The inevitability of shortfalls

Before examining how risk is socialized, we must first grasp a core fact: forced liquidation is by no means risk-free liquidation. In theory, a liquidation engine always seems able to close a position precisely before the price reaches the bankruptcy line, containing the loss perfectly within the trader's margin. In reality, however, a vast gulf—created by multiple sources of friction—separates the idealized liquidation model from the chaos of actual markets. It is precisely these frictions that make shortfalls structurally inevitable, and the existence of shortfalls is the fundamental reason the entire risk-socialization mechanism and all of its institutional arrangements come into being. Understanding these frictions is the logical starting point for understanding why insurance funds must exist.

12.1.1 Four frictions in the liquidation process

When markets enter a state of extreme volatility, the liquidation mechanism tends to encounter four core frictions. These frictions do not stem from design flaws in the liquidation engine itself; they are rooted in the intrinsic limitations of market microstructure and infrastructure. They are the direct drivers of shortfalls and a key entry point for understanding why insurance funds are necessary.

The most direct friction arises from insufficient order-book depth. When a large-scale liquidation occurs, enormous market-order closeouts instantly exhaust the liquidity resting on the order book. Market makers, as the suppliers of market liquidity, build their core business model on spread income rather than on bearing directional risk. When the market moves violently in one direction, market makers—seeking to avoid excessive

directional exposure of their own—often widen their bid-ask spreads sharply or withdraw their quotes altogether. This sudden evaporation of liquidity forces liquidation orders to fill at markedly unfavorable prices, and the final execution price is often far worse than the trader's bankruptcy price. This phenomenon was especially pronounced during the 2025 "10/10" event. According to data from the research firm Amberdata, during a mere 40-minute phase of violent cascading (20:50 to 21:30 UTC), positions totaling $6.93 billion were liquidated, and the liquidation rate surged to $10.39 billion per hour—about 866 times the $12 million per hour of the normal period in the eight hours before the event [1]. Under such concentrated selling pressure, the top-of-book depth for Bitcoin shrank by more than 90%, and the bid-ask spread jumped from a normal single-digit number of basis points to double-digit percentages [2]. In such an extreme liquidity vacuum, the liquidation engine simply could not complete closeouts without triggering shortfalls. Figure 12-1 places the 10/10 event in a longitudinal comparison with major liquidation events in the history of crypto derivatives; it shows that the liquidation scale of October 10, 2025 (about $19.37 billion on CoinGlass's basis, exceeding $19 billion) was nearly double the previous record (about $10 billion on April 18, 2021) and almost 20 times that of the 2020 "Black Thursday," reflecting a trend in which liquidation scale accelerates as the market becomes more leveraged.

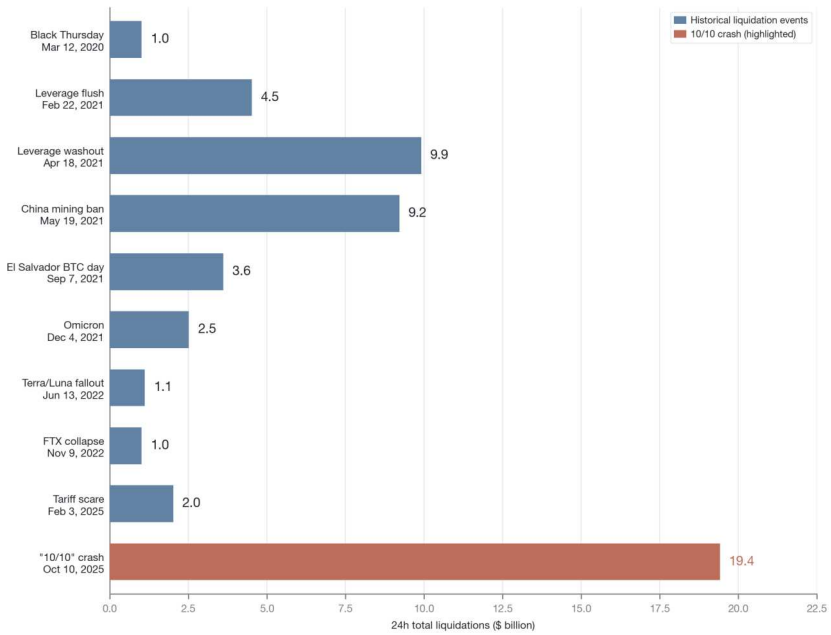


Figure 12-1. Historical comparison of major liquidation events in crypto derivatives (CoinGlass basis, subject to historical underestimation; the April 18, 2021 and May 19, 2021 events were of similar scale, both in the range of roughly $9 billion to $10 billion, with April 18 slightly higher when reconciled at present value)

Comparison across periods based on absolute liquidation volume alone can mislead, because the overall size of the cryptocurrency market has differed enormously across periods. A more analytically useful, standardized metric is the ratio of liquidation volume to the total cryptocurrency market capitalization of the same period. Take

the March 2020 "Black Thursday": about $10 billion in liquidations corresponded to a total crypto market capitalization of roughly $200 billion at the time, a normalized ratio of about 0.5%. By contrast, the roughly $19.37 billion in visible liquidations on October 10, 2025 (CoinGlass basis; note that this visible notional may be systematically underestimated because of platform rate-limiting) corresponded to a total market capitalization of about $3.5 trillion, a normalized ratio of about 0.55%. The closeness of the two events on this standardized metric offers a preliminary hint that the relative severity of extreme liquidation events across different market cycles may follow some regularity worth further verification. This preliminary observation must, however, be treated with caution: two data points are not enough to establish a statistically robust regularity; the normalized ratios of other historical extreme events (such as May 19, 2021, or the liquidation event triggered by the Bank of Japan's rate hike on August 5, 2024) may deviate substantially from this range; and the heterogeneity of macro shocks (global systemic panic versus a regional geopolitical shock may differ fundamentally in transmission path and liquidation severity) means it would be unwise to attribute the consistency of the two ratios simply to an endogenous structural property of the crypto market. Moreover, using total open interest (OI) rather than total market capitalization as the normalization denominator could yield different conclusions. If this preliminary observation holds up under a larger sample and more diverse standardization methods, it would carry important implications for calibrating the actuarial models of insurance funds.

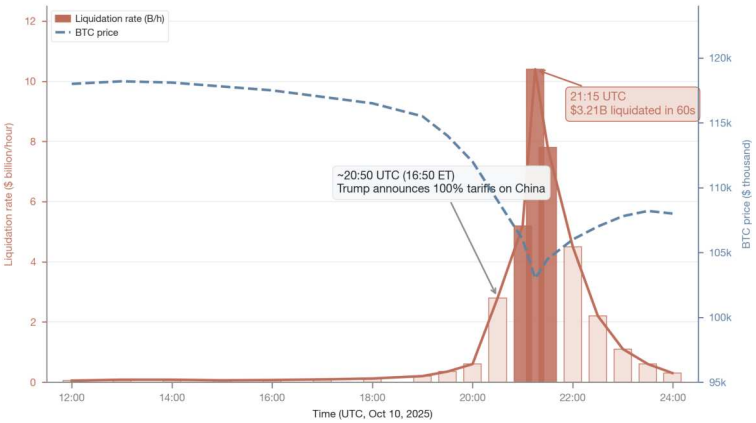


Figure 12-2. Liquidation rate and BTC price timeline for the 10/10 event (the minute-by-minute liquidation-rate and price series is a reconstructed illustration based on characteristic parameters from Amberdata and CoinGecko, not raw tick-by-tick data; endpoints and peaks have been verified)

As shown in Figure 12-2, within the 60 seconds around 21:15 UTC, positions worth $3.21 billion evaporated instantly—a liquidation rate that far exceeded the market's liquidity-bearing capacity. Insufficient order-book depth is a problem not only during extreme events; in day-to-day trading, the distribution of liquidity is itself highly time-varying and non-uniform, and extreme conditions merely amplify this ever-present fragility in its most violent form. The retreat of market makers, though rational risk avoidance at the individual level, constitutes at the system level a *procyclicality of liquidity*: the

moment the market most needs liquidity is precisely the moment when the supply of liquidity is scarcest.

A second class of friction, running parallel to insufficient liquidity, is the price gap. In certain extreme conditions, an asset's price does not move continuously and smoothly but plunges in a cliff-like fashion. The price may instantly leap past a trader's liquidation price or even bankruptcy price, so that the position has no chance to fill at the preset protective level. This phenomenon exists in traditional financial markets as well (the price limits and circuit breakers of stock markets were designed precisely to address such price discontinuities), but in a cryptocurrency market that trades 24/7 without circuit breakers, the losses caused by price gaps are markedly larger. The "Black Thursday" of March 12, 2020, is the most typical case. That day, amid the global panic selling triggered by the COVID-19 pandemic, the price of Bitcoin plunged from about $7,900 to around $3,800 within roughly 24 hours (spanning the two trading days of March 12 and 13), a drop of more than 50% [3]. In such an extreme price gap, large numbers of highly leveraged long positions fell directly into deep shortfall. According to a research report by Ledger Prime, BitMEX alone saw liquidations exceeding $1.4 billion within a roughly 24-hour rolling window; during its downtime it deployed 1,627 deleveraging positions, and its insurance fund fell from about 35,500 BTC to about 33,880 BTC (a decrease of about 1,620 BTC) [4]. The essence of a price gap is that it renders wholly invalid the assumption of price continuity on which the liquidation engine relies. When prices no longer move continuously, any liquidation logic based on step-by-step triggering leads directly to a shortfall because it cannot find an executable intermediate price.

The divergence between mark price and execution price constitutes a third class of friction. To prevent market manipulation and abnormal price swings on a single exchange from triggering unwarranted liquidations, most exchanges use a weighted average of external spot indices as the *mark price* that triggers liquidation. When the exchange's internal derivatives market collapses, however, the internal contract execution price diverges enormously from the external index price. This divergence creates a dangerous *scissors gap*: relying on the relatively stable external mark price, the liquidation engine judges that a position has not yet reached the liquidation line, but when that position must finally be closed in the collapsing internal market, the actual execution price may already be far below the bankruptcy price. Conversely, in other cases abnormal swings in the internal price may trigger liquidations that should never have occurred. During a contract flash crash on OKEx (now OKX) on March 30, 2018, the external spot index price held at about $7,000, but the exchange's internal quarterly contract execution price instantly plunged to about $4,755 (some sources round this to about $4,800). This divergence of more than 25% relative to the index (about 30% to 32% at the prices above) caused large numbers of traders who should not have been liquidated to suffer irrational forced liquidation, producing severe shortfall losses (OKEx subsequently rolled all contracts back to 04:07 that day and introduced price-limit rules). The mark-price mechanism is itself an important protective measure, but it cannot eliminate the risk of the internal and external markets decoupling under extreme stress—and that decoupling is yet another important source of shortfalls.

Infrastructure congestion aggravates shortfall risk along another dimension. Even if the

liquidation mechanism is flawless in its logical design, bottlenecks in the underlying infrastructure on which it runs can likewise lead to serious consequences. This problem is especially acute in decentralized finance, because the execution speed and throughput of on-chain transactions are hard-constrained by the underlying blockchain's performance. During the 2020 "Black Thursday," the Ethereum network became severely congested under a flood of transaction requests, and gas fees surged from about 20 Gwei in normal times to more than 500 Gwei, a 25-fold increase. This left MakerDAO's liquidation bots unable to submit liquidation transactions in time, because the gas fees they bid were insufficient for miners to include them within a reasonable period. As a result, over a window of about 40 minutes, some speculators seized the opportunity to "win" auctions of Ethereum collateral worth millions of dollars at the extremely low cost of 0 DAI; the total value of collateral swept away in these zero-bid auctions reached about $8.32 million (in a normal liquidation auction, collateral should be sold competitively to recover an equivalent amount of DAI to repay the debt, but because the liquidation bots could not participate in the bidding in time, speculators won with bids of 0 DAI for ETH collateral that should have been worth millions). After deducting the collateral taken for nothing, the MakerDAO protocol was ultimately left with about 5.67 million DAI (roughly $4.5 million) in uncollateralized bad debt, a gap eventually covered by an auction that minted new MKR tokens [5]. The success of the 0 DAI auctions was also partly attributable to a design flaw in MakerDAO's auction contract—namely, a starting bid price of 0 with no reserve-price protection—which allowed a speculator to win at an arbitrarily low price when no competitors participated. This event demonstrated that in DeFi, the effectiveness of a liquidation mechanism depends not only on the logical correctness of the smart contract but also on the availability and reliability of the underlying blockchain network under extreme stress. Since 2020, the infrastructure here has changed substantially. Most current decentralized perpetual-futures platforms (such as Hyperliquid and dYdX v4) run on dedicated app-chains or high-performance Layer 2s, and the problem of gas congestion has largely been resolved. The new architecture, however, introduces new forms of infrastructure risk—including the single-point-of-failure risk of the sequencer, app-chain consensus latency, and cross-chain bridge dependencies—and these new bottlenecks can likewise cause liquidation delays and shortfalls in extreme conditions.

These four frictions point to a core fact: no matter how well the mechanism itself is designed, infrastructure bottlenecks, the endogenous limits of market liquidity, and the discontinuity of price discovery can all lead to serious shortfalls. A shortfall is by no means a flaw of system design; it is a structural inevitability under real market frictions. This structural inevitability gives rise directly to the fundamental need for a risk-socialization mechanism.

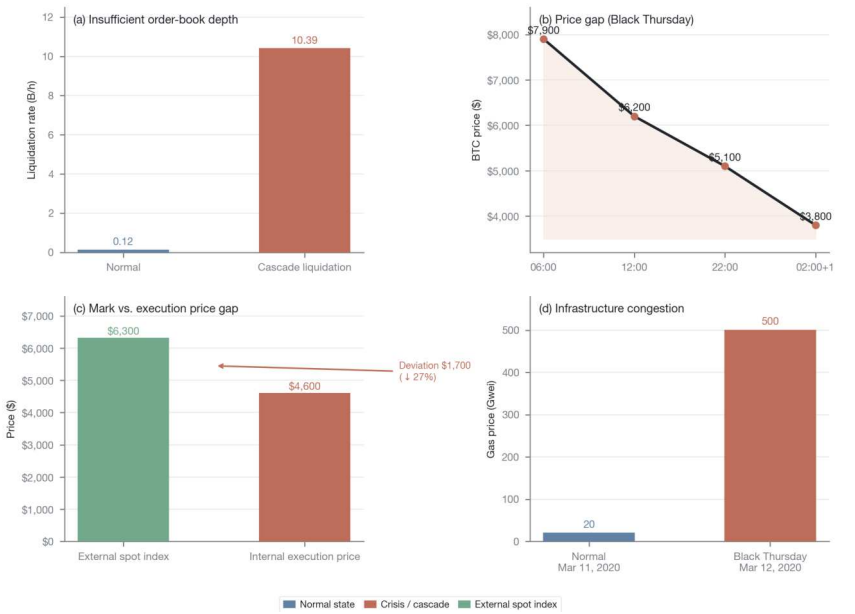


Figure 12-3. The four frictions in the liquidation process and their representative cases (panel ① draws on the 2025 "10/10" event; panels ② through ④ draw on the "Black Thursday" of March 12, 2020; the panels are dimensionally heterogeneous, and panel ① approximates order-book depth depletion with the liquidation rate)

12.1.2 Ex ante pooling and ex post allocation

The inevitability of shortfalls means their economic consequences must be handled systematically. A shortfall is, in essence, a default by a debtor (the liquidated trader whose equity is negative) on a creditor (the profitable trader). In traditional financial markets, a broker can pursue a defaulting client for the amount owed through legal channels, because the client's true identity and asset situation are known. In cryptocurrency markets, however, because of the anonymity (or at least pseudonymity) of accounts and the absence of real-world legal recourse, a platform cannot pursue defaulters for the shortfall as a traditional broker would. At the behavioral level, this institutional feature constitutes an implicit moral-hazard incentive: a highly leveraged trader faces an asymmetric payoff structure of "limited liability plus unlimited upside"—profits accrue to the individual, while losses beyond the margin are borne by the insurance fund or by other traders. As Jensen and Meckling (1976) noted in their classic agency-theoretic analysis, limited liability systematically incentivizes an upward shift in risk appetite [6]. As Know Your Customer (KYC) requirements spread and on-chain identity-tracking technologies (such as Chainalysis and Arkham) advance, this "no recourse" constraint is gradually eroding—yet it remains the general reality of crypto derivatives markets at the current stage. This means the loss produced by a shortfall cannot be pushed outside the system; it must be absorbed within it. In other words, the loss must be socialized—borne collectively by some group of participants in the system. To understand this inevitability, one must recognize the (aggregate) zero-sum nature of the perpetual-futures market once funding rates and fees are removed: in the aggregate, a shortfall trader's excess loss corresponds exactly to the receivable, unrealized profit

of the holders on the other side of the contract (typically the profitable side). The shortfall trader's loss exceeds what their margin can cover, and this excess loss corresponds in accounting terms to the profitable side's receivable profit. If the system does not apportion this "bad debt" through some mechanism, the profitable side's profit becomes a paper figure that cannot be honored. Risk socialization is therefore not a subjective choice of the system's designers but a logical consequence of shortfalls under a zero-sum structure. Faced with this challenge, crypto derivatives markets have evolved two distinct paths of risk socialization.

The first path is ex ante pooling, whose core vehicle is the insurance fund. Before a shortfall event occurs, the insurance fund gradually accumulates a common pool of capital by continuously collecting liquidation penalties from the risk-takers who trigger liquidation. Besides liquidation penalties, the fund's revenue sources also include liquidation surplus (the spread that arises when the liquidation execution price is better than the bankruptcy price, which under normal market conditions is typically the main driver of the fund's growth) and, at some platforms, dedicated subsidies drawn from fee income. When a shortfall actually occurs, this pool serves as the first line of defense, collectively covering the deficit. The core advantage of this approach is that it has no direct effect on profitable traders: it does not erode their profits or disrupt the hedging strategies they have carefully constructed. More importantly, because the rules of the insurance fund are set in advance, it provides the market with a high degree of rule predictability. Before entering, traders know clearly that under normal circumstances their profits will not be eroded by others' shortfalls.

The second path is ex post allocation, which appears mainly in two mechanisms: socialized loss and auto-deleveraging. When the insurance fund is exhausted, the system has no choice but to impose the loss on profitable traders. Socialized loss uses a global-apportionment approach, deducting the total loss proportionally from all profitable accounts. Although this makes each person's loss relatively small, the amount is entirely unpredictable, and traders discover only at settlement that part of their profit has been deducted. Auto-deleveraging, by contrast, takes a targeted-selection approach: the system ranks positions by a combined score of return and leverage and forcibly closes the top-ranked profitable positions to offset the defaulter's loss. A trader subject to auto-deleveraging (ADL) is typically force-closed at the shortfall trader's bankruptcy price, which means the selected profitable trader not only loses the position but may be executed at a price far worse than the current market price, so that the actual loss can far exceed a mere give-back of book profit. This mechanism can be highly destructive for the selected trader: a cross-market hedged portfolio carefully constructed across several exchanges can turn instantly from "market-neutral" into naked directional exposure because one of its legs is force-closed by ADL. The execution challenges faced after being ADL'd go well beyond this: the hedging leg on another exchange must be urgently closed in an equally illiquid market; cross-platform capital rebalancing depends on on-chain transfers (which may be delayed by tens of minutes during network congestion); and the compounding of slippage in extreme conditions makes the total loss far exceed the book impact of the ADL itself. The real destructiveness of ADL to cross-platform hedging strategies is far greater than the loss on a single position, and its unpredictability severely undermines the foundation of market trust.

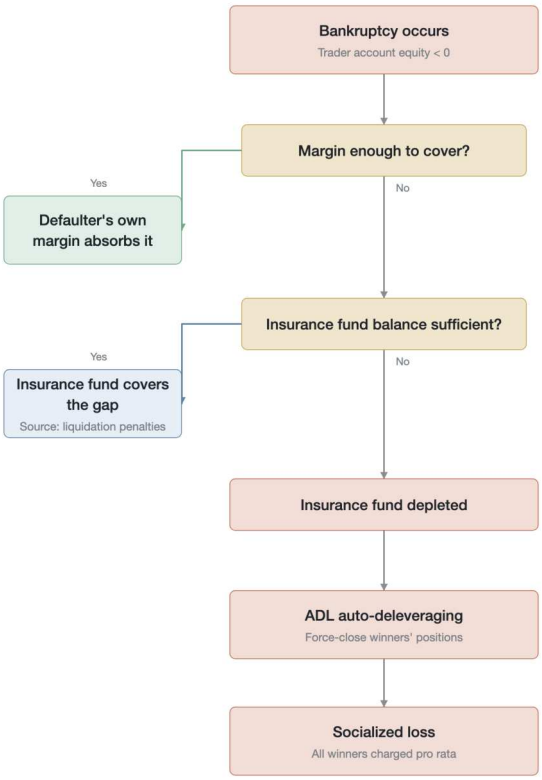


Figure 12-4. The liquidation and risk-socialization waterfall for perpetual futures

The comparison makes clear that both socialized loss and auto-deleveraging resolve the immediate crisis by undermining rule predictability. Socialized loss makes every profitable trader's profit uncertain; auto-deleveraging makes some profitable traders' continued holding uncertain. This uncertainty is a fundamental flaw that any mature financial market strives to avoid, because it directly erodes participants' confidence in the settlement system. For precisely this reason, the ex ante insurance fund is decisive: it is not merely a capital buffer but the ultimate buffer that reduces the probability of ex post allocation being triggered. The more adequate the insurance fund, the lower the probability that ex post allocation is triggered, and the higher the market's settlement credibility and rule predictability.

Ex ante pooling and ex post allocation are not the only conceivable architectures for risk socialization. In traditional finance and the academic literature, there exist at least three alternative or complementary mechanisms. The first is the circuit breaker: when a price falls by more than a preset threshold within a short period, the system halts trading to buy a window of time for liquidity to recover and information to be digested. Traditional stock markets widely adopt this mechanism, but in a crypto market that runs 24/7 and treats censorship resistance as a core principle, "who has the authority to press the pause button" is itself a governance problem. The second

is a market-maker backstop obligation, whereby designated market makers assume an obligation of last resort to absorb order flow in extreme conditions, analogous to the New York Stock Exchange's designated market maker system. Crypto market makers, however, are generally not bound by such obligations and are often the first to withdraw liquidity in extreme conditions. The third is a mandatory hedging requirement, whereby one-sided positions above a certain size must hold an offsetting position in an external market, limiting shortfall risk at its source. This mechanism is conceptually the most thorough, but it lacks a technical basis for enforcement in an anonymized crypto market. The practical limitations of these alternatives argue, by contrast, for the institutional rationality of the current two-tier "insurance fund plus ADL" architecture in crypto derivatives markets. Far from perfect, it is nonetheless the most feasible risk-socialization solution under existing technical and governance constraints.

12.1.3 Internalizing negative externalities and supplying a public good

From a deeper economic perspective, an insurance fund is in essence a mechanism for internalizing negative externalities. In the perpetual-futures market, highly leveraged traders take on great risk in pursuit of high returns. When the market moves against them and they fall into shortfall, that loss—if not handled properly—translates directly into losses for other, profitable traders. This phenomenon, in which the risk-taking of some inflicts losses on others, is what economics calls a classic negative externality. Absent any intervening mechanism, risk-takers need not pay the full social cost of their behavior, which leads to systematic excessive risk-taking: the gains from risk-taking accrue to the individual, while the losses are borne by the collective.

The insurance fund solves this problem by institutional design. It levies a fee—the liquidation penalty—on traders who engage in high-risk behavior and are ultimately liquidated. This fee can be viewed as a "risk tax" that forces risk-takers to pay in advance for the externalities they may cause. The system aggregates these taxes into a common buffer pool dedicated to handling the consequences of shortfalls in extreme situations. In this way, the insurance fund forms, within the system, a closed loop of risk pricing and cost absorption, successfully internalizing the negative externalities of individual speculation. In economic principle, this mechanism closely resembles a Pigouvian tax: by taxing behavior that generates negative externalities, it brings private cost closer to social cost and thereby corrects market failure. The core function of a Pigouvian tax is to change incentives ex ante so as to reduce the volume of negative-externality activity, not to compensate victims ex post. If the current level of liquidation penalties is insufficient to effectively deter excessive leverage (that is, if it falls short of the Pigouvian optimal tax rate), then the insurance fund is closer to a "compulsory risk-mutualization fund" than to a Pigouvian tax in the strict sense.

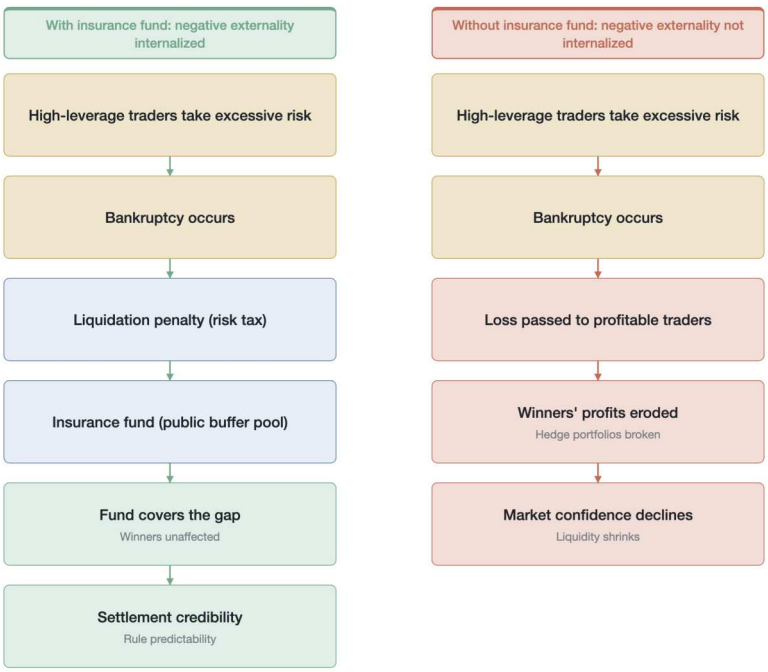


Figure 12-5. How an insurance fund internalizes negative externalities: a mechanism comparison

As shown in Figure 12-5, the presence of an insurance fund corresponds to a virtuous closed loop of cost absorption, whereas the absence of one corresponds to negative externalities spilling outward along the chain of counterparties.

The core output of an insurance fund is not the return earned by investing its capital but two decisive public goods it provides to the entire market. The first public good is settlement credibility. Only when the insurance fund serves as a solid backstop can profitable traders be confident that their book profits will be honored in full and without exception, without having to worry constantly that others' failures will erode their gains. This confidence is especially critical for market makers: the reason they are willing to provide continuous two-sided quotes is that they trust their profits can be reliably settled. If that trust is broken, market makers will withdraw or sharply widen their spreads, and market liquidity will contract dramatically. The second public good is rule predictability. In its Principles for Financial Market Infrastructures, the Bank for International Settlements states explicitly that ensuring the fulfillment of payment obligations to non-defaulting parties is a core function of any financial market infrastructure [7]. In crypto derivatives markets, the insurance fund is precisely the institutional vehicle for this core function. It makes the seemingly self-evident promise that "if you make money, you will surely be able to collect it" possible in a market full of anonymous traders and extreme volatility.

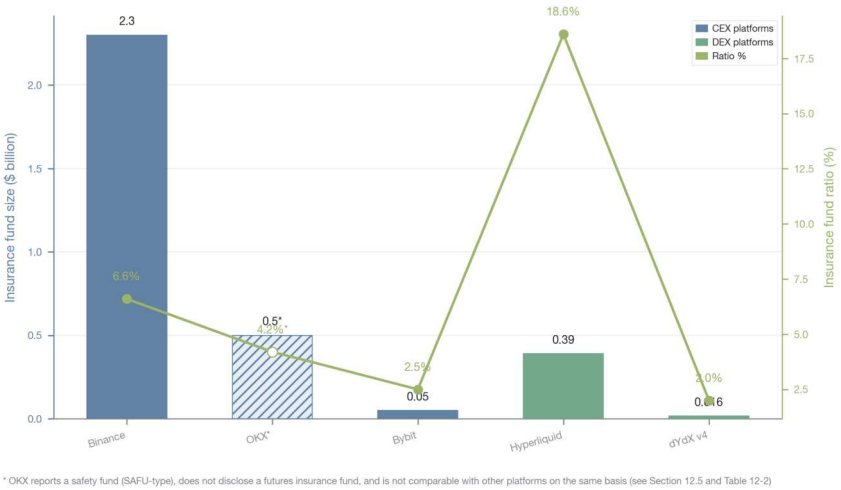


Figure 12-6. Comparison of major exchanges' insurance-fund size and fund-to-open-interest ratio (the fund-to-open-interest ratio is the author's estimate; platforms do not publish this ratio, and the Hyperliquid ratio is an upper-bound estimate based on total HLP TVL; fund sizes are an as-of-2025 snapshot, and present values have since changed and are highly time-sensitive)

Figure 12-6 compares the absolute and relative sizes of major exchanges' insurance funds. It shows that the platform with the largest absolute size (such as Binance, exceeding $2.3 billion) actually has a lower fund-to-open-interest ratio (about 6.6%) than a smaller decentralized platform (such as Hyperliquid, about 18.6%). This difference shows that absolute size alone can be misleading, and that the ratio of the fund to total open interest (here "total open interest" refers to the sum of the notional value of all outstanding contracts across all trading pairs on the platform—that is, the USD-denominated aggregate of open interest) is the effective metric for measuring the buffer thickness per unit of risk exposure. Note that Hyperliquid's roughly $390 million here is the total value locked (TVL) of the HLP vault, which encompasses a shared capital pool for both the market-making strategy and the liquidation-vault strategy. After the 2025 JELLY incident, the liquidation vault was segregated from the HLP main pool, and its standalone size was capped at a "small fraction" of the total HLP balance (the specific proportion has not been publicly disclosed), significantly below the total HLP TVL. Using total HLP TVL directly as the insurance-fund size for cross-platform comparison may therefore overstate Hyperliquid's actual risk-buffer thickness, and the 18.6% ratio should be understood as an upper-bound estimate based on total HLP TVL.

An insurance fund is not, however, a flawless mechanism. In solving the negative-externality problem, it introduces a series of new and complex challenges concerning adequacy of size, transparency of management, and moral hazard. Is the fund large enough to withstand extreme events? Are the rules for its use transparent and verifiable? Does the very existence of the fund instead encourage traders to take on greater risk, because they know a "safety net" underwrites their risk-taking? From a principal-agent perspective, the operation of an insurance fund is nested within two layers of principal-agent relationship. The first layer is between the platform and traders: traders, as principals, delegate the authority for risk management to the insurance fund

operated by the platform, but the platform, as agent, holds an informational advantage regarding the fund's true condition, and this information asymmetry may allow the platform to conceal the fund's true risk exposure or misappropriate its assets (hidden-action moral hazard). The second layer is between the platform and liquidity providers (LPs), who deposit funds into the market-making vault or insurance pool and thereby delegate risk underwriting to the platform, but who lack sufficient observability of the platform's liquidation strategy, risk-parameter adjustments, and discretionary decisions in extreme situations (hidden-information adverse selection). This dual principal-agent structure means that the institutional design of an insurance fund must not only solve the technical question of whether the fund is large enough but also mitigate the moral hazard inherent in each layer of agency through incentive-compatible mechanism design. These challenges test not only a platform's technical design capabilities but, more deeply, its governance wisdom. The sections that follow take up these questions across several dimensions: the institutional architecture of the liquidation waterfall, actuarial adequacy-ratio models, the economic consequences of auto-deleveraging, the political economy at stake, the transmission between confidence and liquidity, dual moral hazard, and the governance spectrum.

12.2 Auto-deleveraging: the final mechanism of risk socialization

When extreme conditions leave the liquidation engine's closeout orders unexecutable for lack of liquidity, and the insurance fund is fully exhausted, the system triggers its last line of defense: auto-deleveraging. As the bottom layer of the liquidation waterfall, ADL is the most contested institutional design in crypto derivatives markets. It no longer attempts to resolve risk through market-based means; instead, through forced intervention, it directly strips some profitable traders of their profits and positions in exchange for the survival of the entire system. ADL is not a simple "risk-bearing switch" but a complex mechanism-design problem, involving the fairness of the ranking algorithm, the differentiated impact on different participants, and an inescapable trilemma among solvency, revenue, and fairness. To understand ADL is not merely to understand a risk-control algorithm; it is an entry point for analyzing the trade-offs among efficiency, fairness, and solvency in crypto markets under extreme conditions.

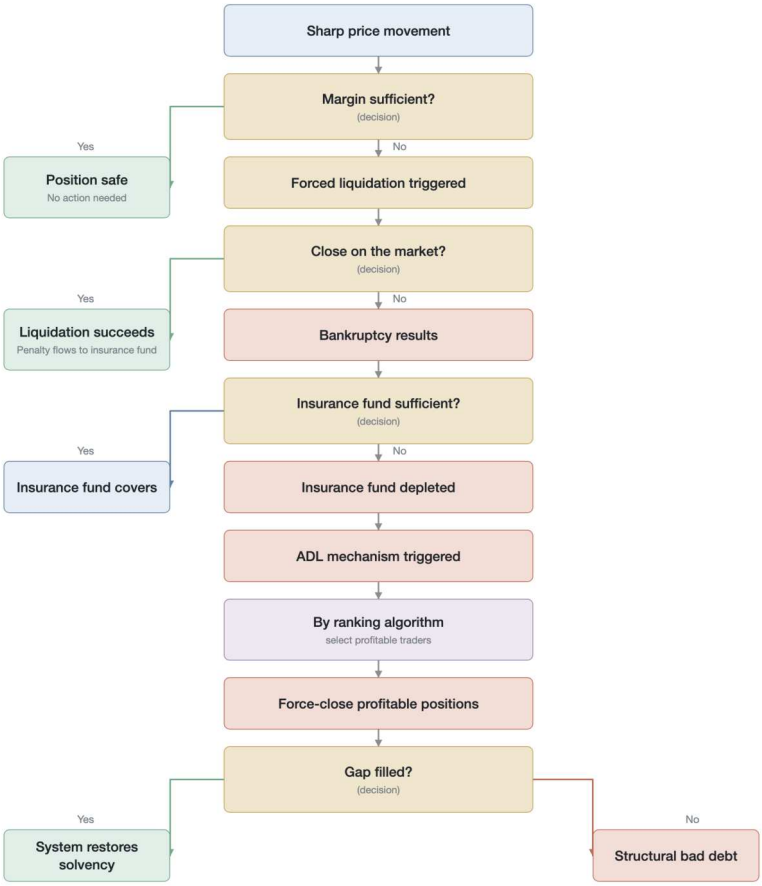


Figure 12-7. The decision process from price movement to an ADL trigger

12.2.1 Ranking algorithms

After shortfall losses break through the three buffers of margin, the liquidation engine, and the insurance fund in turn, the system finally enters the ADL stage (as shown in Figure 12-7). The core of ADL lies in deciding who bears the loss. When a shortfall loss exceeds the insurance fund's current balance, or the fund balance falls below a preset threshold, ADL is triggered. The system first identifies the size of the shortfall gap, then determines a deleveraging priority among the holders of opposing profitable positions according to a ranking algorithm, and force-closes at the bankruptcy price starting from the highest-ranked holder until the gap is completely filled. In mainstream implementations, these "opposing profitable positions" are strictly limited to opposing holdings in the same contract: when a BTC long falls into shortfall, the system selects ADL targets only among the profitable holders of BTC shorts, and never executes across contracts. A direct consequence of this design is that for low-liquidity long-tail assets, if long and short positions are highly imbalanced, the opposing profitable positions available for ADL are extremely limited—which is precisely the root cause

of the structural bad debt in the WLD case discussed in Section 12.2.3. When facing a systemic shortfall gap, an exchange must establish a clear set of rules to select the "sacrifices" from among its many profitable traders. The concrete embodiment of these rules is the ADL ranking algorithm. At present, the market's ADL algorithms have evolved into three main paradigms, each embodying a distinct design philosophy.

The oldest and most widely used is the queue-ranking algorithm based on the "product of profit rate and leverage." This ranking formula was introduced by BitMEX in 2016 (Huobi first introduced the ADL mechanism itself in 2015, but it lacked an estimable ranking), and it has since become the de facto industry standard, adopted by Binance (which adopted the formula in 2019), Bybit, and the decentralized platform Hyperliquid, among others [8]. Its core logic is that those who use the highest leverage and reap the largest unrealized profits are regarded as the group that has benefited most from the current extreme conditions and, simultaneously, contributed most to system risk. The system therefore scores and ranks all profitable accounts by the product of return and effective leverage. When ADL is required, the system force-closes the positions of these "top winners" one by one, strictly from highest to lowest, until the shortfall gap is completely filled. The advantage of this algorithm is its extremely high execution efficiency: it can eliminate the largest risk exposure with the fewest operations. Its core defect, however, is that it excessively concentrates its punishment on the most successful traders, giving rise to strong moral hazard and adverse selection. A selected trader may lose all of their profit or more, while the vast majority of profitable traders are entirely unaffected. More seriously, the queue-ranking algorithm has extremely low Sybil resistance: a sophisticated trader can split one large account into several small accounts to lower each account's ranking score, effectively evading the risk of being selected for ADL, so that those who ultimately bear the ADL impact are often the less sophisticated participants. Sybil-evasion strategies are not, however, costless: on centralized exchanges, operating multiple accounts runs up against the hard constraint of KYC identity verification; on decentralized platforms, spreading positions across sub-accounts requires each to maintain margin independently (forfeiting the capital efficiency of cross-margining), and managing multiple accounts increases the probability of operational error in extreme conditions.

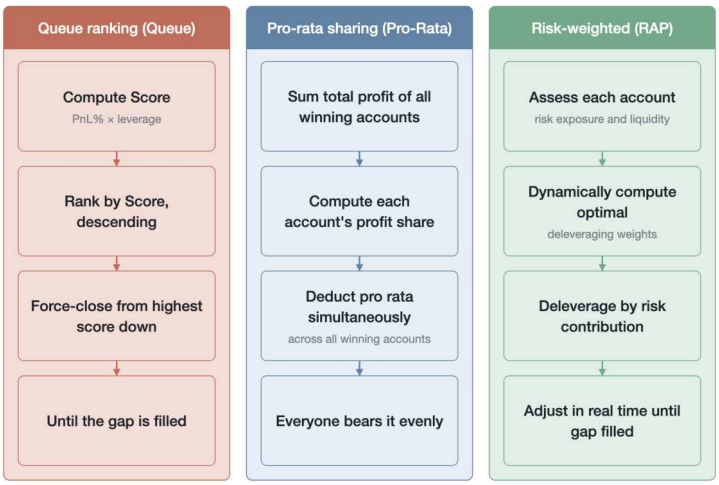


Figure 12-8. A comparison of the execution logic of three classes of ADL ranking algorithms

As shown in Figure 12-8, queue ranking concentrates its blows on the top winners; pro-rata allocation distributes losses evenly; and risk-weighted allocation dynamically allocates losses according to each account's aggregate risk exposure. To mitigate the extreme unfairness of queue ranking, some decentralized exchanges (DEXs, such as Drift and Paradex) have begun to explore pro-rata allocation [9]. This algorithm abandons the precise targeting of the "largest winners" and instead apportions the shortfall gap evenly across all profitable accounts in proportion to their profit. This means that regardless of leverage, any account in profit must pay a "compulsory tax" for the survival of the system. On its face, pro-rata allocation appears more moderate and fair—each person's loss is smaller—and its Sybil resistance is extremely high (splitting accounts does not change the pro-rata outcome). In actual execution, however, because it involves the simultaneous handling of a large number of tiny accounts, its computational complexity and execution latency increase markedly, and it often proves inadequate in the split-second urgency of extreme conditions. A deeper problem is that if pro-rata allocation ignores differences in risk, it causes low-risk traders to provide an implicit subsidy to high-risk traders. When this subsidy cost exceeds the trading utility that low-risk traders derive from the platform, they will rationally choose to exit, raising the platform's average risk level, which in turn forces still more low-risk traders to exit—a self-reinforcing chain of exits.

In recent years, the academic community has proposed a more advanced risk-weighted allocation algorithm. Represented by the research Tarun Chitra published in late 2025 [10], the RAP algorithm seeks a dynamic balance between efficiency and fairness. Rather than relying solely on static profit or leverage metrics, it incorporates an account's aggregate risk exposure, historical volatility, and the current market's liquidity depth, using a complex mathematical model to compute in real time each account's optimal deleveraging share. Its core idea is to align risk with cost: accounts

that contribute more to system risk should bear a larger deleveraging share, rather than being ranked simply by profit or leverage. Chitra's research further proposes the MDIC algorithm, which achieves an $O(\sqrt{T})$ regret bound, whereas any static strategy incurs linear regret $\Omega(T)$ in an adversarial environment. The core distinction here is this: in the face of an adaptive adversary, a static strategy (such as a fixed queue ranking or a fixed pro-rata allocation) accumulates regret that grows linearly with the time step T, meaning the gap between the strategy's performance and the optimal ex post benchmark widens uncontrollably; MDIC, as an online-learning algorithm, dynamically adjusts the allocation weights after each ADL event based on the observed market state, compressing cumulative regret to $O(\sqrt{T})$—that is, sublinear growth. This jump in order from $\Omega(T)$ to $O(\sqrt{T})$ is used in the paper's formal model to establish two propositions (these conclusions come from a preprint and await peer-review verification): first, under the premise of an adversarial market environment, any static ADL rule that does not adjust to the market state cannot achieve sublinear regret and is therefore suboptimal; and second, there exists a computable dynamic strategy that can asymptotically approach optimal performance [10]. Moreover, the asymptotic optimality of the online-learning framework (the advantage of an $O(\sqrt{T})$ regret bound relative to $\Omega(T)$) rests on the premise that T tends to infinity, whereas ADL events are extremely low-frequency (the 12-minute window of the 10/10 event produced only about 1,000 shocks, and most platforms experience only single-digit numbers of ADL events per year); when T is small (for example, $T < 100$), the practical numerical difference between the two is negligible. The engineering value of the MDIC algorithm may therefore lie more in its idea of dynamically adjusting weights than in a strict guarantee of asymptotic optimality. Combined with the computational complexity of the MDIC algorithm (each round requires solving a convex optimization problem) and the challenges of engineering implementation, it remains difficult to deploy at scale in a production environment and remains largely a subject of academic exploration.

12.2.2 The trilemma

A close analysis of the ADL mechanism reveals that it faces an inescapable theoretical dilemma: the ADL trilemma. In designing any ADL strategy, an exchange tries to achieve three core objectives simultaneously: maintaining the system's solvency, protecting the platform's long-term revenue, and ensuring fairness to traders. In his 2025 research, however, Chitra proves within his formal model that—under the objectives and the class of feasible strategies specified in that paper—no ADL algorithm can simultaneously dominate on all three dimensions of solvency, long-term revenue, and fairness (the three constitute a Pareto trade-off), and an exchange must make a difficult trade-off among them [10].

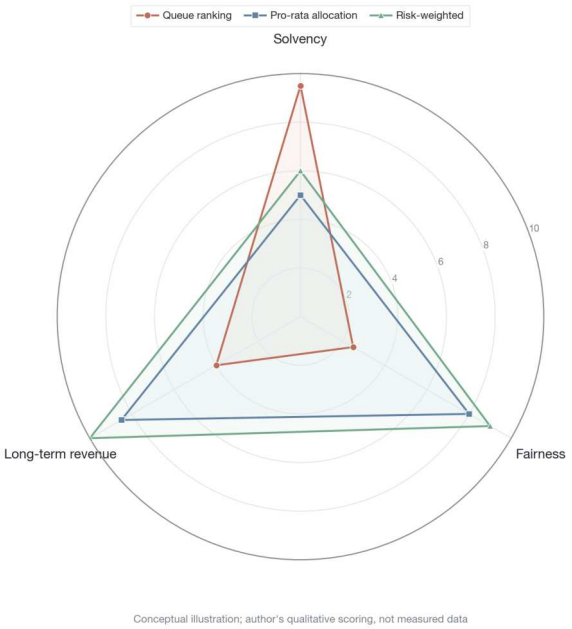


Figure 12-9. The trilemma performance of different ADL strategies (conceptual illustration, qualitative scoring by the author, not measured data)

Figure 12-9 presents, in radar-chart form, the trade-offs of different ADL strategies across the three dimensions: queue ranking has the highest solvency and the lowest fairness, pro-rata allocation is the reverse, and risk-weighted allocation seeks a balance. Solvency is the exchange's core constraint, requiring that ADL rapidly and thoroughly eliminate every shortfall gap. The most effective way to achieve this is to adopt an aggressive queue-ranking algorithm that directly cuts off the largest winners' positions. This approach, however, severely damages fairness, because it makes a tiny minority bear a disproportionate share of the systemic cost. Worse, this crude expropriation markedly reduces the willingness of high-net-worth traders to participate, causing them to leave the platform permanently after suffering ADL (that is, user attrition). The attrition of high-frequency traders and market makers directly drains platform liquidity and sharply reduces long-term fee revenue, thereby damaging the exchange's long-term revenue objective.

If, to protect long-term revenue and preserve fairness, an exchange chooses a gentler pro-rata allocation or sets a deleveraging cap, then in the face of a massive shortfall gap the force of ADL execution may fall far short. This not only fails to restore the system's solvency in a short time but may, through the continued spread of risk, trigger even larger-scale panic and a run. And if the exchange prioritizes protecting whale clients (high-net-worth traders) to preserve revenue, then small and medium traders must bear a larger share of the loss, fairness is further distorted, and systemic risk may keep accumulating because large positions are not disposed of in time. Within this trilemma, the prevailing mainstream queue-ranking algorithm has in effect chosen to preserve solvency at all costs, sacrificing fairness and long-term revenue.

To quantify this unfairness, Chitra introduces the profit-to-solvency ratio (PTSR) [10].

The PTSR measures the relationship between the profit a trader retains after suffering ADL and their contribution to the system's solvency. Mathematical analysis shows that the PTSR is of order $O(b_n/n)$, where b_n is the largest winner's profit and n is the total number of traders. For a light-tailed (thin-tailed) probability distribution, the largest winner's profit is asymptotically negligible relative to the total scale ($b_n = o(n)$), which means that as an exchange grows larger, ADL's "expropriation" of top winners tends toward completeness: in a sufficiently large exchange, the most successful traders tend to lose almost all of their profit. This conclusion is premised on the light-tail assumption: when the loss/profit distribution exhibits the power-law heavy tail described in Section 12.3.5 of this chapter (tail index $\alpha < 2$), b_n may be of the same order as the total scale, in which case this expropriation conclusion is weakened or even reversed—and this is precisely the scenario most likely to arise for long-tail/meme assets. Ideally, the PTSR should be close to 0.5, meaning risk-sharing is relatively balanced (the 0.5 benchmark here implicitly assumes a symmetric fairness criterion, namely that about half of the profit should be retained per unit of solvency contribution; under different theories of fairness—for example, the Rawlsian maximin principle that prioritizes protecting the weakest participant, or a utilitarian framework that maximizes total utility—the ideal PTSR value may differ). Under the traditional queue-ranking algorithm, however, the largest winner's PTSR is often below 0.001, reaching the extreme level of one in a thousand, meaning they are stripped of almost all of their legitimate profit. By contrast, the PTSR of pro-rata and risk-weighted allocation strategies ranges between 0.3 and 0.5, which is markedly fairer.

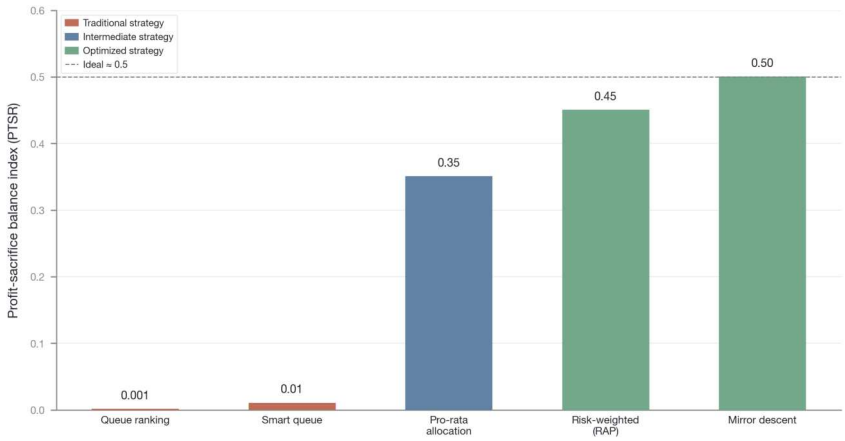


Figure 12-10. PTSR comparison across ADL strategies (PTSR values are constructed or illustrative by the author, based on the ADL research of Chitra and Gauntlet; the primary paper has not been verified; "smart queue" has no independent source, and only the qualitative ranking holds)

Figure 12-10 quantifies this difference in unfairness. From the perspective of a Stackelberg game, this trilemma carries a deeper dynamic implication. The game structure of ADL can be modeled as a sequential game. In a simultaneous game without commitment (Nash), a "bad equilibrium" exists: traders expect the severity of ADL to be high and therefore withdraw or reduce their positions; their withdrawal in turn reduces the total profitable positions available for deleveraging, forcing the system to impose a higher deleveraging intensity on the remaining participants in the next crisis—a vi-

cious cycle. In a Stackelberg setting, by contrast, the exchange can commit in advance to an incentive-compatible strategy (such as RAP or MDIC), sending a credible signal to the market that coordinates participants toward a "good equilibrium," in which traders believe deleveraging will be fair and are therefore willing to stay on the platform, providing the system with a more ample deleveraging buffer [10].

12.2.3 Formal analysis and empirical testing

For a long time, discussion of ADL remained largely at the level of experience and intuition, lacking a rigorous mathematical framework and large-scale empirical testing. Only in late 2025, when Tarun Chitra published the first rigorous formal analysis of the ADL mechanism [10], did the field gain a rigorous scientific footing. Chitra's research not only proved the existence of the ADL trilemma in theory but also, through high-fidelity tick-by-tick replay of a real historical event, revealed the enormous defects of existing mechanisms, providing a solid empirical basis for optimizing ADL algorithms.

The core empirical basis of this research is the extreme market crash that occurred on October 10, 2025 (known in the industry as the "10/10 event"). On that Black Thursday, hit by a sudden shock from macro geopolitical news, the crypto market triggered more than $19 billion in visible liquidation notional within just 24 hours, affecting roughly 1.62 million traders and setting the highest visible liquidation record on record [11]. On decentralized platforms such as Hyperliquid, the ADL mechanism was triggered frequently. Chitra's research focuses on the 12-minute window from 21:16 to 21:27 UTC on October 10, 2025; in that brief but violent period, Hyperliquid saw about $2.1 billion in position liquidations, involving about 160 assets and 19,337 wallets, and producing $T = 1{,}097$ ADL shocks, of which 201 produced a positive shortfall gap (that is, a genuine shortfall) [10].

By replaying tick-by-tick the 1,097 ADL shocks that occurred within these 12 minutes, Chitra's analysis reveals a striking phenomenon: over-deleveraging. The queue-ranking algorithm running in production, lacking dynamic awareness of the global gap, tends to mechanically close far more positions than actually needed. The data show that within these 12 minutes, the cumulative actual loss (that is, the dynamically accumulated value of the shortfall—the liability gap arising after account equity falls below zero across all shortfall events) was about $23.24 million, whereas the total profit available for deleveraging from all profitable traders (that is, the sum of the unrealized profit of the opposing profitable positions within the same time window, i.e., the reducible P&L allowance) was only $6.62 million. Yet the production queue-ranking strategy actually executed cumulative over-deleveraging as high as about $653.6 million (here "over-deleveraging" is measured by the **equity value** of the force-closed positions—that is, margin plus unrealized profit—rather than by a profit measure), with a single largest over-deleveraging of $47.1 million [10]. The seemingly contradictory difference in magnitude among these three figures (millions versus hundreds of millions) stems from two different measurement bases: the shortfall gap and the available profit are measured in "profit and loss (P&L) dollars" (counting only the profit/loss portion), whereas over-deleveraging is measured in "equity dollars" (the full position value, comprising margin principal plus profit). Based on public data, this book estimates that the average

equity-to-profit ratio of the affected traders was about 6.66 times (the paper itself does not directly provide this ratio), meaning that each $1 of reducible profit corresponds to about $6.66 of position equity [10]. The core problem of the queue-ranking algorithm is that it force-closes the **entire position** of the highest-ranked profitable trader in full, rather than reducing only the minimum share of profit needed to fill the gap. To take a simplified numerical example: suppose a BTC short falls into shortfall and produces a $100,000 shortfall gap, while the largest opposing winner holds a long position with $5 million of equity value (of which $4.25 million is margin and $750,000 is unrealized profit). The queue-ranking algorithm force-closes the entire $5 million equity position of that long to fill the $100,000 gap, of which $4.9 million of equity-value closeout is unnecessary over-deleveraging. It is precisely this crude logic of "closing a large position to eliminate a small gap" that leads the system to impose forced closeouts tens of times larger than necessary on profitable traders. All of the optimized alternative strategies (including Smart Queue, Mirror Descent, Vector Optimization, and Hybrid Blend) keep over-deleveraging near zero; measured by the total equity force-closed, this indicates that about 98% was over-deleveraging unnecessary for filling the gap—pure additional harm inflicted on profitable traders by the crude design of the ranking algorithm.

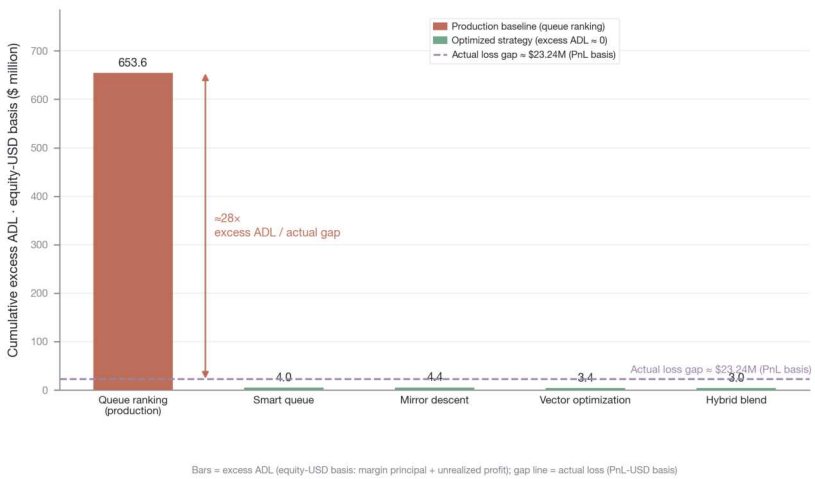


Figure 12-11. Comparison of cumulative over-deleveraging across ADL strategies (bar height is the equity-dollar measure of force-closed positions, including margin principal and unrealized profit, not the P&L actual shortfall gap; the actual loss was about $23.24 million, and over-deleveraging was about 28 times that)

Figure 12-11 compares the cumulative over-deleveraging of different strategies; the enormous gap between the production queue-ranking strategy and the near-zero optimized strategies (Smart Queue, Mirror Descent, and others) confirms the crudeness of static ranking rules. In an on-chain environment, executing 1,097 ADL shocks within 12 minutes involves a large number of state changes and fund transfers, and the operational risk is not to be underestimated: the uncertainty of transaction ordering, partial execution failures, and the challenge of guaranteeing state consistency—risks specific to on-chain systems—make the engineering implementation of ADL far more complex than a database transaction in a centralized environment.

More significant theoretically, Chitra's research identifies the absolute failure boundary of ADL in mathematical terms: the structural bad-debt floor. In the 10/10 event, some small-cap assets encountered a severe liquidity vacuum. The most extreme case appeared in the WLD asset: at one instant, that asset produced a single shortfall gap as high as about $11.56 million, while the maximum deleveraging capacity that all profitable holders of that asset could provide at that moment was only about $2,500 [10]. The gap between the two exceeded 4,600 times, for a coverage ratio of only 0.02%. Under such extreme asymmetry in microstructure, no ADL algorithm, however ingenious (whether queue ranking, pro-rata allocation, or the most advanced risk-weighted allocation), can mathematically fill the loss. This structural bad-debt floor shows that ADL is no panacea. When the market's one-sided tilt reaches an extreme—when a given asset's long and short positions are so severely imbalanced that nearly all participants stand on the same side—the socialization mechanism fails completely, and the system inevitably produces bad debt. Across the entire 12-minute window, at least $16.6 million in losses were structurally uncoverable by any ADL strategy, constituting an irreducible "hard bad-debt floor" [10].

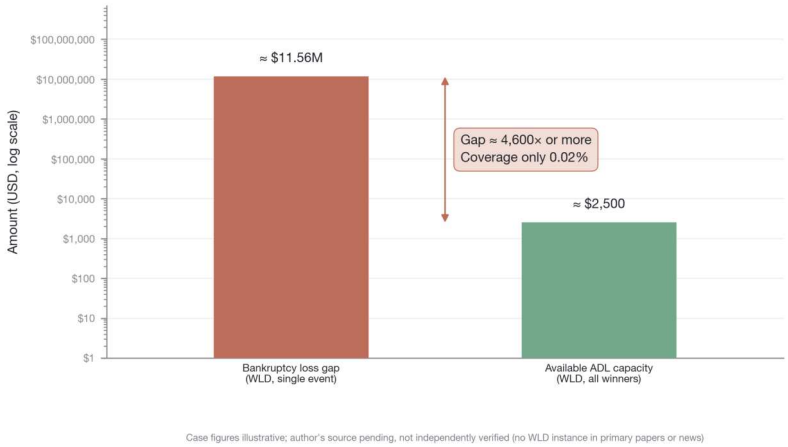


Figure 12-12. WLD's shortfall gap versus available deleveraging capacity (case values are the author's, with sources to be supplemented, or illustrative; neither the primary paper nor news reports contain the WLD instance, which is not publicly measured)

Figure 12-12 presents the stark contrast in this extreme WLD case, illustrating the mathematical failure of the ADL mechanism in the face of a structural liquidity imbalance. This finding carries far-reaching implications for institutional design. It means that even if the ADL algorithm itself is optimized to the extreme, the system's resilience still depends on the upstream mechanisms—including the insurance fund's adequacy ratio, the liquidation engine's efficiency, the prudent setting of risk parameters, and the quality of liquidity management. ADL is only the last resort, not a cure for the underlying condition. As the WLD case reveals, when the underlying liquidity structure itself has a critical flaw, no socialization-allocation mechanism, however ingenious, can avert the loss. This recognition redirects attention from "how to optimize ADL" back to "how to strengthen the upstream defenses," laying an empirical foundation for the later sections' discussion of insurance-fund governance and a re-engineering of the risk

waterfall.

12.2.4 ADL as a contrarian indicator

Although ADL means a passive loss of profit and position for traders, in the eyes of professional quantitative institutions and high-frequency market makers the frequency and scale of ADL triggers are a microstructural market signal with analytical value. From the perspective of market microstructure, ADL is in essence the concentrated eruption, at the liquidation level, of a severe scarcity of liquidity and a high concentration of one-sided positions. It is not only a contrarian indicator of an exchange's risk-management level but also a potential leading signal of turning points in the market sentiment cycle. This dual signaling value makes ADL a unique observational window connecting a platform's microstructural risk controls with the market's macro state: by analyzing the pattern of ADL triggers, one can infer both the mechanistic health of the platform and the sentiment extremes of the market.

ADL trigger frequency and scale are a key contrarian indicator of an exchange's overall risk-management level. When a trading platform begins to trigger ADL frequently, it sends the market several clear signals: the platform's insurance fund is exhausted or severely inadequate; its market-maker cohort has retreated, leaving the liquidation engine unable to find sufficient liquidity in the market to absorb bankruptcy orders; its risk parameters may be too aggressive (for example, allowing excessive leverage or listing illiquid long-tail assets); and the efficiency of its liquidation engine may be in question. Frequent ADL triggers suggest an inadequate insurance fund, an inefficient liquidation engine, overly aggressive risk parameters, or poor liquidity management. Conversely, a platform that has never triggered ADL does not necessarily have a well-designed ADL mechanism; rather, it means that its earlier layers of risk control (the margin system, the liquidation engine, and the insurance fund) are effective enough to have absorbed the risk at a more upstream stage.

The value of ADL as a contrarian indicator, however, goes well beyond assessing platform risk. Astute traders often treat ADL as a strong signal of a reversal in market sentiment. When a market crash is accompanied by large-scale ADL alerts, this usually means that panic selling has been exhausted and the short side's momentum is being forcibly consumed through this destructive process. A profitable short position closed by ADL is not executed by sending a buy order to the order book; it is settled off-book through a matched offset (match-off) between counterparties. ADL therefore does not directly generate buy-side order flow and produces no immediate buy-side price impact on the market. Nonetheless, by forcibly reducing short positions, ADL changes the balance of long and short forces in the market; moreover, if a trader who has been ADL'd loses their position and chooses not to rebuild it, this is tantamount to indirectly removing part of the sell-side force in the market. This indirect rebalancing of forces often produces a supportive effect on price after the peak of an ADL wave, driving a rebound or reversal. Conversely, at a bull-market top, if long positions begin to suffer ADL, it indicates that long leverage has reached a ceiling the system cannot bear, unsupported by real capital. This forced deleveraging, triggered by the collapse of the liquidation mechanism, is often a precursor to the end of the trend and an im-

pending deep correction. In this sense, ADL is not only an exchange's final mechanism for maintaining solvency but also a structural catalyst for the market sentiment cycle to shift from excessive optimism toward mean reversion.

As the WLD case reveals, however, even an optimal ADL algorithm is powerless in the face of a structural shortage of liquidity. The effectiveness of ADL as a contrarian indicator rests on one premise: that sufficient opposing positions still exist in the market to be deleveraged. When this premise breaks down—when a given asset's long and short positions are extremely imbalanced and nearly all participants stand on the same side—ADL can neither save the system nor provide a meaningful reversal signal. In such an extreme case, an ADL trigger marks a structural market failure rather than a cyclical sentiment extreme. Therefore, when using ADL as a trading signal, one must combine it with an in-depth analysis of the underlying liquidity structure, distinguishing a "sentiment-driven extreme" (a high probability of reversal after ADL) from a "structure-driven collapse" (possible further deterioration after ADL). This discriminating ability constitutes a core competitive advantage for professional traders in extreme conditions. The effectiveness of ADL as a contrarian indicator may itself decay as market participants learn—a classic manifestation of Goodhart's law in the domain of trading signals. If the pattern of "the market reverses after an ADL trigger" becomes widely recognized and traded upon, the signal's information content will gradually be replaced by a coordination effect. Worse, an attacker could deliberately trigger ADL to induce herd buying and then take profit in the rebound. The signaling value of ADL must therefore be evaluated dynamically within the framework of Lo's (2004) adaptive markets hypothesis, rather than treated as a static structural regularity [12].

12.3 The confidence-as-liquidity hypothesis

In analyzing systemic risk in derivatives markets, traditional financial models tend to focus on the absolute size of capital and the efficiency of the liquidation algorithm. In a crypto market that lacks a central bank as lender of last resort, however, the effectiveness of an insurance fund depends not merely on its book balance but, more importantly, on the confidence it establishes in the minds of market participants. This section proposes the confidence-as-liquidity hypothesis, which argues that under extreme market conditions the size of the insurance fund is itself a strong market signal: by shaping the confidence of traders and market makers, it directly determines the market's liquidity depth and, in turn, the system's true capacity to withstand shortfall risk. The insurance fund's "signaling function" carries more economic weight than its "absorption function." Confidence management is, in essence, risk management.

12.3.1 The transmission mechanism

As one of the market's few publicly observable indicators of systemic health, the insurance-fund balance transmits complex signals with every fluctuation of its figures. This signal acts directly on the belief layer of market participants, which changes their behavior and ultimately reshapes the entire market's liquidity environment. We can decompose this transmission process into four layers: the signal layer, the belief layer, the behavior layer, and the outcome layer.

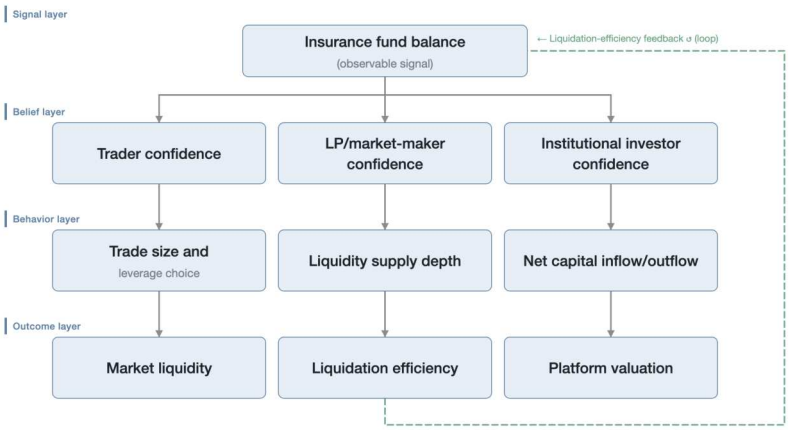


Figure 12-13. The confidence transmission mechanism of the insurance-fund balance

As shown in Figure 12-13, the observable insurance-fund balance signal acts in turn on the belief layer, the behavior layer, and the outcome layer, and forms a closed loop through the feedback of liquidation efficiency on the fund balance.

In the positive transmission chain, when market participants regard the insurance fund as "ample," a series of positive chain reactions unfolds. Users believe the system can withstand extreme events; market makers, facing lower shortfall risk, are willing to quote narrower bid-ask spreads; liquidity providers are willing to keep more capital in the vault (because they believe extreme losses will be absorbed by the fund); and traders are willing to hold larger positions and choose higher leverage. The consolidation of these beliefs translates directly into positive market behavior: trading volume expands, liquidity supply deepens, and capital keeps flowing in on a net basis. Finally, at the outcome layer, the market exhibits abundant liquidity and excellent liquidation efficiency—deeper liquidity, narrower spreads, and higher market quality. Efficient liquidation in turn keeps shortfall losses to a minimum, further protecting the insurance-fund balance and forming a virtuous ecological cycle.

The reverse transmission chain presents the exact opposite picture. When the insurance fund is regarded as "inadequate," or its governance is called into question, fear propagates in reverse along the same chain. Users worry that an extreme event could trigger ADL or a system collapse; market makers quickly widen spreads and cut quoting depth to demand higher risk compensation; liquidity providers withdraw capital from the vault ("I am unwilling to bear a risk I do not understand"); and traders reduce their positions ("I am not sure the platform is safe"). The end result is shrinking liquidity, widening spreads, and a sharp decline in market quality. The deterioration of liquidity in turn further lowers liquidation efficiency, raises shortfall risk, and places even greater pressure on the insurance fund.

12.3.2 The positive feedback loop

When the market suffers a sudden black-swan event that causes the insurance fund to shrink markedly, an extremely dangerous positive feedback loop of confidence collapse can be triggered. In traditional financial markets, such a loop is usually broken forcibly by emergency intervention from regulators or by a central bank's liquidity injection; but in a decentralized or inadequately regulated crypto market, this loop often evolves in free fall until some external force intervenes or the system collapses completely. This process closely parallels the classic mechanism of a bank run. More specifically, Diamond and Dybvig (1983), in their classic bank-run model, revealed the existence of multiple equilibria: in the "good equilibrium," depositors believe the bank is solvent and do not run, so the bank continues to operate; in the "bad equilibrium," depositors expect others to run and therefore rationally choose to withdraw first, so the run becomes a self-fulfilling prophecy [13]. Mapping this framework onto the behavior of insurance-fund LPs: when LPs believe the HLP vault's risk exposure is manageable, keeping capital in place constitutes the "good equilibrium"; but once LPs expect other LPs to withdraw in panic, even if the fund remains fundamentally adequate, a rational LP will choose to exit early to avoid being the last one out—which is precisely the dynamic depicted in the Hyperliquid whale incident of Section 12.3.3.

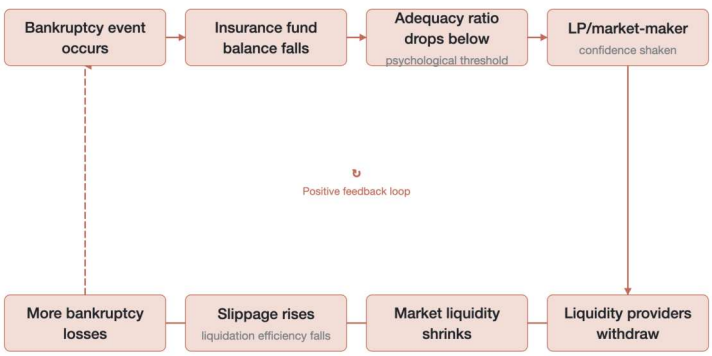


Figure 12-14. The positive feedback loop of confidence collapse and liquidity evaporation

The starting point of the loop is usually a massive liquidation the market fails to absorb smoothly, so that shortfall losses hit the insurance fund directly. When the fund balance falls and drops below participants' psychological expectation threshold, panic begins to spread among liquidity providers. To avoid potential systemic risk, market makers and liquidity providers instinctively choose to cancel orders and withdraw capital. The sudden withdrawal of liquidity further worsens an already fragile market depth. With insufficient buyers to absorb the flow, subsequent liquidation orders can only fill with greater slippage, which not only accelerates the crash in asset prices but also produces more severe shortfall losses. These new shortfall losses are again borne by the insurance fund, causing its balance to plunge further and triggering a new, more intense round of panic and withdrawal.

In this vicious cycle, the loss of confidence is directly equivalent to the evaporation of liquidity. The positive feedback loop keeps running until the vault shrinks to some minimum viable size, or confidence is restored by some external intervention—whether an emergency adjustment of protocol parameters, an injection of external capital, or transparent communication from the core team. Once the spiral begins, however, the cost and difficulty of breaking it rise sharply. This is precisely why preventive confidence management is far more important than after-the-fact crisis rescue.

12.3.3 Empirical evidence of the confidence multiplier

The Hyperliquid whale liquidation incident of March 2025 [14] provides quantitative empirical support for the confidence-as-liquidity hypothesis. The incident showed how a book loss can be amplified dozens of times through the confidence transmission mechanism, ultimately evolving into a severe liquidity crisis.

The incident began when a whale trader, at an address starting with 0xf3F, built a long Ethereum position on the Hyperliquid platform with a notional value exceeding $300 million (Arkham basis; at a liquidation price of about $1,915 and roughly 113,000 ETH, this works out to about $210 million, the difference stemming from different pricing at the position peak versus at the liquidation moment), opening at a notional leverage as high as 50x. When the market price moved unfavorably, the trader did not close out in the usual way but actively triggered the system's forced liquidation by strategically withdrawing collateral. The trader's choice of a "strategic liquidation" path rather than closing directly on the order book had a clear economic rationale: for a massive position with a notional value exceeding $300 million, closing directly at market would have caused a price impact of hundreds of basis points on the liquidity-limited order book, whereas transferring it to HLP through liquidation externalized this slippage cost to the liquidity providers. In essence, this was a form of arbitrage against the liquidation mechanism, using a gap in the rules to convert a private trading cost into a socialized loss borne by the public pool. Ultimately, the position was liquidated at a price of about $1,915, and the trader exited having cashed out a profit of about $1.8 million. Because the position was too large, however, the liquidity of the platform's public order book could not fully absorb this liquidation order, forcing the remaining position to be transferred to Hyperliquid's liquidity provider vault for backstop liquidation.

In handling this massive position, the HLP vault suffered an actual loss of about $4 million due to enormous market slippage, roughly 1% of the vault's total size (about $450 million to $480 million before the event). In a crypto derivatives market with trading volumes routinely in the hundreds of millions, an absolute loss of $4 million hardly seems fatal. Yet this loss triggered a strong psychological shock among liquidity providers. Vault participants suddenly realized that, as passive risk-bearers of an automated market maker, they not only shared in the trading-fee income but could also be forced to absorb this kind of hard-to-handle, high-risk liquidation order flow in extreme situations, bearing enormous socialized losses. Even more unsettling, the scale and frequency of this risk had not been fully understood or anticipated beforehand. The driver of the panic was not that "a $4 million loss is unacceptable," but rather the sudden realization that one's capital was exposed to a risk one did not fully

understand, that even greater losses might occur in the future, and that it would be better to withdraw first.

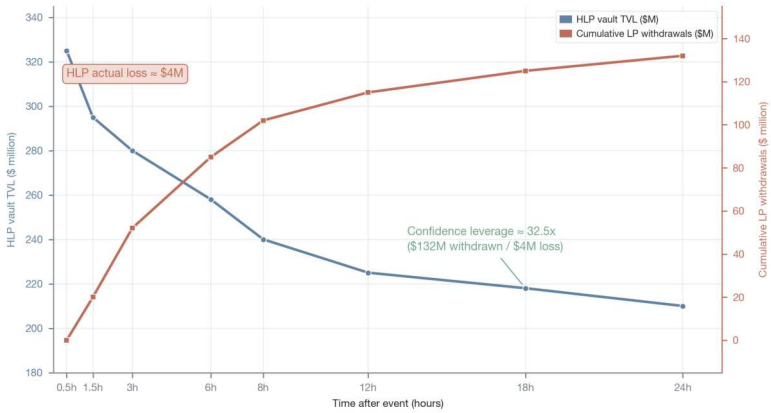


Figure 12-15. The confidence-leverage effect in the Hyperliquid whale incident (the intermediate series of HLP vault TVL and cumulative withdrawals is an on-chain estimate for illustration, with only the endpoints and the 32.5-times figure verified; 32.5 times is a correlational amplification, not strict causation)

The scissors gap between the two curves in Figure 12-15 captures the divergence between the decline in the HLP vault's TVL and the cumulative LP withdrawals.

Panic quickly turned into a run. Within 24 hours, more than $130 million was withdrawn in panic from the liquidity provider vault. A book loss of $4 million was accompanied by a liquidity outflow of $130 million, a ratio of about 32.5 times (this book's estimate, defined as the secondary panic-redemption amount divided by the primary direct vault loss). This striking correlational amplification factor reveals how fragile liquidity providers' confidence is in a decentralized architecture.

> The correlational amplification factor of about 32.5 times (a correlation measure, not a strict causal estimate) far exceeds the amplification of capital leverage: 50x leverage merely magnifies margin by 50 times, whereas confidence-related amplification expands the book loss more than 30-fold—and this "leverage" is not constrained by any margin system.

The 32.5-times confidence multiplier above comes from a single event (the March 2025 whale liquidation) on a single platform (Hyperliquid), and its external validity is subject to several limitations. First, as a newly emerging decentralized platform, Hyperliquid's LP cohort may differ systematically in risk perception and behavior from the LPs of mature centralized-exchange (CEX) platforms, who typically have richer experience with extreme events and higher risk tolerance. Second, the event occurred during Hyperliquid's early growth stage, when the platform had not yet experienced a similar shortfall event, and the LPs' panic stemmed partly from the "cognitive shock of first risk exposure," a factor that may weaken in subsequent similar events. Moreover, the 32.5-times amplification factor must also be treated with caution in causal attribution: LP withdrawal behavior may have been driven simultaneously by a contemporaneous

decline in macro risk appetite (such as the Fed policy-expectation shift of March 2025), the viral spread of panic narratives on social media, changes in competing platforms' yields, and other factors, rather than being a pure confidence response to Hyperliquid's single event. If other DeFi protocols also experienced synchronous TVL declines over the same period, then part of the LP withdrawals can be attributed to a systemic deterioration in risk appetite rather than a platform-specific confidence shock. The 32.5 times should therefore be understood as a "correlational amplification factor" after multiple factors are superimposed, and the multiplier actually attributable to pure confidence transmission may be considerably lower; it should be regarded as an observed value under specific conditions rather than a universal constant. Building a more robust confidence-leverage model would require systematic empirical comparison across platforms (CEX and DEX), across event types (manipulative shortfalls versus market-driven shortfalls), and across market cycles (bull and bear). Even so, even if the true confidence multiplier varies substantially across contexts, the core qualitative conclusion revealed by the Hyperliquid incident—that the amplification effect of a confidence shock far exceeds the book loss itself—remains robust.

The core mechanism revealed by the Hyperliquid whale incident is that the loss of confidence is not a linear response to "losses already incurred" but a fearful amplification of "losses that may occur in the future." The magnitude of the fear is severely disproportionate to the loss already incurred. A tiny crack in the books, once it touches participants' trust in the system's safety boundary, triggers a large-scale flight of capital. This nonlinear amplification effect has a solid theoretical basis in behavioral finance. Kahneman and Tversky's prospect theory holds that the psychological weight of a loss is about twice that of an equivalent gain [15], and in scenarios involving the perception of systemic risk this asymmetry may be amplified further. Liquidity providers' decisions are based not on a precise calculation of realized losses but on a subjective probability estimate of the worst future scenario. When an unexpected loss event updates their cognitive model of the system's fragility, their subjective probability of extreme future losses rises sharply, triggering a withdrawal far exceeding the actual loss. Prospect theory explains the psychological basis of panic from the angle of individual cognitive bias, but it overlooks the key amplifier of panic transmission in crypto markets: the information-cascade effect of social media. The herd-behavior theory of Banerjee (1992) and Bikhchandani et al. (1992) reveals that in an environment of information uncertainty, individuals rationally choose to follow others' behavior while ignoring their own private information [16][17]. Among crypto LP cohorts sharing the same Discord channels and Twitter feeds, the behavior of a few early withdrawers may, through an information cascade, trigger exponentially growing follow-on withdrawals, with an amplification effect far exceeding loss aversion at the individual level. If the temporal distribution of LP withdrawals exhibits the typical features of a cascade (a few early movers triggering exponential growth in followers), then social-media-driven herd behavior may be the dominant amplification mechanism behind the correlational amplification factor. This evidence shows that the "signaling function" of an insurance fund (or market-making vault) carries more economic weight than its "absorption function."

12.3.4 Confidence management as risk management

Faced with the enormous destructive power of the correlational amplification effect of confidence transmission, exchanges and decentralized protocols must elevate "confidence management" to a strategic priority on par with "balance-sheet management." For the operators of insurance funds and market-making vaults, managing confidence may be as important as managing capital—perhaps more so. The core of confidence management lies in stabilizing participants' expectations of the system's risk-absorption capacity through institutional design and transparency-building.

Effective confidence management depends first on a transparent communication mechanism. When a shortfall event occurs, the platform must promptly and accurately inform liquidity providers and users of the scale of the loss, its cause, and the countermeasures. An information vacuum breeds panic. As Shiller's (2019) narrative economics points out, the "viral narratives" that form in an information vacuum (such as "the protocol is about to go bankrupt" or "the team has already absconded") propagate in a dynamic closer to an epidemiological model, with contagiousness correlated with emotional arousal rather than with factual accuracy [18]. When participants cannot obtain authoritative firsthand information, speculation and rumor on social media quickly fill the void and amplify fear. In the Hyperliquid whale incident, the community's misreading of the nature of the event and its excessive worry about the system's safety stemmed in large part from the lag and incompleteness of information dissemination. Equally important as after-the-fact communication is the ex ante setting of expectations: before liquidity providers deposit capital, a platform should clearly convey "the extreme loss scenarios your capital may face," so that LPs make an "informed" decision when depositing rather than "suddenly discovering" in a crisis that they are exposed to unanticipated risk. In the Hyperliquid incident, many LPs panicked precisely because they had not previously understood the tail risk that the HLP vault bears as the backstop liquidator. Had these risks been fully disclosed and quantified in advance, the degree of panic would very likely have been much lower.

Another core issue in confidence management is the selective disclosure of the adequacy ratio. Publishing the insurance fund's actuarial adequacy ratio can enhance transparency, but if the adequacy ratio is poor, publishing it may instead trigger panic. This trade-off between transparency and stability parallels the "constructive ambiguity" strategy of central banks in traditional finance. A central bank does not always publish its full assessment of the banking system's fragility, because excessive transparency can, in some circumstances, become a self-fulfilling prophecy. For crypto-market insurance-fund managers, the key is to find a balance point: providing enough information for market participants to make rational judgments while avoiding the release, at sensitive moments, of signals that could be over-interpreted.

In addition, platforms need to build multi-layered defense in depth. A single insurance-fund pool is easily and quickly exhausted in the face of an extreme shock, triggering panic. By introducing an auto-deleveraging mechanism as a safety valve, establishing tiered and segregated insurance pools, and introducing external insurance capital, a platform can effectively slow the spread of risk and buy a precious window of time for market confidence to recover. The FTX affair, from the opposite direction, confirms

the institutional value of confidence management. According to the court testimony of co-founder Gary Wang [19], the insurance-fund size that FTX displayed to the outside world was entirely fabricated data generated by a formula containing a random number, bearing no relation to the fund's true balance. At one moment, FTX claimed the fund held $5.5 million and 5 million FTT tokens, when in fact the fund held no FTT at all. This fraud reduced the insurance fund from a risk-management tool to a mere instrument of false marketing, and when the truth came out it ultimately dissolved the foundation of user trust in the platform.

12.3.5 The cliff effect

A pronounced nonlinear feature exists between insurance-fund size and market confidence, which can be summarized as the "cliff effect." This effect indicates that market confidence does not decay smoothly and linearly as the insurance fund's adequacy ratio declines; rather, it collapses in cliff-like fashion within a specific threshold range. This nonlinear feature closely parallels the effect of bank capital adequacy on depositor confidence in traditional finance. The reason the minimum capital requirements under the Basel III framework are set at specific levels is precisely that regulators recognized that the impact of a decline in capital adequacy on market confidence is not linearly incremental but exhibits an exponential amplification near a critical point. In crypto derivatives markets, because there is no deposit-insurance system and no lender of last resort, this nonlinear feature manifests even more dramatically.

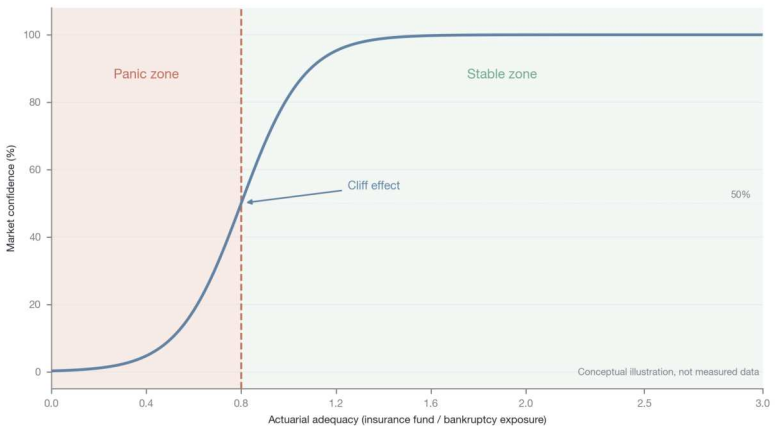


Figure 12-16. The nonlinear relationship between insurance-fund adequacy ratio and market confidence (conceptual illustration, not measured data; the threshold is an illustrative critical band, not a precise critical point, and the text likewise stresses that the threshold cannot be known precisely in advance)

We can define the actuarial adequacy ratio as the ratio of the insurance fund's current size to the maximum shortfall loss that might arise under expected extreme market conditions (measured by value at risk, VaR, or expected shortfall, ES). The academic roots of this definition trace back to the classic Cramér-Lundberg ruin model in actuarial science [20]. In that model, the insurance fund's balance is modeled as a stochastic process: premium income (corresponding to the steady inflow of liquidation penalties)

grows at a constant rate, while claims (corresponding to shortfall losses) arrive as a compound Poisson process. The ruin probability (the probability that the fund balance first falls below zero) depends on the ratio of the premium rate to the claim intensity and on the tail distribution of claim sizes. Applying this framework to a crypto derivatives insurance fund, the key modeling choice lies in the selection of the tail-risk measure: VaR measures the quantile threshold of losses at a given confidence level (the loss level exceeded with only a $1-\alpha$ probability) but ignores the severity of losses beyond that threshold, whereas ES (also known as CVaR) measures the conditional expected loss beyond the VaR threshold and better captures the "tail of the tail" extreme risk. Given the heavy-tailed character of the crypto market's shortfall-loss distribution (for example, WLD's single shortfall gap of about $11.56 million in the 10/10 event far exceeded that asset's VaR estimate of shortfall loss), ES is therefore a more prudent and appropriate choice of measure than VaR. Furthermore, if the shortfall-loss distribution exhibits a power-law tail (tail index $\alpha < 2$), then even ES may face the challenge of unstable estimation, requiring extreme-value-theory methods (such as fitting a generalized Pareto distribution) for more robust tail-risk modeling. The chapter's later adequacy analysis will therefore prioritize ES as the benchmark denominator. In operationalizing this definition, key parameter choices include the confidence level (99.5% or 99.9%, whose estimates may differ severalfold under a heavy-tailed distribution) and the time window of the extreme scenario (the maximum single-day loss or the cumulative loss over several consecutive days). These parameter choices have an enormous impact on the numerical value of the adequacy ratio, so cross-platform comparisons must ensure the use of a consistent parameter baseline. Two core assumptions of the classic Cramér-Lundberg model face significant departures in crypto markets. First, the model assumes that claim arrivals follow a compound Poisson process (that is, claim events are mutually independent), but shortfall events in crypto markets are highly clustered—for example, 1,097 ADL shocks within 12 minutes in the 10/10 event—and this tail clustering severely violates the independence assumption of Poisson arrivals, requiring self-exciting models such as the Hawkes process to more accurately capture the clustering of shortfall losses. Second, the model assumes a constant premium rate (corresponding to liquidation-penalty income), but insurance-fund income is itself procyclical (higher in bull markets, and potentially negative in extreme conditions precisely because liquidation volume surges while payouts explode simultaneously), and this comovement of income and expenditure further weakens the model's applicability. When the adequacy ratio is at a high level (that is, in the stable zone), market participants generally take the system's safety for granted. Within this zone, whether the insurance fund grows or shrinks by tens of millions of dollars has no material effect on market confidence. Traders' and market makers' attention is focused mainly on finding trading opportunities and optimizing market-making strategies, and the fund's existence draws no attention.

When frequent shortfall events cause the insurance fund to be continuously depleted, however, and the adequacy ratio approaches and drops into a certain psychological threshold range, the situation changes dramatically. Within this narrow range, even a tiny further loss is amplified by the market and interpreted as a precursor to the system's imminent bankruptcy. At this point, market sentiment flips instantly from

"absolutely safe" to "extreme panic," and participants' behavior shifts from seeking profit to fleeing at any cost. This sharp collapse of confidence directly triggers the liquidity withdrawal and positive feedback loop described earlier, ultimately leading to a comprehensive collapse of the system. From "no one cares" to "everyone panics," there is almost no smooth transition zone in between.

A key feature of the cliff effect is that the location of the threshold cannot be known precisely in advance. It depends on multiple factors: user psychology, the market environment, the memory effect of recent events, and the emotional amplification of social media. In a bull market, participants' risk tolerance is higher, and the threshold may be pushed to a lower level; in a bear market, or just after a major security incident, the threshold shifts markedly upward, and the system becomes more fragile. This uncertainty means that insurance-fund managers face a challenge that "cannot be precisely optimized": staying above the threshold requires an ample safety margin, but too high a safety margin means a loss of capital efficiency.

As shown in Figure 12-6, different exchanges differ enormously in insurance-fund size and relative adequacy ratio. Binance, with an absolute size exceeding $2.3 billion, far surpasses other platforms, but in terms of the ratio of fund size to total open interest, the picture changes markedly. Hyperliquid's relative ratio of about 18.6% (calculated on total HLP TVL) far exceeds Binance's 6.6% (OKX, because it discloses only a security fund rather than a futures liquidation insurance fund, has a ratio that is not comparable on a like-for-like basis; see Section 12.5), but as noted earlier this ratio may be overstated because of the distinction between total HLP TVL and the liquidation vault's standalone size. Even so, this difference still reflects that decentralized platforms, lacking centralized credit backing, must rely on a thicker capital buffer to sustain community confidence. For a decentralized protocol, the relative size of the insurance fund is the on-chain equivalent of its credit rating.

The significance of understanding the cliff effect lies in its warning to risk managers: one must never wait until the insurance fund is about to be exhausted before taking remedial measures. Once the system falls off the cliff, the cost of rebuilding confidence will be astronomical, far exceeding the capital cost of maintaining an ample safety margin in advance. A platform must therefore set an ample safety buffer at the edge of the cliff and, when the adequacy ratio touches the warning line, decisively initiate capital replenishment or a risk circuit breaker, resolutely defending that fragile yet decisive psychological line. It is thus evident that managing an insurance fund is both an actuarial problem and a problem of confidence governance and institutional constraint.

Synthesizing the foregoing analysis of confidence transmission, the positive feedback loop, and the cliff effect, we can construct a unified theoretical framework to explain how insurance-fund depletion triggers a nonlinear system collapse. This framework comprises two levels of mechanism: a macro-level threshold effect and a micro-level amplification mechanism. At the macro level, the cliff effect implies that the insurance fund's adequacy ratio has a critical threshold, below which market confidence undergoes a phase transition, jumping irreversibly from a "stable equilibrium" to a "panic equilibrium." At the micro level, confidence collapse is transmitted through three amplification channels: market-maker withdrawal leading to liquidity evaporation (price-

impact amplification), LP runs further weakening the insurance buffer (capital-buffer amplification), and trader deleveraging aggravating order-flow imbalance (liquidation-cascade amplification). The coupling of the macro threshold and micro amplification produces the characteristic nonlinear collapse dynamic: once the adequacy ratio breaks below the critical point, the three micro amplification channels activate simultaneously, and the positive-feedback interaction among them makes the system's recovery path far longer than its collapse path—which is precisely why "the cost of rebuilding confidence is dozens of times the cost of maintaining it." This "macro threshold plus micro amplification equals nonlinear collapse" framework provides a unified analytical lens for understanding the failure modes of insurance funds in successive crypto-market crises, and it also furnishes a theoretical anchor for the later sections' discussion of re-engineering the risk waterfall.

12.4 The governance dilemma and the governance spectrum

One of the core contradictions of crypto derivatives markets is that the system must respond to extreme and unpredictable market risk, which often requires granting managers some form of discretion; yet at the same time, the core of the crypto ethos is "code is law," which stresses the elimination of human intervention and trust assumptions. This contradiction is especially prominent in the management of insurance funds and the design of liquidation mechanisms, forming a governance dilemma that is difficult to reconcile. Just as central banks in traditional finance perpetually oscillate between "rules" and "discretion," crypto derivatives platforms likewise face the challenge of finding a balance between certainty and flexibility. Because of the immutability of blockchain technology and the automatic execution of smart contracts, however, the search for this balance becomes more complex and fraught with tension.

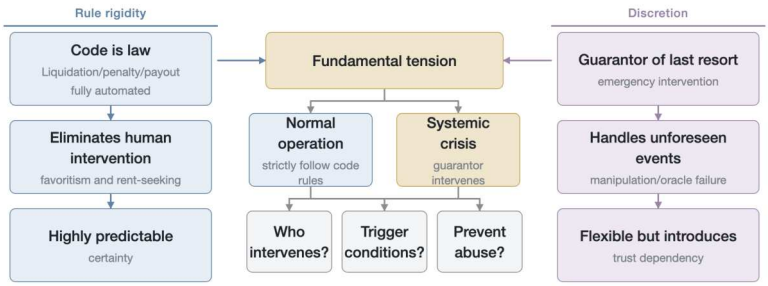


Figure 12-17. The tension between rule rigidity and discretion in insurance-fund governance

12.4.1 The limits of rule rigidity

"Code is law" is the cornerstone of decentralized finance. In an ideal decentralized derivatives platform, the triggering of liquidation, the collection of penalties, insurance-fund payouts, and the execution of auto-deleveraging should all be completed automatically by pre-written smart contracts. This rule rigidity brings a high degree of certainty and transparency, thoroughly eliminating the moral hazards common on centralized platforms—black-box operations, favoritism toward large clients, or internal rent-seeking. Before participating in the market, a trader can know clearly at what price they will be liquidated and how large a penalty they will face, and no one can arbitrarily change these rules.

The complexity of financial markets, however, often exceeds the preset range of code logic. The greatest limitation of rule rigidity is its inability to respond effectively to "unknown unknowns." A smart contract can only execute according to preset parameters and lacks the ability to understand the context of extreme anomalies. When the market encounters an unprecedented liquidity vacuum, when congestion on the underlying blockchain network prevents transactions from being included, or—more commonly—when the oracle price feed is maliciously manipulated, strict execution of the code may instead lead to systemic catastrophe. In these extreme scenarios, if a platform adheres entirely to rule rigidity and lets the liquidation engine blindly dump assets, it will not only aggravate the vicious cycle of prices but may also exhaust the insurance fund in an instant, ultimately triggering large-scale auto-deleveraging that makes innocent profitable traders bear enormous losses.

A further problem of rule rigidity lies in its "predictability" to attackers. When all rules are transparently encoded on-chain, an attacker can precisely calculate the cost and expected return of manipulating the market and thereby design a targeted attack strategy. This predictability does not exist in traditional finance, because regulators and market makers can increase attackers' uncertainty through unpredictable discretionary actions. Therefore, even a platform that touts a high degree of decentralization often has to find some form of emergency intervention mechanism when facing a life-or-death crisis—which brings us to the problem of the "guarantor of last resort."

12.4.2 The guarantor of last resort and discretion

To make up for the deficiencies of rule rigidity, a system often needs to introduce the role of a "guarantor of last resort" or "emergency administrator." On centralized exchanges, this role is naturally played by the platform operator; on decentralized platforms, it may take the form of a core team holding multi-signature authority or a specific governance committee. The guarantor of last resort is granted discretion to suspend trading, modify liquidation parameters, or even roll back transactions in emergencies.

This discretion has an irreplaceable function in responding to a systemic crisis. It endows the system with a kind of "elasticity," allowing the platform to block the spread of risk through human judgment when the code logic fails. Introducing discretion, however, inevitably reintroduces trust assumptions, which directly contravenes the original trustless intent of cryptocurrency. More seriously, once the power to intervene in the system is granted to some person or organization, one inevitably faces a series of thorny

governance problems: who intervenes, what the trigger conditions are, and how to prevent the power from being abused.

The Hyperliquid JELLY incident of March 2025 provides a canonical empirical case of this dilemma. The attacker built two-way positions on the JELLY perpetual futures and actively withdrew margin to trigger self-liquidation, forcing the platform's liquidity provider pool to take over the short side; the attacker then pumped the JELLY spot price on the Raydium exchange on the Solana chain, causing HLP to bear an enormous unrealized loss (the mechanical details—the number of addresses, the size of the two-way positions, the magnitude of the pump, the loss amount, and the rollback price—are elaborated in Section 12.6.2) [21]. Here it is necessary to distinguish precisely between two manipulations of different natures: what the attacker carried out was "manipulation of the underlying asset price" (changing the asset's actual execution price by putting real money into a thin spot market), not "oracle signal injection" (directly tampering with the oracle's data transmission through technical means). In the JELLY incident, the oracle faithfully relayed the manipulated "real" market price, and its own technical integrity was not compromised. This distinction carries important implications for institutional design: defending against manipulation of the underlying asset price requires market-structure measures such as liquidity thresholds and position caps, whereas defending against oracle signal injection requires technical measures such as cryptographic verification and multi-source aggregation. Faced with this crisis, Hyperliquid's validator nodes voted in an emergency to override the oracle price, forcibly rolling back JELLY's settlement price to the level before the attack occurred, thereby avoiding a massive loss to the insurance fund [22].

This incident displays the two-sidedness of the guarantor-of-last-resort role. On the positive side, the validators' emergency intervention did indeed stop a planned market-manipulation attack and protected the interests of the insurance fund and the vast majority of users. On the negative side, however, this forced overwriting of the price triggered widespread doubt in the community: if validators can override the oracle price at any time, then what is the essential difference between Hyperliquid and a centralized exchange? Binance founder Changpeng Zhao went so far as to publicly criticize Hyperliquid's approach, questioning the authenticity of its decentralization. Viewed from a legal perspective, validators forcibly settling users' positions at a non-market price may create legal risk in multiple jurisdictions. At the level of contract law, even if the user agreement contains an "emergency" clause, the enforceability of that clause varies considerably across jurisdictions. At the level of property rights, for non-attackers holding long JELLY positions, the choice of the price benchmark for a "full refund" is itself a point of legal dispute. At the regulatory level, an entity with central counterparty (CCP)–like functions under the EU's Markets in Crypto-Assets (MiCA) framework unilaterally changing a settlement price may trigger scrutiny. The back-and-forth of Eisenberg being charged, convicted, and then having the conviction vacated in the Mango Markets case (see Section 12.4.3) exposes precisely the legal uncertainty of criminal prosecution for crypto-market manipulation on questions of venue and the elements of fraud; at the same time, whether the platform's intervention exceeded the reasonable scope of self-help could also become the focus of future litigation. This legal-dimension uncertainty adds the further complexity of cross-border legal coor-

dination to the "residual control rights" problem discussed in Section 12.4.4. Historical experience shows that unconstrained discretion easily degenerates into an infringement of users' interests. Even with the purest of motives, the boundary of intervention is highly ambiguous. For example, when a large trader's collapse might trigger a chain of liquidations, a platform's choice to suspend liquidation to protect the insurance fund is in effect a disguised bailout of that large trader—and grossly unfair to other users who trade by the rules. This "too big to fail" phenomenon, long criticized in the traditional financial system, now reappears in another form in crypto markets.

12.4.3 Oracle risk

Oracle risk is a central topic in the governance dilemma. The oracle is the bridge connecting off-chain real-world price data with on-chain smart contracts, and it is also the most fragile link in a decentralized derivatives platform. No matter how tightly a platform's internal liquidation logic is designed, once the mark price the oracle feeds in is distorted, the entire system's risk-control mechanism fails in an instant. Take Hyperliquid: it uses a self-built oracle system that feeds prices aggregated from the spot prices of multiple external exchanges. For a low-liquidity asset like JELLY, however, its spot trading may be concentrated in the liquidity pools of just one or two decentralized exchanges, leaving the oracle's data sources severely lacking in diversity, so that the cost of manipulation is correspondingly extremely low. This reveals a key design flaw: for low-liquidity assets, an oracle should set a minimum data-source-diversity threshold, and if the available data sources fall below the threshold, the system should automatically limit the maximum position size on that asset or suspend its derivatives trading. The reason oracle risk occupies a special place in insurance-fund governance is that the shortfall losses it triggers arise not from normal market risk but from malicious manipulation of external inputs, which renders traditional risk-management frameworks almost entirely ineffective against such attacks.

Oracle manipulation attacks take multiple forms in practice. The most direct is spot-price manipulation, in which an attacker artificially pumps or dumps a token's price in a poorly liquid spot market with enormous capital, causing the oracle to capture the abnormal price and transmit it to the derivatives platform. A more covert approach exploits the flash-loan function of DeFi protocols to borrow an enormous sum within a single transaction to manipulate the price, repaying the loan immediately after completing the attack, with the entire process completed within one block. In addition, an attacker can exploit the gap in the oracle's update frequency to conduct arbitrage before the oracle reflects the latest price. According to Chainalysis estimates, in 2022 alone DeFi protocols lost about $403.2 million to oracle manipulation attacks across 41 separate incidents (this figure is explicitly labeled by Chainalysis as an estimate, and its statistical scope covers the entire DeFi ecosystem rather than being specific to perpetual futures) [23]. According to other estimates by security researchers, since 2020 oracle manipulation has accounted for about 15% to 20% of major DeFi security incidents (this proportion lacks a uniform definition of "major incident" and a public methodology, and should be treated with caution) [24]. The statistics above come mainly from publicly disclosed security-incident reports and on-chain data analysis, and the actual loss scale may be underestimated because of undisclosed incidents.

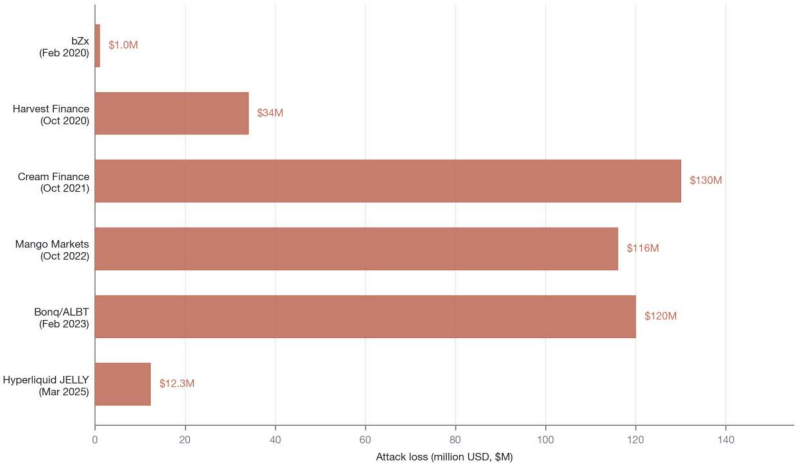


Figure 12-18. The loss scale of major oracle manipulation attacks in DeFi history (Bonq/ALBT's $120 million is the minted notional; constrained by wALBT liquidity, the actual amount extracted was about $1.7 million to $4 million)

Figure 12-18 shows the evolution of the loss scale of oracle manipulation attacks—from the millions to the hundreds of millions since the 2020 bZx incident—reflecting the persistent fragility of DeFi protocols in protecting against external price inputs.

The Mango Markets incident of October 2022 is a paradigmatic case of oracle manipulation. The attacker, Avraham Eisenberg, exploited the weak spot liquidity of the MNGO token to artificially manipulate the oracle price by building positions simultaneously in multiple markets, then used the inflated book profit to borrow various crypto assets worth as much as $116 million from the protocol [25]. Although the protocol's smart contract ran entirely according to its established rules, with no code vulnerability, the false price fed in from outside directly bankrupted the protocol. Eisenberg was charged by the U.S. Department of Justice in January 2023 with commodities fraud, commodities manipulation, and wire fraud, and was convicted by a jury in April 2024; but on May 23, 2025, the U.S. District Court for the Southern District of New York, invoking Rule 29 of the Federal Rules of Criminal Procedure, vacated all criminal convictions on the grounds of improper venue and insufficient elements of wire fraud (Mango Markets had no terms of service, did not prohibit the relevant operations, and had no repayment agreement, so there was no material misrepresentation); the prosecution has appealed, and as of this writing the case remains before the Second Circuit Court of Appeals [26][27].

Oracle risk poses a fundamental problem for insurance-fund governance: how should shortfall losses caused by non-market risk be handled? If the insurance fund pays out according to the established rules, then the attacker is in effect "legally" stealing funds from the insurance fund; if the platform chooses to intervene and refuse to pay, then we return to the discretion-abuse problem discussed earlier. There is no perfect solution to this dilemma; one can only reduce the probability of a successful attack through more refined oracle design (such as time-weighted average price, multi-source aggregation, and outlier filtering) while, through governance mechanisms, standardizing the trigger

conditions and execution process of emergency intervention.

12.4.4 The theory of incomplete contracts

To understand this governance dilemma deeply from a theoretical standpoint, we can invoke the "incomplete contracts" theory from economics. Proposed by Oliver Hart (co-winner of the 2016 Nobel Prize in Economics with Bengt Holmström) and John Moore, this theory holds that because of humans' bounded rationality and the unpredictability of the future, contracting parties cannot foresee, ex ante, all the situations that might arise and write them into the contract [28]. All contracts are therefore "incomplete" in essence.

In the cryptocurrency context, the smart contract is the ultimate embodiment of such an "incomplete contract." Although a smart contract is called a "contract," it is in fact merely a piece of computer code deployed on a blockchain. When writing the code, developers can only set parameters based on their understanding of past market behavior and their predictions of a limited set of future scenarios. They cannot foresee every form of market manipulation, hacker attack, or extreme black-swan event. When these situations not covered by the code occur, the smart contract exhibits a "blank" or "loophole." The developers of Mango Markets could not have foreseen that someone would manipulate the price of a small-cap token in such an extreme way; nor could Hyperliquid's designers have exhaustively enumerated every possible cross-chain attack vector in the code.

The Hart-Moore model further points out that when a contract is incomplete, the key question is who holds the "residual control rights"—the power to make decisions in situations the contract does not specify. On a traditional centralized exchange, the platform holds absolute residual control rights, which gives it great flexibility but also creates the platform moral hazard discussed earlier. In the ideal model of full decentralization, an attempt is made to eliminate residual control rights entirely, leaving everything to code execution. Because of the inherent incompleteness of smart contracts, however, completely eliminating residual control rights is impractical. When unanticipated events such as oracle failure or a market collapse occur, someone or some mechanism must handle the situations not covered by the code, exercising residual control rights to save the system.

This theoretical framework offers a key perspective for understanding the governance dilemma of crypto derivatives: the core of the problem is not "whether discretion is needed" but "how to allocate and constrain discretion." An ideal institutional design should follow a "two-tier architecture": in day-to-day operation, strictly adhere to the code rules to ensure certainty and fairness; and when a systemic crisis occurs, authorize a guarantor of last resort—constrained by strict procedures—to intervene, but with the scope of intervention, the trigger conditions, and the after-the-fact accountability mechanism all clearly specified in advance. This is precisely why even the most hardcore DeFi protocols ultimately tend to retain some form of governance mechanism or emergency pause function. The essence of the governance dilemma is how, under the ideal of decentralization, to appropriately allocate and constrain this unavoidable residual control right.

12.4.5 Information asymmetry and signaling games

Before turning to the governance spectrum, consider the deeper logic of insurance-fund governance from the perspective of information economics. In his seminal paper on the "market for lemons," George Akerlof revealed how information asymmetry leads to market failure: when the seller holds quality information that the buyer cannot observe, high-quality products are driven out of the market by low-quality ones [29]. Mapping this framework onto the governance of crypto derivatives insurance funds, the "lemon problem" appears in a distinctive form: the platform (the seller) holds the fund's true condition (including its actual balance, risk exposure, historical payout record, actuarial adequacy ratio, and other private information), while users and LPs (the buyers) can observe only the public signals the platform chooses to disclose selectively.

Under this information asymmetry, Spence's signaling-game theory [30] provides a key perspective for understanding platform behavior. A high-quality platform (with an adequate insurance fund and transparent governance) has an incentive to distinguish itself from low-quality platforms by sending "costly signals." These signals include publishing on-chain proof of reserves, or PoR (which can be independently verified and is costly to fake), accepting third-party actuarial audits (which require real data to support them), and encoding liquidation rules completely into auditable smart contracts (which, once deployed, cannot be unilaterally tampered with). These signals are effective because the cost for a low-quality platform (one whose fund is inadequate or that engages in misappropriation) to mimic them is extremely high: a platform with a fictitious insurance fund cannot pass the verification of on-chain proof of reserves, and a platform with opaque rules cannot bear the exposure risk of a smart-contract audit. The FTX case is precisely a cautionary counterexample: it sent false signals by fabricating insurance-fund data, and its eventual implosion proved that false signals are unsustainable in the long run.

The logic of signaling games leads directly to the analysis of the governance spectrum. The progression from G0 to G3 is, in essence, a gradual upgrade of the "quality signal" a platform sends to the market, from cheap talk to a verifiable commitment.

12.4.6 The governance spectrum G0–G3

Faced with the tension between rule rigidity and discretion, crypto derivatives platforms have not remained stuck in a black-or-white binary but have explored, in practice, governance models of varying degrees. Taking governance transparency, the degree of decentralization, and the constraints on discretion as yardsticks, one can depict a "governance spectrum" running from a fully centralized black box to fully on-chain autonomy. We can divide it into four main levels: G0 through G3 (Table 12-1). This classification framework echoes the verifiability ladder discussed in Chapter 5 of this book; the governance level of the insurance fund is, in essence, a key dimension of a platform's overall verifiability.

Level	Name	Transparency	Constraint on discretion	Representative platforms
G0	Fully opaque	Size/source/rules all undisclosed	No constraint whatsoever	Early BitMEX, FTX
G1	Size visible	Publishes PoR or on-chain balance	Governance process opaque	Binance, OKX (current)
G2	Size + rules visible	Rules encoded on-chain, auditable	Upgrade authority concentrated in the core team	Hyperliquid
G2–G3	Rules visible + partial on-chain governance	On-chain voting + timelock already running in production	Upgrade authority constrained by DAO governance, but emergency powers retained	dYdX v4
G3	Fully on-chain governance	Upgrades require community vote + timelock	Discretion programmatically constrained	Theoretical standard (none yet fully achieved)

Table 12-1. Definition of the G0–G3 insurance-fund governance spectrum and representative platforms (Data source: constructed by the author)

The G0 level represents fully opaque black-box governance. At this level, the insurance fund's size, funding sources, payout rules, and liquidation logic are all undisclosed. The platform operator holds absolute discretion subject to no oversight whatsoever. Early BitMEX and the later-collapsed FTX are typical representatives of this level. In the G0 model, users can only blindly trust the platform's moral integrity, and history has repeatedly shown that such trust is often betrayed, with serious consequences. The FTX insurance-fund fraud case shows that at the G0 level the insurance fund can even be entirely fictitious, with users utterly unable to detect it.

The G1 level introduces a preliminary degree of transparency and can be described as "size visible, governance opaque." Platforms at this level begin to publish proof of reserves periodically or disclose the on-chain wallet address of the insurance fund, allowing the public to verify that the funds truly exist. How these funds are used, however, along with the specific parameters of the liquidation engine and the decision process in emergencies, still rests in the hands of a centralized team. Most mainstream centralized exchanges on the market today, such as Binance and OKX, sit largely at this level. They build trust by displaying a huge insurance-fund size, but they still retain broad discretion in rule enforcement. The advance of G1 over G0 is that users can at least verify "whether the money is really there," though "how the money will be used" remains unknowable.

The G2 level is a key step toward decentralization, characterized by "both size and rules visible, but upgrade authority concentrated." At this level are mainly the new generation of decentralized derivatives protocols. They encode liquidation logic, penalty collection, and the insurance fund's operating rules completely into smart contracts deployed on-chain for anyone to audit. The system's day-to-day operation achieves "code is law." To cope with the risk brought by incomplete contracts, however, the core

team still retains the power to modify key parameters or upgrade contracts through means such as multi-signature. Hyperliquid's performance in the JELLY incident is a clear depiction of the G2 level: day-to-day operation is fully transparent and automated, but at the moment of crisis the validators (in substance controlled by the core team) can intervene in an emergency. This model represents a qualitative leap in transparency, but it still carries the single-point risk of the team acting unilaterally in bad faith or being coerced by regulators.

The G3 level represents the highest standard the industry is currently exploring: "fully on-chain governance with timelock constraints." At this level, not only are the rules and funds fully transparent, but residual control rights are handed to the decentralized community (usually through token voting). Any modification of system parameters, upgrade of contracts, or use of the insurance fund must go through an on-chain proposal and voting process. More importantly, the system introduces a timelock mechanism: after a vote passes, a change to the code must go through a mandatory waiting period (usually 24 to 72 hours) before taking effect. This gives users who disagree with the decision ample time to withdraw their funds, providing an exit-right guarantee of "voting with one's feet." dYdX v4, through its independent app-chain architecture based on the Cosmos SDK, is actively exploring this direction. The G3 model, through a programmatic democratic process and a time buffer, constrains the abuse of discretion to the greatest extent possible while preserving the system's elasticity.

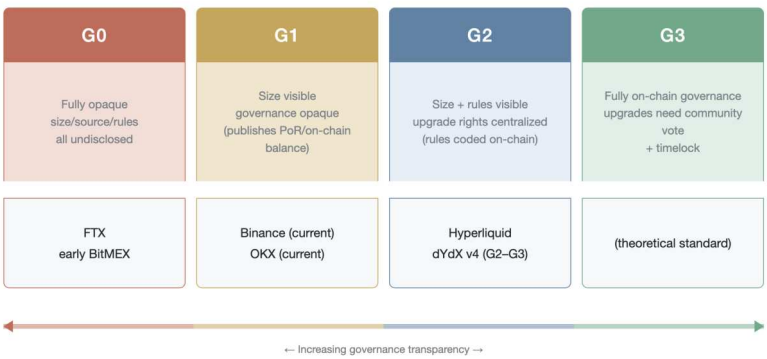


Figure 12-19. The governance spectrum of crypto derivatives platforms and the positioning of major platforms

Figure 12-19 shows the positioning of various platforms on the G0–G3 governance spectrum, with the industry as a whole exhibiting a gradual evolution from low to high transparency.

Even the G3 level, however, is not flawless. Fully on-chain governance faces real-world challenges such as low voter turnout, concentration of governance tokens, and insufficient emergency-response speed. When the system is under attack, waiting for a community vote and the expiry of the timelock may mean a delay of hours or even days, during which the attacker may already have completed the transfer of funds. Some protocols therefore introduce an "emergency multisig" as a safety valve within the G3 frame-

work, allowing intervention that bypasses the normal governance process in extreme situations, subject to after-the-fact community review and ratification. Although this design sacrifices some purity of decentralization, it strikes a pragmatic balance between practicality and safety. In addition, the DAO voting mechanism itself faces the challenge of various behavioral biases: the concentrated holding of governance tokens leads to "whale governance" (plutocratic governance), in which the interests of a few large holders may be systematically inconsistent with those of ordinary LPs (for example, market makers holding large amounts of governance tokens may vote against lowering the leverage cap, because high-leverage trading brings more fee income); chronically low voter turnout allows a few active participants to dominate decisions; and token holders may vote for short-term token-price gains rather than the system's long-term robustness. These behavioral factors mean that G3 may not always be superior in practice to a well-designed G2 mechanism.

Examining insurance-fund governance from a regulatory perspective is likewise important. The Principles for Financial Market Infrastructures, jointly issued by the Bank for International Settlements' Committee on Payments and Market Infrastructures and the International Organization of Securities Commissions, sets an international standard for the default management of central counterparty clearinghouses, requiring CCPs with more complex business risk or systemic importance across multiple jurisdictions to maintain prefunded resources sufficient to cover the "default of the two largest participants and their affiliates" scenario (the industry's "Cover 2" standard), while other CCPs must at least cover the default of the single largest participant ("Cover 1"; the EU's EMIR subsequently raised Cover 2 to a hard minimum for all CCPs within its jurisdiction). This standard provides a reference benchmark for assessing the adequacy of crypto derivatives insurance funds. Although a crypto-market insurance fund is not, in legal nature, a CCP guarantee fund, its functional role is closely analogous. At the regulatory-framework level, the EU's MiCA regulation (Regulation (EU) 2023/1114, in force June 29, 2023, with the core rules for crypto-asset service providers applying from December 30, 2024) imposes capital-adequacy and prudential-operation requirements on crypto-asset service providers, but its specific regulatory standards for derivatives insurance funds remain in the process of detailed rulemaking [31]. Platforms with lower governance transparency on the G0–G3 spectrum (G0–G1) may evade these requirements by registering in loosely regulated jurisdictions, creating a "regulatory arbitrage" in which a platform acquires users in strictly regulated markets while operating its core risk-management mechanism in a regulatory void. This arbitrage not only undermines the consistency of global financial stability but also makes cross-border regulation of insurance funds an urgent issue in crypto-finance regulation. Beyond the applicability of the regulatory framework, the legal characterization of the insurance fund itself carries unresolved uncertainty. Core disputes include: whether the insurance fund constitutes "client assets" and is thereby subject to stricter segregation and fiduciary obligations; the fund's legal position in the priority of claims in a platform's bankruptcy liquidation (as in the FTX case)—whether it is common property belonging to all traders or the platform's own asset; and how liquidation penalties are classified for tax purposes (insurance premium, fee, or fine), with different classifications carrying starkly different compliance obligations for the platform. The unresolved state of these legal

questions constitutes yet another dimension of uncertainty for the sustainability of the insurance-fund institution.

The progression from G0 to G3 is not merely an upgrade of technical architecture but a profound reconstruction of the trust mechanism of crypto derivatives markets. On this spectrum there is no absolutely perfect endpoint, only a process of continually seeking a better balance among transparency, efficiency, safety, and the ideal of decentralization. The governance of the insurance fund, like the weakest link in a chain, determines the ceiling of the entire platform's governance system. A platform that has achieved a high degree of transparency in both trade matching and asset custody will have the credibility of its whole system dragged down by this shortcoming if its insurance-fund management remains stuck at the G0 or G1 level. The political economy of insurance funds ultimately comes down to how, through institutional design, to build a governance consensus—in a market environment full of uncertainty—that can both withstand systemic collapse and defend against human greed.

12.5 An evaluation framework and cross-platform comparison

To comprehensively assess the merits of different derivatives platforms' insurance-fund mechanism designs, we must move beyond a single size metric and establish a multi-dimensional composite evaluation system. As argued in Section 12.1.3, simply comparing "whose insurance fund is larger" is not only superficial but potentially misleading: the platform with the largest absolute size does not necessarily have the greatest buffer thickness per unit of risk exposure. Through a systematic examination of four dimensions—capital adequacy, incentive compatibility, process quality and efficiency, and governance and verifiability—we can more clearly discern each platform's true defensive capacity in responding to extreme risk, and provide investors and protocol designers with a structured analytical tool.

12.5.1 Constructing the four-dimensional evaluation system

Capital adequacy is the primary basis for measuring an insurance fund's defensive capacity, and also the most intuitive but most easily misread dimension. This dimension involves three levels of examination. The basic metric is the ratio of the fund's size to the platform's total open interest; this relative metric reveals the thickness of the capital reserve per unit of risk exposure better than absolute size does. In traditional finance, central counterparty clearinghouses typically maintain strict margin and guarantee-fund ratios and conduct periodic stress tests to ensure solvency under extreme scenarios; in crypto markets, because asset volatility is higher, the market structure is more fragmented, and there is no unified regulatory standard, the reasonable range of this ratio is still being explored dynamically. Above this is stress-test coverage—whether the fund can cover the maximum total loss under historical extreme conditions (such as the "Black Thursday" of March 2020, the liquidation cascade of May 2021, and the "10/10 event" of October 2025) or under simulated future extreme scenarios. This requires combining the liquidation-reflexivity equation (and its divergence-criticality

condition) established in Chapter 11 with the tail-risk-distribution modeling of Section 12.3.5 of this chapter to make a quantitative estimate. Further, there is the reliability of the replenishment mechanism—whether, after the fund is substantially depleted in an extreme event, there is a clear, credible, and rapid recapitalization mechanism. An insurance fund that takes months to recover after a crisis will have its defensive capacity in successive extreme events greatly diminished.

Incentive compatibility examines whether a platform's mechanism can effectively align the interests of various participants with the system's long-term robustness. This dimension covers three key sub-metrics. The reasonableness of the liquidation-penalty structure is the first key point—whether the penalty effectively deters high-risk behavior without excessively punishing normal traders. Too low a penalty cannot constrain high-leverage speculation, while too high a penalty may cause traders to adopt a more aggressive "gamble it all" strategy as they approach the liquidation line, instead aggravating system risk. On this basis, the ability to suppress moral hazard is likewise important, particularly at the platform level—whether the platform aligns its own interests with the system's robustness through "interest-binding." If a platform controls the insurance fund but bears no risk of loss, then the fund may degenerate into a marketing tool or even a fraud tool. Finally, the degree of interest alignment is also a core consideration—whether the distribution of interests across funding sources, use rules, and value appreciation matches the actual risk-bearers. If liquidity providers bear high tail risk without receiving a corresponding risk premium, or if the platform lacks skin-in-the-game constraints when handling shortfall losses, such a mechanism will inevitably face the risk of collapse in extreme conditions.

Process quality and efficiency focuses on the liquidation mechanism's performance in actual operation, especially its response speed and execution accuracy in the face of extreme events such as network congestion, oracle latency, or malicious manipulation. This includes liquidation execution speed and slippage (whether risky positions can be disposed of quickly and effectively in extreme volatility without producing an excessive market impact); the shortfall rate and ADL frequency (as the ultimate outcome metrics of problems in the earlier layers of the liquidation chain, a higher shortfall rate means weaker upstream defenses of the liquidation mechanism); and system performance under extreme conditions (whether problems such as downtime, API latency, or unfillable orders occur). The market-wide liquidation cascade of October 10, 2025 (the "10/10 event") provides empirical data of significant analytical value for this dimension: in that event, several mainstream exchanges suffered varying degrees of system-performance degradation, and some platforms' liquidation engines had insufficient throughput to handle the instantaneous burst of forced-liquidation orders, causing shortfall losses far above normal levels [32].

Beyond the above four dimensions, an increasingly important but often overlooked supplementary dimension is the systemic risk contribution. As the crypto derivatives market grows in scale and cross-platform interconnection deepens, the failure of a single platform's insurance fund may produce spillover effects on the entire ecosystem through multiple transmission channels (including market makers' cross-platform exposure, systemic redemption pressure on stablecoins, and the cross-market spread of panic). Traditional finance has developed mature tools for measuring systemic risk:

among them, the CoVaR (conditional value at risk) proposed by Adrian and Brunnermeier [33] measures the increment in the entire system's value at risk when a single institution falls into distress, while the MES (marginal expected shortfall) proposed by Acharya et al. [34] measures a single institution's expected loss contribution during a systemic crisis. Transplanting these tools to the crypto derivatives domain makes it possible to assess the core question: "if platform X's insurance fund is broken through, how much additional risk will other platforms and the entire market face?" Directly transplanting these traditional tools, however, faces obstacles specific to crypto markets: the extreme insufficiency of transparency in cross-platform position data makes the conditional tail-dependence structure difficult to estimate; the crypto market's volatility-regime switches are more frequent, so conditional-quantile estimates based on historical data may lag severely; and the correlation among exchanges may be transmitted mainly through market makers' cross-platform inventory channel rather than through traditional balance-sheet linkages. Applying CoVaR and MES to the crypto derivatives domain therefore requires modeling adaptations tailored to the above distinctive transmission channels. Although the data availability and transparency of cross-platform positions remain major obstacles, the introduction of the systemic-risk-contribution dimension extends insurance-fund evaluation from "the robustness of a single platform" to "the resilience of the entire ecosystem," connecting with the logical framework of macroprudential regulation.

Governance and verifiability concern the transparency of the insurance fund and the certainty of rule enforcement. This requires not only that the fund balance be verifiable on-chain but also that the liquidation rules, the conditions for using funds, and the emergency-intervention mechanism have a high degree of transparency and predictability. Specifically, this dimension examines rule transparency (whether the inflow, depletion, and replenishment mechanisms of funds are public, clear, and codified), fund auditability (whether the balance and every income-and-expenditure flow can be independently verified in real time by a third party), and the governance mechanism for parameter changes (whether these are decided unilaterally by the platform or through a decentralized checks-and-balances process such as DAO voting). As the G0–G3 governance spectrum established in Section 12.4 of this chapter reveals, a platform's position on this spectrum—from a fully black-box operation (G0) to fully decentralized on-chain governance (G3)—directly determines whether it can maintain user trust at the moment of crisis. An "on-chain visible balance" does not equal "governance transparency"; a visible size with unverifiable use rules is still incomplete transparency. The insurance fund is usually the "weakest link" in a platform's governance chain, because it involves the largest scale of funds, the highest urgency of decisions, and yet transparency requirements that are often lowered in the name of "safety."

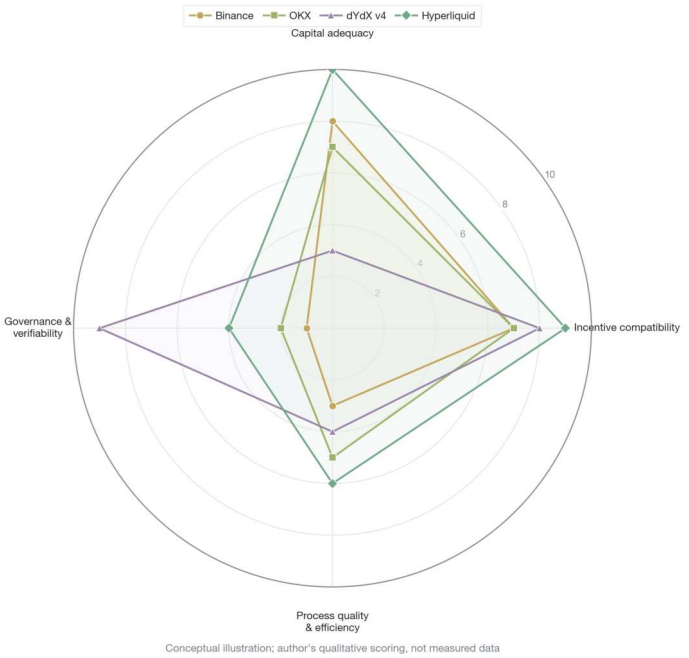


Figure 12-20. A four-dimensional radar chart of major derivatives platforms' insurance-fund mechanisms (conceptual illustration, qualitative scoring by the author, not measured data)

12.5.2 Platform comparison analysis

Figure 12-20 presents, in radar-chart form, the performance profile of each platform across the four dimensions: centralized platforms hold an advantage in capital adequacy but have a clear shortcoming in governance verifiability, while decentralized platforms exhibit the opposite pattern. Applying this four-dimensional evaluation framework to the current market's mainstream platforms, we can observe markedly different mechanism choices and capability emphases. No single platform holds an absolute advantage across all dimensions, and this "no-platform-optimal-on-all" pattern essentially reflects the deep trade-offs in insurance-fund design among capital efficiency, incentive compatibility, process reliability, and governance transparency.

Centralized exchanges, by virtue of their enormous trading volumes and long-term operational accumulation, possess insurance funds of enormous absolute size. According to public data from 2025–2026, Binance's futures insurance fund (an aggregation of segregated funds across contract groups, CoinGlass basis, as of 2025) has remained above $2.3 billion year-round and should be distinguished from its roughly $1 billion Secure Asset Fund for Users (SAFU) user-protection fund; it is the largest single futures insurance fund in the industry [35]. Because of its enormous open-interest base (total OI across all contracts of about $35 billion, fluctuating with market conditions), however, its fund-to-open-interest ratio typically holds around 6.6%. Binance's liquidation penalties adopt a tiered, escalating structure, setting different penalty ratios according to leverage and position size, which deters high-risk behavior to some extent. In terms of process quality, Binance's centralized matching engine performs relatively stably under normal market conditions. But in the 10/10 event, Binance too suffered API latency

and partial functional downtime, exposing its performance bottleneck under extreme load. In terms of governance and verifiability, Binance remains at the G0–G1 level: although some fund addresses have been made public, the rules for using the fund, the adjustment logic of liquidation parameters, and the emergency-intervention mechanism remain a black box to outside observers, and all parameter changes are decided unilaterally by the platform. OKX has not publicly disclosed the size of its futures liquidation insurance fund; the roughly $500 million it discloses (now expanded to over $700 million) is a Security Fund (mainly covering user losses from platform-side hacks and vulnerabilities, comparable to Binance's SAFU) and should not be used directly as a futures insurance fund or in a fund-to-open-interest ratio. OKX has made some improvements in transparency (disclosing more operational data), but its core governance mechanism remains controlled by the platform, and its overall evaluation is similar to Binance's.

Among decentralized derivatives protocols, dYdX v4, by deeply binding its insurance fund to the community treasury, has achieved a high degree of governance verifiability (G2–G3 level). Although its roughly $16 million insurance fund is relatively small in absolute size, with a fund-to-open-interest ratio of only about 2.0%, it is entirely controlled by DAO governance, and the rules for injecting and using funds are transparent and predictable [36]. The liquidation rules are fully codified, parameter changes require a DAO vote, and any community member can put forward a governance proposal and participate in the discussion. This highly transparent governance model puts dYdX v4 far ahead of its centralized competitors on the governance dimension. This community-treasury-based model of socialized risk-bearing, however, still has room for improvement in incentive compatibility: token holders passively bear the system's tail risk without a refined risk-pricing mechanism to compensate for it. In the 10/10 event, dYdX v4 experienced a chain halt, exposing its process-quality shortcoming under extreme load. Moreover, the low fund-to-open-interest ratio means its buffer space is limited in the face of a large-scale liquidation cascade, and the speed of its replenishment mechanism is also constrained by the DAO governance process.

Hyperliquid, for its part, has explored a new direction in market-based risk underwriting. Its HLP vault's total TVL was about $390 million in the second half of 2025 (peaking above $500 million on March 11, 2025, and falling back to about $270 million by mid-2026), and the fund-to-open-interest ratio calculated on this basis was as high as about 18.6% [37]. As noted in Section 12.1.3, however, using total HLP TVL as a whole as the insurance-fund size may overstate the actual risk-buffer thickness. Moreover, the tail risk of long-tail assets (such as low-liquidity meme coins like JELLY and POPCAT) is far greater than that of mainstream assets such as BTC/ETH, and the average ratio of 18.6% masks a coverage that may be severely inadequate for specific asset classes—precisely the "risk-pool commingling" problem revealed by the case studies in Section 12.6. More importantly, by opening its liquidation vault and allowing liquidity providers to participate in risk underwriting autonomously on market-based principles, Hyperliquid makes HLP's annualized return directly reflect the market's dynamic pricing of liquidation risk, creating a marked differentiation from other platforms in incentive compatibility. Its liquidation rules are on-chain verifiable, and the validator consensus mechanism provides a degree of decentralized checks and balances

on parameter governance, so its overall governance level sits at G2. As we see in the case in Section 12.6, however, this complex market-based mechanism still faces severe tests of process quality and efficiency in the face of extreme manipulation of long-tail assets: in the 10/10 event Hyperliquid triggered its ADL mechanism for the first time, and the 2025 JELLY incident further exposed the structural flaw of risk-pool commingling. In addition, although the validator consensus mechanism provides emergency-intervention capability, the degree of centralization and transparency of its decision process still needs improvement.

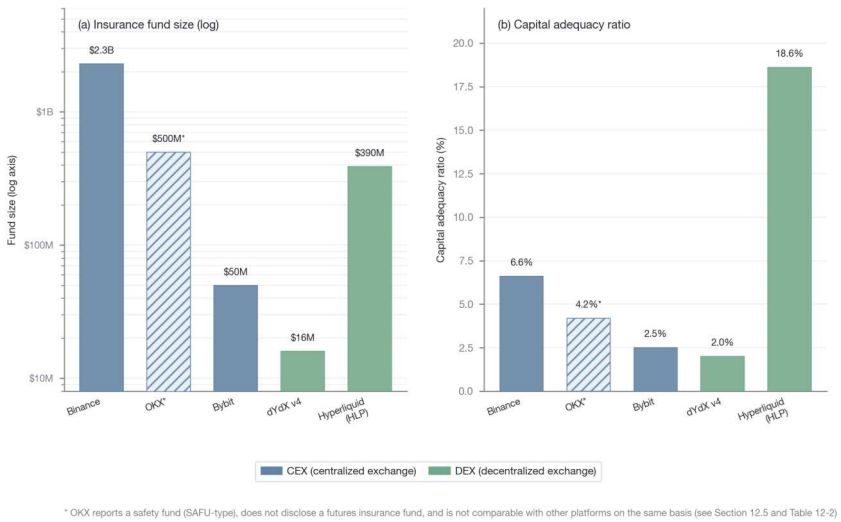


Figure 12-21. Comparison of major derivatives platforms' insurance-fund size and capital adequacy ratio (the capital adequacy ratio is the author's estimate; platforms do not publish this ratio; fund sizes are an as-of-2025 snapshot, and present values have since changed and are highly time-sensitive)

Figure 12-21 translates the above analysis into a visual comparison, intuitively confirming the earlier argument that "absolute size can be misleading" (see Section 12.1.3). Table 12-2 summarizes the systematic comparison of the four major platforms under the four-dimensional evaluation framework, providing readers with a structured frame of reference.

Evaluation dimension	Key metric	Binance	OKX	dYdX v4	Hyperliquid
Capital adequacy	Fund size	About $2.3 billion (futures insurance fund, distinct from the roughly $1 billion SAFU)	Not disclosed‡	About $16 million	HLP total TVL about $390 million (market-making + liquidation)†
	Fund-to-OI ratio	About 6.6%	Not comparable‡	About 2.0%	About 18.6% (upper-bound estimate based on total HLP TVL)†
	Replenishment mechanism	Opaque	Opaque	DAO governance	Market-based replenishment
Incentive compatibility	Penalty structure	Tiered, escalating	Tiered, escalating	On-chain rules	Uniform low fee
	Interest alignment	Low	Low	Medium	High
Process quality	10/10 event performance	Downtime and API latency	Partial latency	Chain halt	ADL triggered
	ADL frequency	Extremely rare	Extremely rare	Chain halt as substitute	First triggered in 10/10
Governance verifiability	Fund transparency	Black box (G0–G1)	Partially transparent (G1)	Fully on-chain (G2–G3)	On-chain verifiable (G2)
	Parameter governance	Platform decides unilaterally	Platform decides unilaterally	DAO vote	Validator consensus

Table 12-2. Four-dimensional evaluation summary of the four major platforms' insurance-fund mechanisms (Data source: compiled by the author from each platform's public documentation)

> Note: † HLP total TVL encompasses a shared capital pool for both the market-making and liquidation strategies, and using it as a whole as the insurance-fund size may overstate the actual risk-buffer thickness; 18.6% should be understood as an upper-bound estimate based on total HLP TVL (see Section 12.1.3). ‡ OKX has not publicly disclosed the size of its futures liquidation insurance fund; the roughly $500 million it discloses is a security fund (comparable to SAFU, covering platform hacks), not comparable on a like-for-like basis with other platforms' futures insurance funds, so the "fund size/ratio" columns are recorded as "not disclosed/not comparable."

Several key insights can be distilled from this cross-platform comparison. Capital adequacy and governance transparency currently exhibit an observable descriptive association in the market: the platforms with the largest capital tend to have the least transparent governance, while the platform with the most transparent governance has the smallest capital. This association, however, may mainly reflect historical path dependence rather than a causal institutional trade-off: centralized platforms were founded earlier and have operated longer, naturally operating in an opaque manner and accumulating a size advantage in the early industry environment that lacked DeFi

infrastructure; decentralized platforms are newer and have not yet experienced enough bull-and-bear cycles to accumulate a large-scale fund. Moreover, opaque and underfunded platforms (such as FTX) have already collapsed and are absent from the current sample, so survivorship bias may also distort the observed association. Whether, as decentralized platforms mature, a platform combining "high transparency plus high capital adequacy" will emerge and break the current association pattern is a proposition worth tracking over the long term. On this basis, a trade-off also exists between incentive compatibility and process reliability: Hyperliquid's market-based mechanism is more refined in incentive design than other platforms, but its complexity also brings a larger attack surface and more edge cases. On balance, no platform has reached an ideal level on all dimensions, and the choice depends on a user's prioritization of the different dimensions: a large institution pursuing capital safety may prefer Binance's absolute size, a DeFi-native user pursuing transparent governance may prefer dYdX v4, and a professional trader pursuing capital efficiency and innovative mechanisms may prefer Hyperliquid.

12.6 Case study: the JELLY incident

The merits of a theoretical framework must ultimately be tested by real market extreme stress. The series of risk events on the Hyperliquid platform in 2025—particularly the JELLY token manipulation case in March—offers us an extremely rare and research-worthy empirical case study. These events not only exposed the hidden fragility of a complex decentralized liquidation architecture but also laid bare the difficult choices decentralized governance faces at the moment of crisis, and further revealed the inherent limitations of a market-based risk-underwriting mechanism in the face of deliberate malicious gaming. The JELLY incident is worth studying in depth not only because of the sophistication of its attack method and the severity of its consequences, but also because it concentrates the technical design, governance mechanism, and political-economy tensions of the insurance fund in a single event, making it the fullest illustration of this chapter's theoretical arguments.

12.6.1 The HLP architecture

To understand the essence of the JELLY incident, we must first dissect the architecture of Hyperliquid Vaults. Hyperliquid's core innovation lies in its HLP vault system—a community-owned protocol vault that executes multiple automated strategies. After a user deposits USDC to become an LP, the capital is allocated across multiple sub-strategies, sharing the overall profit and loss of the strategy portfolio. Among the most critical sub-strategies are the market-making strategy (continuously providing two-sided quotes on the order book to earn the bid-ask spread) and the liquidation vault. The insurance-fund function is executed precisely as a core strategy embedded within HLP—namely, the liquidation vault. This design of integrating the market-making function and the insurance function within the same vault is the key feature distinguishing Hyperliquid from other protocols and the source of its high capital efficiency: the same pool of capital can both earn market-making returns and serve as a liquidation buffer when needed.

The liquidation vault was designed to solve the problem of liquidity evaporation in a decentralized order book under extreme conditions. Its risk-underwriting process follows a strict hierarchical logic: when a leveraged position's equity falls below the maintenance-margin requirement, the system first attempts to close it via a market order on the public order book—an open, competitive process in which any market participant can act as counterparty to absorb the liquidation order. If order-book depth is insufficient, or excessive price slippage drives the position's equity below two-thirds of the maintenance margin, the position is transferred directly to the liquidation vault. After the liquidation vault takes over all of the shortfall account's positions and remaining margin, it uses its enormous capital reserve to absorb the instantaneous shock, then gradually closes out the position in the market via an algorithm to minimize the damage to the market price. The profit or loss is shared by all HLP LPs [38]. The operating model of the HLP liquidation vault differs fundamentally from a centralized exchange's insurance fund. A CEX insurance fund typically uses a "deficit-covering" model: after the liquidation engine executes the closeout at market, the fund makes up the shortfall, and the fund itself bears no directional risk. The HLP liquidation vault, by contrast, uses a "position-takeover" model: it takes over the shortfall account's positions and remaining margin in full, bearing the market risk of holding directional exposure until it can gradually close out in the market. This design difference exposes HLP to an additional risk dimension that a CEX insurance fund does not face—"position-holding risk": in extreme conditions, the taken-over position may deteriorate further during the closeout process, and the JELLY incident is precisely the extreme manifestation of this risk dimension.

The design's strength lies in converting an instantaneous liquidation shock into a gradual release of liquidity, enhancing the system's antifragility. As a final line of defense, Hyperliquid also designed an auto-deleveraging mechanism: when the potential loss facing the liquidation vault exceeds a certain threshold, the system force-closes profitable users' positions to make up the liquidation loss, ensuring the protocol's overall solvency. Above ADL, there is also an ultimate discretion layer—the validator consensus mechanism—in which validators can decide by vote whether to carry out a special intervention. In theoretical reasoning, this four-tier defense system, composed of "order-book market closeout → liquidation-vault absorption → ADL ultimate risk-bearing → validator consensus," seems to construct a complete line of defense.

HLP's three major mechanism advantages earned it a high composite score in the evaluation framework of Section 12.5. First, market-based pricing of risk: HLP's annualized return directly serves as a dynamic risk-price signal—when the system faces higher liquidation risk, the higher return expectation attracts more risk-underwriting capital, and conversely it squeezes out excess capital, achieving a dynamic balance between the insurance fund's size and the actual risk exposure. Second, preliminary risk tiering and segregation: allowing users to create custom vaults focused on specific assets or strategies lays a foundation for more refined risk management in the future. Third, enhanced incentive compatibility: liquidation profits belong to the LPs, incentivizing the community to keep optimizing the liquidation strategy and turning risk management from "the platform's burden" into "the community's opportunity."

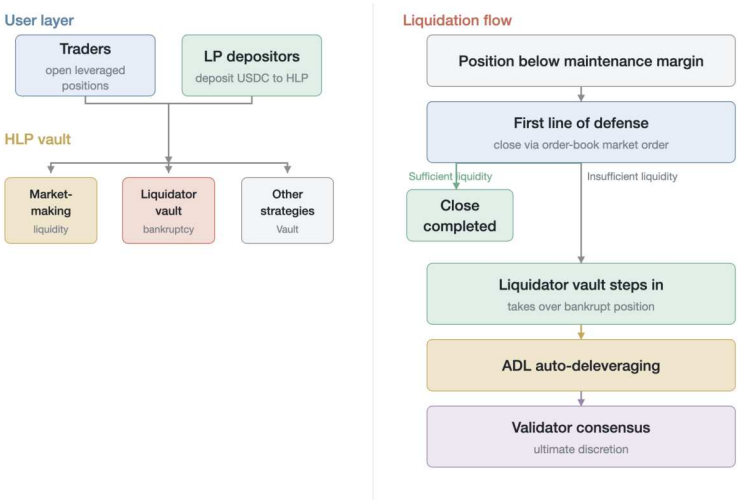


Figure 12-22. The risk-underwriting architecture and liquidation flow of Hyperliquid Vaults

Figure 12-22 presents, with arrows and threshold annotations, the direction of flow and the trigger conditions of this four-tier risk-underwriting architecture. The JELLY incident, however, proved that a significant gap exists between theory and practice.

12.6.2 Timeline of the incident

The JELLY incident of March 26, 2025, however, exposed a structural flaw in this defense system. JELLY was a meme coin issued via Pump.fun on the Solana chain, with a market capitalization of only about $10 million at the time and extremely scarce on-chain liquidity. The attacker precisely exploited a hole in Hyperliquid's risk controls for long-tail assets, carrying out a planned combined attack of cross-chain price manipulation and a liquidation-mechanism loophole [22]. The sophistication of this attack lay in the fact that it did not seek to profit from a directional bet on price but directly targeted the structural flaw of the liquidation mechanism itself.

In the setup phase of the attack, the attacker created three brand-new addresses on-chain and prepared several million dollars in capital. On Hyperliquid, targeting the low-liquidity JELLY, the attacker simultaneously opened a short position with a total value of about $4.5 million (concentrated in one address) and two long positions totaling about $5 million (spread across the other two addresses). This long-short hedged setup left the attacker's net risk exposure to JELLY price movements extremely small. The JELLY incident exposed not only a flaw in the liquidation architecture but, more fundamentally, an absence of basic operational risk controls. In traditional derivatives markets, the pattern of three newly created addresses simultaneously opening opposing enormous positions on the same low-liquidity instrument within a short time would immediately trigger a related-trading surveillance alert; a two-way position of $4.5 million/$5 million against JELLY's market capitalization of only about $10 million means that open interest approached the instrument's entire market capitalization—

an obvious manipulation warning signal; and the pattern of actively and gradually withdrawing margin is itself a clear signal of self-liquidation. The above basic anomaly-detection capabilities are standard equipment on traditional derivatives exchanges, and their absence on Hyperliquid reflects the systematic lag of decentralized platforms in building operational risk-control infrastructure. The real source of the attack's profit was not a directional bet on price but arbitrage against the liquidation mechanism itself—that is, by "injecting" a doomed-to-lose short position into the liquidation vault, then manipulating the price in the external market to make the liquidation vault bear an enormous loss, and finally realizing a net profit through the profit on the long positions.

Next came the liquidation-trigger phase. The attacker began to gradually withdraw margin from the address holding the short position, actively pushing that position toward the liquidation line. The key to this operation was "gradual withdrawal" rather than "one-time withdrawal": by precisely controlling the speed of margin withdrawal, the attacker ensured that the timing of the liquidation was precisely synchronized with the rhythm of the cross-chain manipulation. Because JELLY's depth on the Hyperliquid order book was extremely scarce (for a meme coin with a market capitalization of only $10 million, there was almost no counterparty on the order book capable of absorbing a liquidation order of several million dollars), the liquidation engine could not find sufficient liquidity in the market to complete the closeout. Following the system's design logic, this enormous short position was duly transferred to HLP's liquidation vault. At this point the attacker had successfully "injected" an enormous short risk exposure into HLP, and this was the most critical step in the entire attack chain.

At almost the same moment the short position was taken over by the liquidation vault, the attack entered the cross-chain price-manipulation phase. The attacker deployed enormous capital on a decentralized exchange on the Solana chain (mainly Raydium) to buy up JELLY spot frantically. Because JELLY's on-chain liquidity was extremely thin, this concentrated buying drove the token's price to surge by about 400% in an extremely short time [21]. This violent price swing was rapidly transmitted to the Hyperliquid platform via the oracle. The oracle was designed to ensure consistency between the on-chain price and the off-chain market price, but in this scenario the oracle faithfully relayed an artificially manipulated price signal, instead becoming an indispensable transmission tool in the attack chain. This link reveals a systemic risk in the cross-chain DeFi ecosystem: when a protocol's risk management depends on price signals from an external market, the manipulability of that external market becomes the protocol's security boundary.

As JELLY's price surged, HLP faced severe solvency pressure. The enormous short position the liquidation vault was forced to hold instantly produced about $12 million to $13.5 million in unrealized losses (OAK Research/Halborn put it at about $12 million; CoinDesk and perp.wiki put it at about $13.5 million, equivalent to a temporary drawdown of about 27% of the HLP vault). This figure was nearly enough to exhaust the entire HLP vault's liquidity buffer. At this point, the system should have triggered the ADL mechanism to resolve the crisis—that is, to make up the liquidation loss by force-closing profitable users' positions. A serious architectural flaw, however, caused the ADL mechanism to fail completely.

ADL's trigger logic is based on the ratio of the liquidation vault's unrealized loss to its corresponding capital pool exceeding a preset threshold. In the system design at the time, however, the liquidation vault's capital pool was not fully segregated from the HLP main pool; the liquidation vault was merely a sub-strategy within HLP, not an independent capital entity. This meant that the denominator of the ADL trigger ratio was set to the entire HLP vault's enormous total assets (hundreds of millions of dollars), rather than the liquidation vault's own standalone capital. Against HLP's capital pool of hundreds of millions of dollars, the $12.3 million loss did not reach the ADL trigger threshold. The after-the-fact fix, which segregated the liquidation vault independently, greatly shrank the denominator of the ADL trigger ratio, markedly improving the safety valve's sensitivity. But the specific trigger-threshold value has not been publicly disclosed to the community; Hyperliquid's documentation describes only the general condition that ADL activates when "account value or the value of an isolated position becomes negative," without giving the precise ratio parameter—itself a transparency issue that needs continual improvement in G2-level governance. The system fell into a contradictory deadlock: an enormous shortfall loss was genuinely occurring, yet the final safety valve remained silent because "the pool was too big." This is precisely the fatal manifestation, under extreme conditions, of the "risk-pool commingling" problem analyzed in this section: the capital of the market-making strategy and the capital of the liquidation strategy were mixed in the same pool, weakening the risk signal and lowering the sensitivity of the safety mechanism.

Faced with the imminent systemic solvency crisis, the Hyperliquid team activated the final tier of the four-tier defense system: the validator consensus mechanism. Within just a few minutes of the crisis erupting, the platform's validators unanimously passed an emergency resolution: to disregard the real market price relayed by the oracle and forcibly overwrite JELLY's settlement price to $0.0095, the level before the manipulation occurred. Based on this artificially set price, the system forcibly closed all related positions, instantly eliminating HLP's book loss of over ten million dollars. The essence of this operation was that the validators collectively exercised a "price discretion," replacing the market price and oracle data with human judgment to preserve the system's overall solvency. In the cleanup phase after the crisis was resolved, Hyperliquid gave a full refund to all innocent JELLY long holders at the fair market price before the event, ensuring that ordinary users did not suffer losses because of the platform's mechanism flaw. The attacker's accounts were frozen, with a final net loss of about $910,000 (Arkham basis, an upper-bound estimate under the scenario that the frozen funds cannot be recovered) [21]. Specifically, the attacker's profit-and-loss structure was roughly as follows: the short position lost all of its margin because of the actively triggered liquidation; the two long positions produced a considerable book profit after the price surged, but the validators' price rollback forcibly restored the settlement price to the pre-attack level, wiping out most of the longs' profit; and the buying slippage on Raydium during the cross-chain price manipulation constituted an additional execution cost. The $910,000 net loss was the final result after netting the above items. A counterfactual analysis reveals the economic impact of the intervention: had the validators not intervened, the attacker's net gain—the long profit minus the short loss and manipulation cost—could have been as high as several million dollars, while HLP would

have suffered an actual loss of about $12.3 million. From a purely economic standpoint, the attacker's attack did not succeed in turning a profit; but from the standpoint of system security, the architectural flaw the attack exposed was far more serious than the $910,000 direct loss.

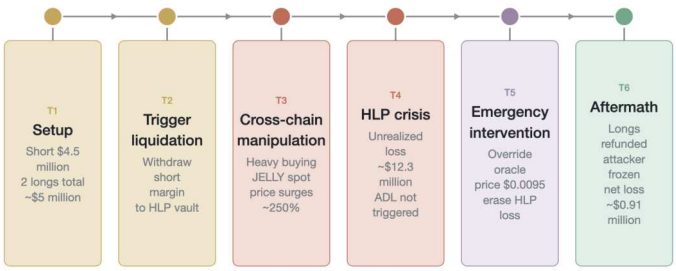


Figure 12-23. Timeline of the JELLY incident's cross-chain manipulation and crisis evolution (the "price surge of about 250%" in the figure is the gain in the initial manipulation window; the cumulative gain to the validator-consensus overwrite point was about 429%, and the "about 400%" in the text lies between the two, reflecting the difference in scope across different time windows of the same event)

Figure 12-23 integrates the time points of each phase, the key operations, and the flow of funds into a complete timeline. The entire attack, from setup to resolution, lasted less than a few hours, showing how fast risk evolves in a cross-chain DeFi environment.

The JELLY incident was not an isolated event but the climax of a series of risk events Hyperliquid experienced in 2025. Earlier, on March 12, 2025, Hyperliquid had already experienced a large whale-manipulation event: a trader used 50x leverage to open a long ETH position of more than $300 million and then triggered liquidation by strategically withdrawing margin, causing HLP to suffer a loss of about $4 million [39]. Although the loss was relatively limited, this event had already exposed HLP's fragility in the face of large directional manipulation and sparked wide community discussion of HLP's safety. And in November 2025, after the JELLY incident, the POPCAT incident further exposed a similar fragility, though its attack vector differed from JELLY's: the attacker withdrew about $3 million in USDC from OKX, split it across 19 wallets, built a POPCAT long with high leverage, placed a fake buy wall of about $20 million on Hyperliquid's own order book to fabricate demand, then abruptly canceled the orders, triggering a liquidation cascade of its own longs and forcing HLP to absorb about $4.9 million in bad debt; the attacker went ahead knowing full well that they would lose all of their roughly $3 million in margin. POPCAT was an on-platform order-book spoofing combined with liquidation-mechanism arbitrage, and did not involve JELLY-style cross-chain spot/oracle manipulation; the two share the common feature of targeting thinly liquid meme coins and being willing to sacrifice their own principal to spill losses onto HLP, and on-chain clustering suggests the two likely came from the same attacker cluster. This "harm yourself to harm others" strategy shows that for premeditated, cost-insensitive directional manipulation, a market-based risk-underwriting mechanism still faces severe challenges: when an attacker's goal is not to profit but to destroy, a defense mechanism based on the assumption of economic rationality will face fundamental

failure.

The POPCAT incident is especially worth analyzing in depth because it occurred after Hyperliquid had implemented a series of fixes in response to the JELLY incident. The three core improvements introduced after the JELLY incident (asset segregation of the liquidation vault from the HLP main pool, dynamic position caps based on market capitalization and order-book depth, and a validator-vote delisting mechanism) all failed to effectively defend against the attack in the POPCAT incident. Asset segregation did allow the ADL mechanism to trigger at an appropriate threshold, avoiding the "pool too big so the safety valve stays silent" problem of the JELLY incident; the dynamic position cap, however, was circumvented by the attacker through dispersing addresses and building positions gradually, and the validators' delisting vote failed to start in time before the loss occurred because of the ambiguity of judging "whether an attack has already happened." This outcome reveals a fundamental limitation of institutional evolution: each round of fixes is essentially a targeted patch against the "last attack vector," while an adaptive attacker actively seeks new attack surfaces. This "attack → patch → new attack" evolutionary dynamic is highly isomorphic to the Red Queen effect in biology: the defender must keep evolving merely to maintain the same relative position against the attacker. From the standpoint of economic theory, this evolutionary dynamic can be understood more precisely as a manifestation of Goodhart's law: when the defense rule (the "measure") is transparently made public, the attacker (the "agent") will adjust its strategy specifically to render that rule ineffective. The logic of the Lucas critique applies equally: a defense strategy formulated on the basis of historical attack patterns will change future attack patterns because of the attacker's expectations about the strategy itself. For a decentralized protocol, this means that risk management cannot rely solely on after-the-fact patching of vulnerabilities but must establish a systematic adversarial mindset: incorporating "how an attacker would exploit this rule" as a core consideration at the mechanism-design stage, and reserving sufficient institutional elasticity for unknown attack vectors.

Event	Whale incident March 12, 2025	JELLY incident March 26, 2025	POPCAT incident November 12, 2025
Trigger asset	ETH	JELLY (memecoin)	POPCAT (memecoin)
HLP loss	~$4 million	peak ~$12.3 million final ~$0 (intervention)	~$4.9 million
Key magnitude	LP withdrawals ~$130 million	final actual loss ~$0	attacker input ~$3 million
Confidence/cost	confidence leverage ~30×	attack cost ~$0.91 million	attacker strategy self-harm to hurt others
Intervention	none (market absorbs)	validators override oracle	manual close + pause bridge
Aftermath	LP confidence shaken	decentralization dispute	mechanism still fragile

Figure 12-24. A comparison of Hyperliquid's three major risk events in 2025

Figure 12-24 provides a structured comparison of the three events across four dimensions: time, attack method, loss scale, and platform response. From it one can observe three progressive trends: the complexity of the attacks escalated from single-platform manipulation (the March 12 whale incident) to more complex attacks on thinly liquid meme coins, with JELLY embodying cross-chain spot/oracle manipulation and POPCAT embodying on-platform order-book spoofing combined with liquidation-mechanism arbitrage; the attackers' risk appetite shifted from arbitrage-for-profit to a cost-insensitive destructive strategy (the POPCAT attacker was willing to sacrifice about $3 million in principal); and the platform's response also evolved from passively absorbing the loss (the whale incident) to actively intervening in the price (the JELLY incident) and after-the-fact architectural reform. This evolutionary path shows that a market-based risk-underwriting mechanism must iterate its defensive capabilities in lockstep when facing continually evolving adversarial attacks.

12.6.3 The boundaries of decentralized governance

The JELLY incident is a textbook "incomplete contracts" case, proving that no rule system designed in advance can exhaust all malicious attack vectors. No matter how sophisticated the code, there always exist edge cases the rule-makers did not foresee. When such an edge case occurs, the system ultimately relies on a "guarantor of last resort" holding discretion. In traditional finance, this role is played by the central bank or the regulator; in Hyperliquid's case, this role is borne collectively by the validators. This fact pushes decentralized derivatives platforms into a fundamental dilemma, which can be stated precisely as follows: should a protocol whose core narrative is "decentralization" and "code is law" possess an emergency centralized "circuit breaker" when facing an attack capable of destroying it? If the answer is yes, then who should hold this circuit breaker? What are the trigger conditions? And how can its abuse be prevented? If the answer is no, then should the protocol accept the possibility of being destroyed by a deliberate attacker in order to preserve the purity of its decentralization?

Hyperliquid's answer is "validator consensus"—a middle path between fully centralized decision-making and fully ungoverned operation. Validators, as stakeholders in the protocol (typically required to stake a large amount of tokens), have an incentive to maintain the protocol's long-term health. By requiring a unanimous or supermajority vote to pass, this mechanism in theory both preserves the capacity for emergency intervention and reduces the risk of single-point abuse through collective decision-making. It must be squarely acknowledged, however, that when the JELLY incident occurred (March 2025, before the permissionless validator set had gone live, when the network still ran a small foundation-bootstrapped validator set), the Hyper Foundation controlled about 81% of the staked HYPE tokens, and the participation of external independent validators was extremely low. It was precisely this highly concentrated validator structure that allowed the emergency resolution to reach "unanimous passage" with zero opposition in about 2 minutes—a speed that in substance is closer to "a unilateral decision by the core team" than to genuine decentralized consensus. This reality means that Hyperliquid's actual position on the governance spectrum may be closer to G1–G2 than to G2. After the incident, the validator set was expanded to 16 nodes, but the foundation still operates five of them and controls the majority of staked tokens, and

the node software remains closed-source. As the validator set decentralizes further, the credibility of this governance mechanism will rise accordingly, but the governance practice at the current stage should be evaluated as it truly is. In the JELLY incident, the validators' decision process was still highly centralized: the decision was made within a few minutes, the community had no time to participate in discussion, and the basis and process of the decision were not adequately recorded or made public. This "act first, explain later" model, though it saved the protocol at the moment of crisis, also posed a severe challenge to the platform's "decentralization" narrative. Critics point out that if validators can overwrite the oracle price and force liquidation within a few minutes, then what is the essential difference between this and the backroom operations of a centralized exchange?

This incident sparked a fierce debate in the community. Supporters argue that in the face of a deliberate malicious attack, protecting the interests of the protocol and users is the highest priority, and the validators' intervention was necessary and justified—just as central banks in traditional finance hold "lender of last resort" discretion during a systemic crisis. Opponents argue that this intervention undermined the fundamental principle of "code is law" and planted the seeds for future selective enforcement and conflicts of interest: if validators can overwrite prices in some situations, then who guarantees they will not abuse this power in other situations for their own benefit?

From a more macro perspective, the JELLY incident reveals an institutional reality of the current stage of DeFi development: absolute decentralization is difficult to achieve fully under current technical conditions. Almost all successful financial systems, whether traditional or crypto, maintain a fragile but necessary balance between rule rigidity and ultimate discretion. Traditional-finance CCPs have "default management committees" to handle extreme situations the rules cannot cover; central banks around the world hold "lender of last resort" discretion to respond to systemic crises; and even Ethereum itself once used a hard fork to "roll back" stolen funds after the 2016 DAO attack. Hyperliquid's validator consensus mechanism is, in essence, a preliminary but necessary practice of this age-old wisdom in a decentralized context. From an operational-security standpoint, the emergency-intervention mechanism itself also constitutes an attack surface: if an attacker could gain control of enough validators' keys or voting power through social engineering or technical means, they could exploit the same emergency-intervention mechanism to carry out a more destructive attack. Moreover, the process of validators completing a unanimous vote "within a few minutes" implies an efficient but possibly insecure communication channel (such as a Telegram group), and whether such channels themselves become a security hazard is worth examining. The design of an emergency-intervention mechanism must solve not only the governance problem of "who has the authority to intervene" but also the operational-security problem of "how the intervention process itself is protected."

After the incident, Hyperliquid quickly implemented a series of systematic fixes, seeking to plug the exposed loopholes without abandoning the market-based risk-underwriting model. It strictly segregated the liquidation vault's assets from the HLP main pool and set an independent ADL trigger threshold for the liquidation vault, thoroughly fixing the mechanism failure caused by risk commingling: this means that even if the liquidation vault's loss is negligible relative to HLP's total assets, ADL will be triggered

as soon as it exceeds the threshold relative to the liquidation vault's own capital. On this basis, it introduced a dynamic position-cap formula based on real market capitalization and order-book depth, curbing the excessive leveraging of long-tail assets at the source: if a token's market capitalization is only $10 million, the system will no longer allow the accumulation of open interest of several million dollars on that instrument, so that an attacker cannot build an enormous position on a low-liquidity asset large enough to threaten the system. Third, it granted validators the power to delist abnormal assets through an on-chain vote, providing an institutionalized tool for responding to similar long-tail-asset manipulation in the future [40].

The POPCAT incident of November 2025, however, showed that although these fixes plugged the specific loopholes exposed in the JELLY incident, they did not fundamentally resolve the fragility of a market-based risk-underwriting mechanism in the face of deliberate manipulation. When an attacker is willing to bear a net loss of several million dollars to create larger-scale systemic destruction, any defense mechanism based on the assumption of economic rationality will face a challenge. This recognition leads us to a deeper proposition: the design of an insurance-fund mechanism is not only a problem of technical optimization but also a political-economy problem concerning risk, power, and trust. As the chapter title reveals, the insurance fund is the node where "technical design" and "political governance" intersect most deeply in the perpetual-futures system; its technical dimension can be optimized with actuarial models, but its governance dimension involves questions of power, trust, and the distribution of interests—questions that cannot be solved by code, but only constrained by institutions.

12.7 Chapter summary

With this, we have completed a systematic analysis of the institutional arrangement that is the insurance fund. As the concluding piece of Part Four, "Design: The Core Mechanisms of Perpetual Futures," this chapter extends the risk management of perpetual futures from the purely technical level to the institutional level of interest gaming, power distribution, and trust-building. The insurance fund and auto-deleveraging together constitute a dual socialization mechanism for responding to extreme risk. Through ex ante pooling, the insurance fund absorbs within the system the negative externalities generated by high-leverage trading, and its core output is not investment return but the irreplaceable settlement credibility and rule predictability—the concrete embodiment, in the crypto derivatives domain, of the core function of financial market infrastructure emphasized in the Bank for International Settlements' Principles for Financial Market Infrastructures. And when this mechanism's funds are exhausted, ADL, as the ultimate means of ex post allocation, is forced to activate, always facing an irreconcilable trilemma among solvency, long-term revenue, and user fairness. The roughly $650 million in excess haircut in the 10/10 event, and the near-zero profit-to-loss symmetry ratio, reveal through empirical data the fundamental inadequacy of static ADL strategies in the face of complex market environments.

The confidence-as-liquidity hypothesis proposed in this chapter reveals the deeper economic significance of the insurance fund. The fund's signaling function is far more important than its mere capacity to absorb capital. The empirical evidence from the

Hyperliquid whale incident—"a $4 million actual loss triggering $130 million in panic withdrawals"—displays a correlational amplification effect of about 32.5 times, far exceeding the amplification of any capital leverage (although this factor may be affected by contemporaneous macro factors, social-media information cascades, and other confounders). The nonlinear "cliff effect" between confidence and the adequacy ratio shows that managing confidence is, in essence, the highest-order form of risk management.

The actuarial adequacy-ratio model transforms the vague qualitative question of "whether the fund is large enough" into a rigorous quantitative analysis based on the tail-risk distribution. The procyclical paradox that the adequacy ratio exhibits, however (accumulating when it is least needed and drying up when it is most needed), directly reveals the structural defect of relying on liquidation penalties as a single revenue source. The maturity mismatch between "the gradual accumulation of revenue" and "the sudden eruption of expenditure" is a fundamental challenge that insurance-fund design must squarely confront.

In addition, the existence of an insurance fund inevitably breeds a dual moral hazard: traders tend to take on greater risk because of implicit risk-bearing, while platform operators face the temptation to misappropriate funds or abuse discretion. The extreme case of FTX fabricating its insurance fund warns us that "being insured" by no means equals "being safer"; insurance itself may create new systemic risk. We therefore introduced the G0-to-G3 governance spectrum, pointing out that the governance of the insurance fund is often the weakest link in a platform's entire verifiability ladder. Visible size does not equal transparent rules; only when the process is open to scrutiny can trust be established.

Based on the above insights, we distill five design principles for building a robust mechanism: risk must be precisely priced to eliminate cross-subsidies; the platform's interests must be deeply bound to the system's robustness to preclude moral hazard; the operating process must have an extremely high degree of verifiability to establish trust; complex risks must be finely tiered and segregated to prevent contagion; and, at the same time, the "guarantor of last resort" holding ultimate discretion must be acknowledged and severely constrained by institutions. Figure 12-25 integrates the causal chain of the three chapters of Part Four into a single complete arc—from the routine regulation of the funding rate, to the eruption of a liquidation cascade, to the absorption by the insurance fund and the final allocation by ADL—intuitively presenting the hierarchical progression among these mechanisms.



Figure 12-25. The complete causal arc of Part Four

With the end of this chapter, the complete arc of Part Four is now closed. We have clearly seen a tight causal chain: from the routine regulatory function of the funding rate in Chapter 10, to the systemic imbalance triggered when the rate mechanism fails under extreme conditions; from the continuous depletion of margin, to the large-scale chain liquidation and cascade reflexivity triggered by the liquidation engine in Chapter 11; and when this series of shocks finally breaks through individual defenses and produces a shortfall gap, this chapter's insurance fund steps in as a risk-absorption layer to fill the loss; and once the fund is exhausted, ADL, as the ultimate risk-allocation mechanism, forcibly apportions the loss to profitable traders. These three chapters are not a parallel enumeration of three independent mechanisms but a complete life-cycle narrative running from normal operation to extreme collapse to crisis cleanup. The design quality of this set of core mechanisms directly determines the safety, capital efficiency, and basic fairness of perpetual futures as a financial product, and it lays a systematic mechanistic foundation for our subsequent discussion of the market's price discovery, arbitrage behavior, liquidity evolution, and volatility characteristics.

Having understood how the various links of the market mechanism operate in coordination, the next part will formally enter the domain of price discovery, exploring how a bewildering welter of information is incorporated into prices, how external price benchmarks invisibly shape the form of the market, and just what level of informational efficiency the market attains in this friction-filled crypto world.

References

[1] Marshall, M. (2025). How $3.21B vanished in 60 seconds: October 2025 crypto crash explained through 7 charts. *Amberdata Research Blog.* https://blog.amberdata.io/how-3.21b-vanished-in-60-seconds-october-2025-crypto-crash-explained-through-7-charts (accessed June 2026)

[2] FTI Consulting. (2025, December). Crypto crash Oct 2025: Leverage meets liquidity. https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity

[3] Decrypt. (2020, April). How Black Thursday reshaped the Bitcoin futures market. https://decrypt.co/26299/black-thursday-reshaped-bitcoin-futures-market-report

[4] Ledger Prime Research. (2021, August). Crypto's Black Thursday: The market crash of March 2020. https://www.research.ledgerprime.com/p/cryptos-black-thursday-the-market

[5] WhiteRabbit. (2020, March). Black Thursday for MakerDAO: $8.32 million was liquidated for 0 DAI. *Medium.* https://medium.com/@whiterabbit_hq/black-thursday-for-makerdao-8-32-million-was-liquidated-for-0-dai-36b83cac56b6

[6] Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360. https://doi.org/10.1016/0304-405x(76)90026-x

[7] Bank for International Settlements. (2012, April). Principles for financial market infrastructures. CPSS-IOSCO. https://www.bis.org/cpmi/publ/d101a.pdf

[8] Binance. (n.d.). Auto-deleveraging. Binance Support. https://www.binance.com/en/support/faq/auto-deleveraging

[9] Drift Protocol. (n.d.). Socialized loss & ADL. Drift Protocol Documentation. https://docs.drift.trade/

[10] Chitra, T. (2025). Autodeleveraging: Impossibilities and optimization. *arXiv preprint*, arXiv:2512.01112. https://arxiv.org/abs/2512.01112 https://doi.org/10.48550/arXiv.2512.01112

[11] CoinGlass. (2025). Liquidation data — October 10, 2025. https://www.coinglass.com/liquidation

[12] Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management*, *30*(5), 15–29.

[13] Diamond, D. W., & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, *91*(3), 401–419. https://doi.org/10.1086/261155

[14] Arkham Intelligence. (2025). Hyperliquid whale passes $4M loss to HLP vault. Arkham Research. https://info.arkm.com/research/hyperliquid-whale-passes-4m-loss-to-hlp-vault

[15] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

[16] Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, *107*(3), 797–817. https://doi.org/10.2307/2118364

[17] Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026. https://doi.org/10.1086/261849

[18] Shiller, R. J. (2019). *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton University Press. https://doi.org/10.1515/9780691189970

[19] The Block. (2023). FTX used random numbers to generate the size of its insurance fund. https://www.theblock.co/post/255352/ftx-used-random-numbers-to-generate-the-size-of-its-insurance-fund

[20] Asmussen, S., & Albrecher, H. (2010). *Ruin probabilities* (2nd ed.). World Scientific.

[21] Oak Research. (2025). Hyperliquid JELLY attack: Context, vulnerability, team solution. https://oakresearch.io/en/analyses/investigations/hyperliquid-jelly-attack-context-vulnerability-team-solution

[22] Halborn. (2025). Explained: The Hyperliquid hack (March 2025). https://www.halborn.com/blog/post/explained-the-hyperliquid-hack-march-2025

[23] Chainalysis. (2023). Oracle manipulation attacks rising: A unique concern for DeFi. https://www.chainalysis.com/blog/oracle-manipulation-attacks-rising/ (archived at web.archive.org)

[24] Smart Contract Hacking. (2025). Oracle manipulation & price manipulation attacks. https://smartcontractshacking.com/attacks/oracle-manipulation-attacks

[25] Solidus Labs. (2022). The Mango Markets exploit: An order book analysis. https://www.soliduslabs.com/post/mango-hack

[26] U.S. Securities and Exchange Commission. (2023). SEC charges Avraham Eisenberg with manipulating Mango Markets [Press release]. https://www.sec.gov/newsroom/press-releases/2023-13

[27] United States v. Eisenberg, No. 23-cr-10 (S.D.N.Y. May 23, 2025) (order granting Rule 29 motion and vacating conviction); see also CoinDesk. (2025, May 24). Mango Markets exploiter Avraham Eisenberg's fraud conviction overturned.

[28] Hart, O., & Moore, J. (1990). Property rights and the nature of the firm. *Journal of Political Economy*, *98*(6), 1119–1158. https://doi.org/10.1086/261729

[29] Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. https://doi.org/10.2307/1879431

[30] Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374. https://doi.org/10.2307/1882010

[31] European Union. (2023). Regulation (EU) 2023/1114 of the European Parliament and of the Council of 31 May 2023 on markets in crypto-assets (MiCA). *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/reg/2023/1114/oj

[32] CoinGecko. (2026, February). What is October 10th? Crypto's 10/10 mass market liquidation event. https://www.coingecko.com/learn/october-10-crypto-crash-explained

[33] Adrian, T., & Brunnermeier, M. K. (2016). CoVaR. *American Economic Review*, *106*(7), 1705–1741.

[34] Acharya, V. V., Pedersen, L. H., Philippon, T., & Richardson, M. (2017). Measuring systemic risk. *The Review of Financial Studies*, *30*(1), 2–47. https://doi.org/10.1142/9789814417501_0003

[35] Binance. (n.d.). Insurance fund & SAFU. Binance Support. https://www.binance.com/en/support/faq/insurance-fund

[36] dYdX. (n.d.). dYdX community dashboard. https://dydx.community/dashboard

[37] Hyperliquid. (n.d.). Hyperliquid vaults dashboard. https://app.hyperliquid.xyz/vaults

[38] Hyperliquid. (n.d.). Hyperliquid documentation. https://hyperliquid.gitbook.io/hyperliquid-docs

[39] The Block. (2025). Hyperliquid whale 50x leverage ETH event. https://www.theblock.co/post/344553/hyperliquid-whale-50x-leverage-eth

[40] Hyperliquid. (2025). Post-JELLY risk-management updates [Documentation, risk-management/clearinghouse section]. Hyperliquid Documentation. https://hyperliquid.gitbook.io/hyperliquid-docs (accessed June 2026; see the documentation's risk-management section)

Part V
Information and Efficiency

Core question: Is the perpetual futures market an efficient information-processing machine, or is it full of predictable regularities? Whoever commands the information advantage commands this market.

The Process of Price Discovery and Information Incorporation

On February 21, 2025, a security incident of unprecedented scale exposed the complex reality of how information propagates through crypto markets. More than $1.5 billion in digital assets was stolen from the cold wallet of Bybit, the world's second-largest exchange [1]. Although the movement of this enormous sum was fully public and transparent on the blockchain, the market's reaction exhibited a pronounced time lag and structural divergence. For nearly two hours after the anomalous transfer, the price of Ether moved only slightly. Only when Bybit's chief executive officer (CEO) formally confirmed the hack on social media did the price swing sharply within a few dozen minutes, with ETH falling rapidly from $2,828 to $2,708, a drop of 4.2% [2].

During this crisis, different markets absorbed the information at markedly different speeds and magnitudes. The perpetual futures market came under pressure first, and the spot market followed. Even more striking was the shock to market depth: within hours of the event's confirmation, the 1% market depth for Bitcoin, Ether, and the top 50 altcoins fell by 59%, plunging from $68 million to $28 million [3]. Why is one and the same piece of public on-chain information digested at different speeds across different trading venues? Why, when the data are fully public, does the market still exhibit an information-asymmetry window lasting as long as two hours? A second, more granular scenario reveals another face of the price discovery mechanism. At the level of market microstructure, price divergence for the same asset across different markets is common. For example, the last traded price of Bitcoin perpetual futures on one leading exchange might be $61,234, while the last traded price in the spot market of another major exchange is $62,087. Within the same second, the two prices differ by about 1.4%. Faced with two different prices for the same asset, which is the "real" one? The standard textbook answer is that the more efficient a market's price discovery, the closer its price lies to true value. Yet at this particular microstructural moment, the lower perpetual futures price did not arise because informed traders possessed fundamental information that Bitcoin should depreciate; it arose because a liquidation cascade was unfolding in the perpetual futures market. Forced-liquidation sell orders pushed the price below its fundamental level. Five minutes later, the liquidations ended, and the perpetual price recovered and reconverged with spot. The $853 gap generated during those five minutes was not *information* but *mechanical noise*. These two contrasting scenarios distill the central questions of this chapter: In digital asset markets dominated by perpetual futures, how does the process of price discovery unfold? And how does it

differ fundamentally from traditional financial markets?

The perpetual futures market plays a dual role in price discovery. On the one hand, under normal market conditions, perpetual futures are an efficient price discovery instrument: high leverage, low barriers to entry, and abundant liquidity make them the venue of choice for informed traders to express their views, and information is incorporated into prices there fastest. On the other hand, under extreme market conditions, their mechanism design injects non-informational noise into prices: the liquidation mechanism, funding rates, and the way the mark price is calculated inject substantial *mechanical noise*, causing prices to deviate systematically from fundamentals. Understanding this dual character—accelerated information incorporation coexisting with noise injection—is key to assessing the efficiency of perpetual futures markets.

Systematically dissecting this process is key to understanding how crypto markets work. Section 13.1 establishes the theoretical foundations of price discovery and examines how information is converted into order flow. Section 13.2 constructs the *three-stage price discovery model*, decomposing price discovery into three separable processes: expression, aggregation, and transmission. Section 13.3 proposes the *venue gravity model* to explain why perpetual futures attract the most informed traders. Section 13.4 analyzes cross-market price competition through the information flow topology. Section 13.5 dissects the three pathways through which perpetual futures inject mechanical noise. Section 13.6 develops the *ceiling-floor model* to delineate the boundaries of efficiency. Sections 13.7 and 13.8 examine, respectively, the obstacles to price discovery under extreme conditions and the new paradigm brought by on-chain order book decentralized exchanges (DEXs). Drawing these analyses together, Section 13.9 summarizes the dual nature of perpetual futures in price discovery.

13.1 The theoretical foundations of price discovery

Price discovery is the most central function of financial markets. In essence, it aggregates the private information dispersed across the minds of countless participants into a single public, consensus price through the interaction of market mechanisms [4]. In classical microstructure theory, the Kyle model and the Glosten-Milgrom model provide the foundational framework for understanding this process. Yet the distinctive three-dimensional information space of crypto markets, their 24/7 continuous trading, extremely high leverage, and the transparency of on-chain data require a systematic reconstruction of these classical models. This section builds all the theoretical tools needed for the analysis that follows and shows how information is priced through order flow.

13.1.1 The three-dimensional information space of crypto markets

Information is the starting point of price discovery. In traditional financial markets, information is typically divided into public information (such as macroeconomic data and corporate earnings) and private information (such as insider knowledge and institutions' in-depth research). The underlying architecture of crypto markets—the blockchain—expands the information space from two dimensions to three, introducing an entirely

new dimension: on-chain observable information.

To understand the structure of this three-dimensional information space systematically, Table 13-1 compares the three information types across four dimensions: accessibility, interpretation threshold, typical sources, and crypto-specific examples. This classification framework reveals a key difference: the boundary between public and private information is relatively clear in traditional finance, whereas the introduction of on-chain observable information blurs that boundary—it is technically open to everyone, yet in practical terms it may carry informational value only for a handful of specialized institutions.

Information type	Accessibility	Interpretation threshold	Typical sources	Crypto-specific examples
Public information	Accessible to everyone	Relatively low	News media, official announcements, macroeconomic data releases	Exchange listing/delisting announcements, regulatory policies (e.g., the SEC or MiCA), on-chain governance voting results of protocols
Private information	Held by only a few	Relatively high	Internal decisions, in-depth investment research, high-frequency trading algorithm signals	A project team's soon-to-be-announced partnership, token unlock schedules, a market maker's current inventory state and risk exposure
On-chain observable information	Technically visible to everyone	Extremely high	Blockchain explorers, the mempool, on-chain analytics tools	Large fund transfers by whale addresses, computable liquidation thresholds in DeFi protocols, pending transactions not yet included in a block

Table 13-1. A multidimensional comparison of three information types: public information, private information, and on-chain observable information (Data source: compiled by the author)

This three-dimensional information space fundamentally alters the nature of information asymmetry. In traditional markets, information asymmetry arises primarily from *who possesses the information*; in crypto markets, especially when confronting vast volumes of on-chain data, it arises more from *who can interpret the information faster and more accurately*. Even among participants of comparable technical ability, cogni-

tive biases can produce systematically divergent readings of the same on-chain data. The anchoring effect leads observers to classify a large transfer as "internal wallet rotation," a high-frequency prior event; confirmation bias prompts people to selectively ignore anomalous features inconsistent with their existing judgment; and the availability heuristic, reinforced by the recent absence of major security incidents, further depresses risk assessments. Together, these cognitive factors and the technical barrier form a multilayered wall to information interpretation.

On-chain data has the characteristic of being "public but non-consensus." Although anyone can view every transaction through a blockchain explorer, the raw hexadecimal data are unreadable to ordinary investors. Only analytical teams with professional data-engineering capabilities—such as Nansen, Arkham, or quantitative hedge funds—can extract valuable trading signals from these vast datasets. As the Bybit hack demonstrated, even though the anomalous transfer had already occurred on-chain and was publicly visible, the market still lagged nearly two hours before reacting. Subsequent investigation revealed this to be a supply chain attack targeting the Safe{Wallet} multisig infrastructure: by compromising a developer's terminal, the attackers injected malicious code into the signing interface, so that the transaction was tampered with at the moment of signing—something Bybit's internal operational monitoring systems could barely detect in real time. This phenomenon is known as the *transparency paradox*: absolute data transparency does not deliver absolute information symmetry. The information delay stemmed not only from the inadequacy of external monitoring tools but also from the operational stealth of the attack itself: when a malicious transaction is disguised as a legitimate operation at the signing layer, even institutions with on-chain monitoring capabilities struggle to link the anomalous transfer to a security incident in the first instant. By raising the threshold of data processing, this paradox widens the information gap between technically advantaged institutions and ordinary retail investors.

Figure 13-1 shows the time lag between the anomalous on-chain transfer and the market price reaction during the Bybit hack.

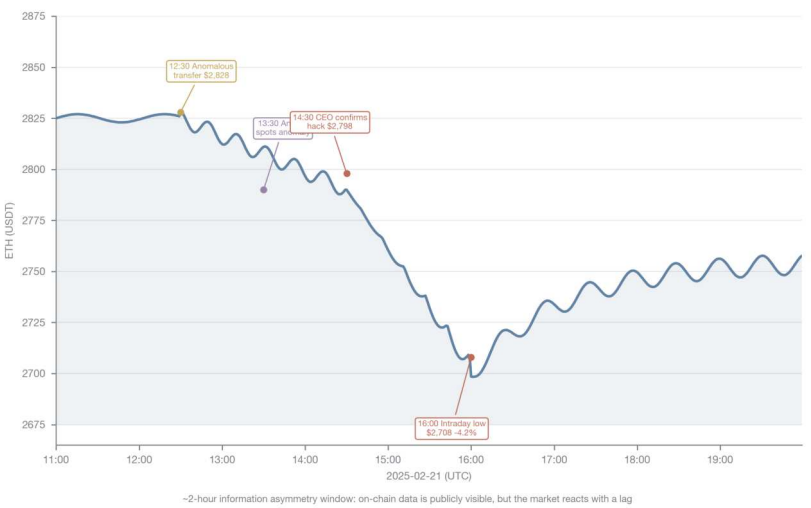


Figure 13-1. The information-asymmetry window and reaction timeline of the ETH price during the Bybit hack (price path is illustrative; key event points are empirical, per Kaiko/CoinDesk; time anchors are calibrated to the official timeline)

As shown in Figure 13-1, even though the on-chain data were fully public, the market reaction still lagged significantly, highlighting the gap between information access and information interpretation. In the nearly two hours after the anomalous on-chain transfer, only a handful of institutions with real-time monitoring and automated analytical capabilities exerted directional pressure on the price, while the vast majority of market participants did not begin to adjust their positions until the CEO publicly confirmed the incident on social media. This delay reveals that information asymmetry in crypto markets stems not from the inaccessibility of data but from the uneven distribution of the capacity to parse it.

Information asymmetry manifests not only in the delay of price reactions but also in dramatic shifts in liquidity structure. Figure 13-2 compares three key liquidity metrics before and after the confirmation of the Bybit hack: 1% market depth, Bybit's share of global liquidity, and ETH perpetual futures open interest.

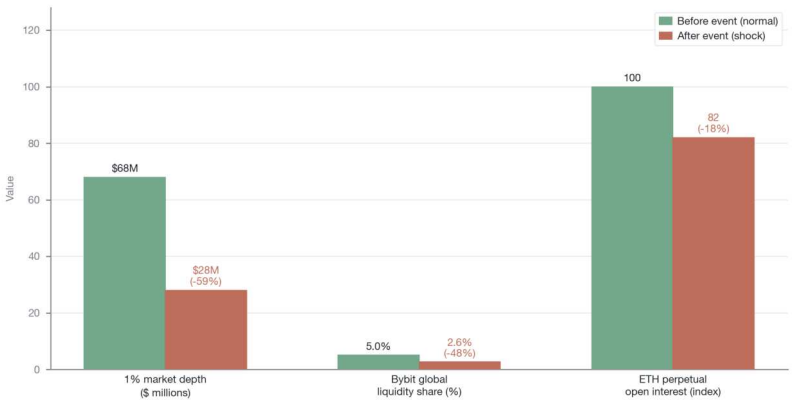


Figure 13-2. Changes in market liquidity depth before and after the Bybit hack (pre- and post-event depth relies on Kaiko paid data, not independently verified)

Figure 13-2 shows that at moments of high information asymmetry, the rational response of market makers is to withdraw liquidity. Market depth fell from $68 million to $28 million not linearly but in a concentrated fashion within the first 30 minutes after confirmation, then held at a low level for several hours; its evaporation took only a few dozen minutes, while recovery to pre-event levels took roughly several days, forming a pronounced asymmetry of "minute-scale evaporation versus day-scale recovery." This pattern reflects market makers' "retreat first, assess later" strategy when facing unquantifiable adverse-selection risk.

Having clarified the distinct features of the three information types, the transmission relationships among them are equally important. Figure 13-3 summarizes the overall architecture of the three-dimensional information space of crypto markets, together

with the transmission paths and convergence nodes of each of the three information types.

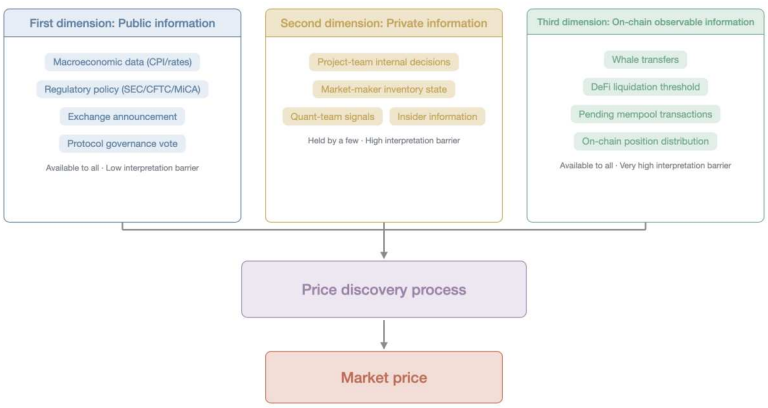


Figure 13-3. The three-dimensional information space of crypto markets and its transmission paths

Figure 13-3 reveals a core paradox of on-chain observable information: although such data are open to everyone, the extremely high interpretation threshold makes them function, in practical terms, much like private information—only a handful of specialized institutions can convert them into a trading advantage. This paradox challenges a classical assumption of the Kyle model: in the traditional framework, public information should be priced in instantly, yet the "public but non-consensus" character of on-chain data in crypto markets blurs the boundary between public and private information. In the price discovery analysis that follows, therefore, information cannot be classified by accessibility alone; the technical-capability threshold required for its interpretation must also be considered.

13.1.2 The Kyle model

To understand how information enters prices, we must look closely at the microstructure level. The classic model proposed by Albert Kyle in 1985 offers a concise and systematic game-theoretic framework [5]. The model divides market participants into three classes: informed traders, noise traders (or liquidity traders), and market makers.

In the crypto context, an informed trader might be an analyst who detects a hack early through on-chain monitoring, or a hedge fund that possesses some quantitative signal. Such traders hold private information about an asset's future value and seek to profit from it through trading. However, if they dump a large order all at once, they immediately alert market makers and trigger a large adverse price move (a substantial price impact). Informed traders therefore behave strategically: they split large orders into smaller ones and inject them into the market gradually over time.

Noise traders, by contrast, are participants who trade for non-informational reasons—

for example, to hedge risk, rebalance a portfolio, or simply as retail investors driven by market sentiment. In the model, they play the role that directly determines market liquidity: they provide "cover" for informed traders. When market makers observe the order flow, they see the aggregate order flow, a blend of informed orders and noise orders, and cannot accurately distinguish the true source of each order. Without noise traders, every action of an informed trader would be perfectly identified by market makers, and the price discovery process would immediately stall.

Market makers, as liquidity providers, have as their core task to "learn from the order flow." They observe net buying or selling pressure and infer the underlying true value through Bayesian updating, adjusting their quotes accordingly. The magnitude by which market makers adjust their quotes is governed by a key parameter: Kyle's lambda (λ).

Mathematically, the price change is linear in net order flow: $\Delta P = \lambda \times \text{Net Order Flow}$.

Kyle's lambda (λ) measures the degree to which order flow impacts price; it simultaneously reflects two key market characteristics: market depth and the degree of information asymmetry. Specifically, the smaller λ is, the smaller the price change caused by absorbing an order of a given size, and the better the market depth; conversely, the larger λ is, the higher the proportion of "toxic" information (i.e., informed trading) that market makers believe the current order flow contains, so they mount a more aggressive price defense against each unit of net order flow. After major shocks such as hacks or severe enforcement actions by regulators, λ typically rises sharply. The measurement convention for net order flow must be made explicit (for example, in units of the base asset or in notional U.S. dollars); the dimension of λ is then determined as "price change per unit quantity." Different conventions can make the numerical value of λ differ by several orders of magnitude, so a consistent convention must be used for cross-market comparisons.

Figure 13-4 illustrates how the Kyle model operates in crypto markets.

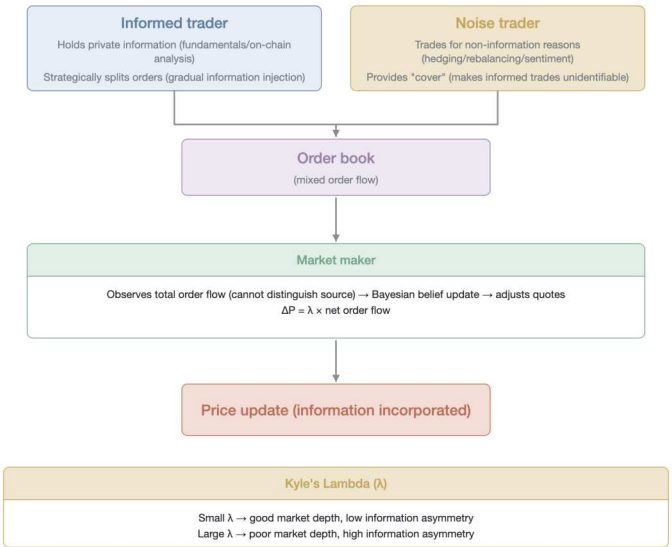


Figure 13-4. How the Kyle model operates in crypto markets

Figure 13-4 shows how market makers dynamically adjust their quotes by observing the aggregate order flow—a blend of informed orders and noise orders—thereby completing the conversion of information into price. In crypto markets, this mechanism is further complicated by the presence of competitive market making: multiple high-frequency trading firms compete in quoting simultaneously, which effectively compresses the bid-ask spread in normal times but can produce synchronized collective withdrawal during sudden shocks. This competitive market-making structure both accelerates price discovery in normal times and amplifies liquidity fragility in times of crisis.

13.1.3 Adverse selection and the bid-ask spread

If the Kyle model explains how prices move with order flow, then the model proposed by Glosten and Milgrom in 1985 answers another fundamental question: why the bid-ask spread exists, and how it reflects information asymmetry [6].

In a cryptocurrency order book, market makers post quotes on both the bid and the ask side, passively waiting for other traders to trade against them. This passivity exposes market makers to a systematic risk: adverse selection.

When an informed trader knows that Bitcoin's true value is above the current market quote, they buy (lifting the market maker's ask); when they know the true value is below the current quote, they sell (hitting the market maker's bid). This means that, when trading against informed traders, market makers suffer a systematic loss in a statistical sense. To offset this inevitable loss, market makers must earn profits from trades with noise traders (who buy and sell randomly).

The Glosten-Milgrom model therefore reaches a core conclusion: the bid-ask spread is, in essence, the market maker's pricing of adverse-selection risk.

The width of the spread is positively correlated with the "toxicity" of informed trading in the order flow (i.e., the proportion of informed traders in total order flow, denoted μ). The market maker's ask and bid quotes equal, respectively, the conditional expected estimate of the asset's true value upon observing the arrival of a buy order or a sell order, and the difference between the two is the bid-ask spread.

In crypto markets, this mechanism is clearly borne out. For highly liquid assets such as Bitcoin and Ether, the sheer volume of noise traders (an extremely low μ) keeps the bid-ask spread at leading exchanges such as Binance typically below one basis point (0.01%). This figure refers to the top-of-book spread during normal trading hours; in low-liquidity weekend periods or on the eve of major news releases, even the bid-ask spread for BTC can widen to several basis points. For small-cap, highly information-asymmetric altcoins, however, market makers face extremely high adverse-selection risk, and their spreads often reach tens or even hundreds of basis points. Moreover, on the eve of major news releases or extreme market moves, once market makers sense that adverse-selection risk is rising, they immediately cancel their tight quotes, causing spreads to widen rapidly.

Figure 13-5 depicts the quote-adjustment logic of market makers facing different types of traders in the Glosten-Milgrom model, and how adverse-selection risk determines the width of the bid-ask spread.

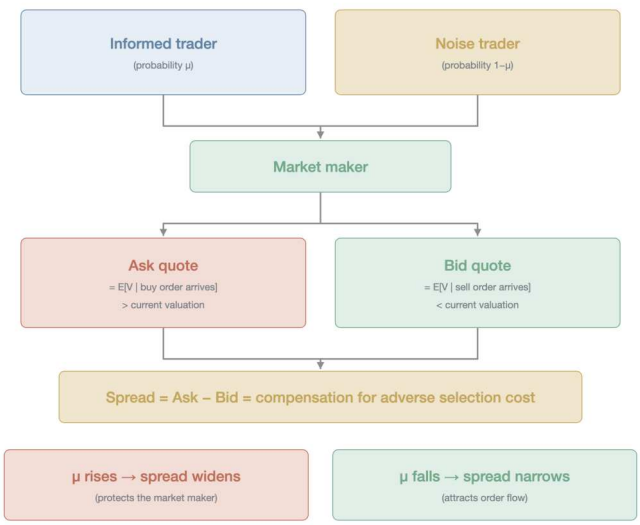


Figure 13-5. Adverse selection and the bid-ask spread formation mechanism in the Glosten-Milgrom model

Figure 13-5 shows that the bid-ask spread is not a simple transaction fee but a necessary compensation that market makers charge for adverse-selection risk in an environment of information asymmetry. In crypto markets, the manifestation of this mechanism differs especially sharply: for highly liquid assets such as BTC and ETH, the bid-ask spread is typically below one basis point, whereas for highly information-asymmetric

altcoins, the spread can reach tens or even hundreds of basis points. This cross-asset divergence in spreads provides a direct quantitative indicator for assessing the quality of the information environment across different markets.

13.1.4 Order flow information metrics

The theoretical models show that order flow is the core carrier of information transmission. In quantitative finance and microstructure research, scholars have developed a variety of metrics to measure precisely the information content of order flow. In crypto derivatives markets, the following three metrics are especially important.

Order flow imbalance. Order flow imbalance (OFI) is the most direct measure of microstructural buy-sell pressure. Unlike traditional metrics that consider only executed trades, OFI, over a given time window, jointly accounts for all events that alter the state of the best bid and ask in the order book; these events include the execution of market orders, the submission of new limit orders, and the cancellation of existing limit orders.

Silantyev's (2019) empirical study of cryptocurrency perpetual futures markets revealed the powerful explanatory power of OFI [7]. In a study using Level 2 data for BitMEX's XBTUSD perpetual futures, OFI, over time windows from 10 seconds to 5 minutes, could explain 40% to 70% of mid-price changes (i.e., the explanatory power of the regression). This directly corroborates the core theory of the Kyle model: order flow carries information, and price impact reflects the process by which information is incorporated. This study is based on BitMEX data from 2019, when the market microstructure differed markedly from that of 2025: BitMEX used a single margin currency (BTC), and the degree of market-maker competition was far lower than in today's multi-exchange, unified-margin market structure. The core conclusion that order flow carries information has continued to be validated in subsequent research, but the specific R^2 values may differ under the more competitive market structure of today.

Trade flow imbalance. Trade flow imbalance (TFI) is a subset of OFI; it computes only the difference between the volume of aggressive buy market orders and the volume of aggressive sell market orders within a specific time period. Because executing a market order requires paying a liquidity premium (crossing the bid-ask spread), such "cost-bearing" trading behavior is generally considered to embody a stronger directional conviction.

Research shows that over very short time windows (such as 1 second or 10 seconds), OFI, which incorporates limit-order dynamics, has stronger explanatory power for price; but as the time window lengthens (beyond 1 minute), the explanatory power of TFI gradually approaches that of OFI, and the two converge. This is because high-frequency limit-order placement and cancellation (or market-maker adjustment) tend to contain more short-term noise, whereas substantive turnover of capital contributes to price discovery over longer time scales and gradually becomes comparable to OFI.

The theoretical difference above is clearly borne out in empirical data. Figure 13-6, based on Silantyev's (2019) BitMEX XBTUSD data, compares the regression explanatory power (R^2) of OFI and TFI for changes in the perpetual futures mid-price across

time windows from 10 seconds to 5 minutes.

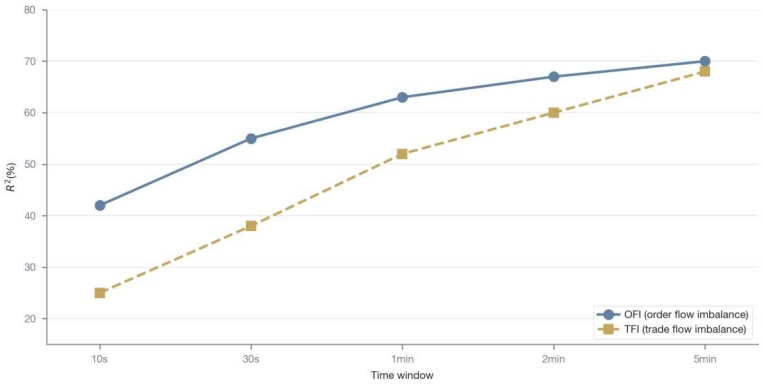


Figure 13-6. The time-window effect in the explanatory power of OFI and TFI for perpetual futures price changes (Data source: Silantyev, 2019 [7])

Figure 13-6 exhibits a pronounced convergence pattern: over the very short time windows of 10 and 30 seconds, the R^2 of OFI is significantly higher than that of TFI, because at high-frequency scales the submission and cancellation of limit orders carry rich instantaneous supply-and-demand information. As the time window lengthens beyond 1 minute, however, the explanatory power of TFI gradually approaches that of OFI, and the two converge. This convergence carries a profound microstructural implication: high-frequency limit-order dynamics contain considerable strategic-adjustment noise from market makers, whereas the substantive turnover of capital measured by market orders contributes to price discovery over longer time scales and gradually becomes comparable to OFI.

Volume-synchronized probability of informed trading. The VPIN metric, proposed by Easley, López de Prado, and O'Hara in 2012, is a major improvement on the traditional probability of informed trading (PIN) model [8]. The core innovation of VPIN is that it abandons "clock time" (such as each minute or each hour) as the unit of analysis and instead adopts "volume time."

Specifically, VPIN divides the continuous trade flow into "buckets" of equal volume. Only when the market has accumulated a specific amount of volume does the system advance one period. Within each volume bucket, it computes the degree of absolute imbalance between buyer-initiated volume and seller-initiated volume to estimate the "toxicity" of the order flow (the probability of informed trading).

This design is markedly well-suited to crypto markets. Crypto markets operate 24/7, and their liquidity and trading activity exhibit enormous intraday-rhythm differences across time zones (such as the Asian session versus the overlapping European and U.S. session). If a fixed time window were used, the metric would be full of noise during quiet periods. By anchoring on volume, VPIN naturally adapts to this time-varying feature: when trading is thin, VPIN updates slowly; during periods of sharp volatility and elevated volume, VPIN samples at high frequency, promptly reflecting the rise in information asymmetry.

In several historic crashes in crypto markets (such as the March 12, 2020 event), VPIN showed a pronounced upward trend hours before prices substantively plunged, indicating that informed traders (or panicked first-movers) had already begun to dump orders unidirectionally and that the adverse-selection risk facing market makers was accumulating sharply. VPIN is therefore not merely an academic tool for understanding price discovery but also an effective early-warning indicator under extreme conditions.

Figure 13-7 shows, respectively, the variation of Kyle's lambda across different market states and the alternation in the explanatory power of OFI and TFI across different time scales.

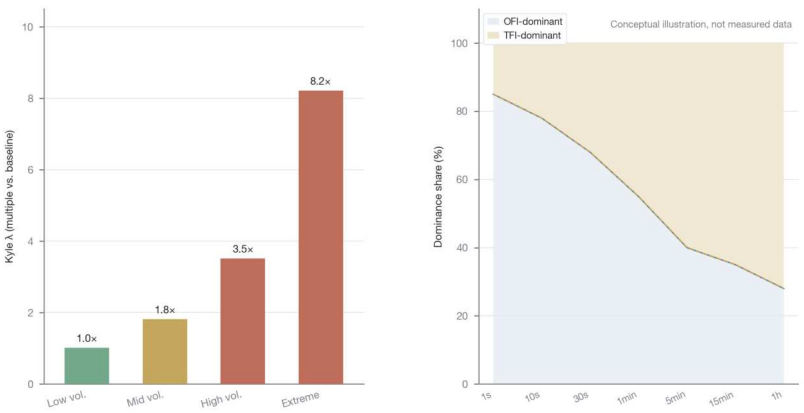


Figure 13-7. The market-state dependence of Kyle's lambda and the time-scale effect of OFI/TFI (conceptual illustration, not empirical data)

The left panel of Figure 13-7 shows that in extreme events, the drying up of liquidity causes the value of lambda to rise sharply and price impact to multiply; the right panel conceptually depicts the alternating relationship between OFI and TFI in leading price discovery across different time scales.

13.1.5 The distinctive features of crypto markets

Although the Kyle and Glosten-Milgrom models provide a solid theoretical foundation, applying them directly to crypto markets is inadequate. The microstructure of digital asset markets challenges and revises the classical assumptions along six key dimensions. The fragmentation across multiple trading venues breaks the classical models' premise of a single trading venue: the same asset trades simultaneously on hundreds of centralized exchanges (CEXs) and decentralized exchanges, turning price discovery into a cross-market network-transmission process in which arbitrageurs play the key role of information carriers. Extremely high leverage (perpetual futures, for instance, allow leverage of up to 100x) is, within the Kyle framework, equivalent to amplifying the capital efficiency of informed traders, enabling them to exert enormous price pressure on the market with very little principal and to accelerate the expression of information; but it also amplifies the destructive power of noise traders in equal proportion, sowing the

seeds of liquidation cascades. The 24/7 continuous trading eliminates the "overnight information backlog" and the opening-auction gaps of traditional equity markets, allowing information to be priced continuously; yet during low-liquidity periods such as weekends or holidays, a small injection of information can trigger violent price swings. The heterogeneity of information types (such as the introduction of on-chain observable information) creates a wholly new information game that is "public yet highly asymmetric." Competitive market making changes the premise of a single monopoly market maker: fierce competition among multiple high-frequency trading firms greatly compresses the bid-ask spread and raises aggregation efficiency in normal times, but can lead to collective withdrawal and a liquidity vacuum during extreme crises. The embedding of mechanical noise is a feature unique to perpetual futures markets: the polarization of funding rates and the mechanical execution of forced-liquidation engines produce a deterministic, structural price impact (σ_{mech}) that differs from the price deviations rooted in irrational behavior found in traditional models.

These revisions do not negate the core logic of the classical models—prices are still discovered through information asymmetry in order flow—but they require us to treat "market structure" itself as a key variable in our analysis. The "multiple trading venues" feature in particular directly gives rise to the "three-stage price discovery model" and the "information flow topology" explored in the sections that follow.

13.2 The three-stage price discovery model

In traditional financial markets, price discovery is usually treated as a relatively concentrated black-box process. In the highly fragmented, multi-centered microstructure of crypto asset markets, however, price discovery must be decomposed into a multi-stage dynamic process. This section presents the first core theoretical contribution of this chapter: the three-stage price discovery model. The model divides the process by which information is transformed from an informed trader's private cognition into a market-wide consensus price into three interconnected stages: expression, aggregation, and transmission.

13.2.1 The expression stage

The starting point of price discovery is the expression of private information by informed traders. When an informed trader acquires new information about an asset's fundamentals (such as a macroeconomic data release, a protocol upgrade, or an anomalous on-chain fund movement), the first decision they face is not to trade directly but to choose a venue.

In crypto asset markets, the same asset often trades simultaneously in the spot and derivatives markets of dozens of centralized and decentralized exchanges. When choosing a venue in which to express information, informed traders are drawn chiefly by four dimensions of gravity: leverage availability, liquidity depth, anonymity, and transaction cost. This selection mechanism is developed in detail in the "venue gravity model" of Section 13.3. Empirical research shows that, because perpetual futures hold significant advantages along these dimensions, informed traders tend to submit orders in the

perpetual futures market first [9].

Once a venue is chosen, informed traders convert their private information into actual trading instructions by submitting market or limit orders. The efficiency of this process is termed expression efficiency, and it depends on how accommodating the market microstructure is to information traders. If the market offers high leverage and low-latency API access, informed traders can express their information at very low capital and time cost, and expression efficiency is high; conversely, if the market lacks liquidity or imposes strict access restrictions, information expression is impeded.

The binary division of traders into informed and noise traders in Kyle's (1985) classic framework requires substantial extension in crypto asset markets. In that framework, informed traders hold private information about an asset's fundamentals while noise traders trade out of liquidity needs or random motives, and the boundary between the two is assumed to be clear and stable. Yet crypto markets contain a large intermediate category with markedly different behavioral characteristics: *narrative-driven traders*. These traders are driven by market narratives circulating on social media, by the guidance of key opinion leaders (KOLs), and by gamified platform elements (such as copy trading, profit leaderboards, and community contests). Subjectively, they are convinced they possess valuable information, but in reality they are responding to a *social amplification loop*: after information is repeatedly forwarded and emotionally colored across Twitter (X), Telegram groups, and Discord communities, its original signal content has been severely diluted or even wholly distorted. The key difference between narrative-driven traders and pure noise traders lies in the persistence and directionality of their behavior: the price impact of pure noise trading is statistically transient, because the directions of random trades cancel one another out; whereas narrative-driven traders, sharing the same narrative frame, trade in a manner highly correlated in both time and direction, generating persistent directional pressure that drives prices away from fundamental value for extended periods. The existence of this intermediate category poses a distinctive challenge to signal extraction in the expression stage: in later stages, market makers and aggregation mechanisms face not only the classic "signal plus white noise" structure but also this "colored noise" component, which carries a persistent bias while conveying no genuine fundamental information.

13.2.2 The aggregation stage

The orders produced in the expression stage flow into a specific venue's central limit order book (CLOB) or automated market maker (AMM) liquidity pool, entering the aggregation stage. At this stage, market microstructure mechanisms (chiefly market makers and the matching engine) process these orders and extract the information embedded within them, forming that venue's local consensus price.

The core challenge of the aggregation stage lies in separating signal from noise. When market makers receive order flow, they face adverse-selection risk: they cannot directly distinguish whether an order comes from an informed trader holding private information (signal) or from a noise trader acting on liquidity needs or sentiment (noise). Market makers respond to this risk by adjusting their bid and ask quotes.

When an informed trader's buy order consumes sell-side liquidity on the order book, it

produces a price impact. According to microstructure theory, this price impact can be decomposed into three components. The information component is permanent: once new information is incorporated into the price, the price does not revert—this is the effective output of price discovery. The liquidity component is transient: it arises from a price shift due to momentary insufficient depth, and once market makers replenish liquidity the price partially reverts. The mechanical component is non-informational volatility: forced trading triggered by mechanisms such as liquidation causes violent price swings, and after the impact passes the price usually reverts completely.

Aggregation efficiency measures the market mechanism's ability to accurately extract the information component from order flow while filtering out the liquidity and mechanical components. In a normal market with ample liquidity, aggregation efficiency is high; but during extreme volatility, the forced-liquidation orders triggered by the liquidation mechanism (mechanical noise) can swamp genuinely informative orders, causing aggregation efficiency to plummet.

13.2.3 The transmission stage

In the highly fragmented crypto market, once information has been aggregated into a price at a particular venue (the leading market), it must diffuse through transmission mechanisms to all other trading venues (the following markets) before a final market-wide consensus price can form. This is the transmission stage.

The transmission stage relies chiefly on two mechanisms. The direct price chain (the arbitrage mechanism) depends on high-frequency arbitrageurs monitoring price differences across markets. When the price in the leading market (such as Binance perpetual futures) rises on new information while the price in a following market (such as certain regional spot exchanges or DEXs) has not yet reacted, arbitrageurs sell in the leading market and buy in the following market. This cross-market arbitrage forces the following market's price to converge toward the leading market's by consuming the following market's liquidity [4]. The indirect price chain (the oracle mechanism) serves decentralized finance protocols: when the leading market's price moves, oracle networks (such as Chainlink) capture these movements and update the on-chain price feed, directly altering the pricing benchmark of on-chain derivatives or lending protocols.

Transmission efficiency depends on the friction costs of cross-market arbitrage (including transaction fees, withdrawal delays, network congestion, and so on) and on the update frequency of oracles. Makarov and Schoar's (2020) empirical study notes that although the arbitrage efficiency of crypto markets keeps improving, during periods of sharp market volatility cross-exchange price differences can persist for minutes or even longer, indicating that transmission efficiency is significantly impaired under stress [10].

The analysis above, however, implicitly treats arbitrageurs as rational agents with unlimited capital and perfect execution—an assumption that is far from holding in reality. As Shleifer and Vishny (1997) argued in their classic paper on "the limits of arbitrage" [11], real-world arbitrageurs face serious capital constraints and agency problems, and these constraints are especially pronounced in the transmission stage of crypto markets. First, crypto arbitrageurs typically use leverage to amplify capital efficiency, their own capital is limited, and in extreme conditions they cannot serve as

a price-correcting force; on the contrary, they may become passive amplifiers of price dislocations through margin calls or even forced liquidation. Second, cross-exchange arbitrage requires arbitrageurs to pre-position capital across multiple exchanges (the pre-positioned inventory model), because crypto assets lack a cross-exchange clearing and settlement mechanism, so once funds are deposited at one exchange they are locked within that exchange's account system. This pre-positioned model entails a high capital-locking cost: arbitrageurs must maintain idle funds at every potential arbitrage venue, and the opportunity cost of those funds is especially significant in a high-interest-rate environment. More importantly, the sources of friction in the transmission stage must be distinguished along two dimensions of entirely different nature. Technical transmission friction—including transfer-confirmation delays caused by blockchain network congestion, cross-chain bridge processing times, and the like—is essentially a physical constraint of infrastructure performance that improves as a trend with technological progress. Institutional transmission friction—that is, exchanges deliberately suspending withdrawal processing, rate-limiting their API interfaces, or even temporarily shutting down during extreme conditions—is by contrast a deliberate decision by exchanges based on their own risk-control logic. The latter's impact in extreme events often far exceeds the former's: when several major exchanges suspend withdrawals simultaneously, the physical channels for cross-exchange arbitrage are fundamentally severed, and no matter how much capital or how fast an execution speed an arbitrageur possesses, they cannot complete the transfer of positions across exchanges. This means that the collapse of transmission efficiency under stress is not merely a matter of rising friction costs but a structural rupture of the arbitrage channels.

13.2.4 Total efficiency

The three-stage price discovery model provides a quantitative framework for assessing the quality of market-wide price discovery. The total price discovery efficiency of the market is not the simple sum of the efficiencies of these three stages but their product:

$$\text{Total efficiency} = \text{Expression efficiency} \times \text{Aggregation efficiency} \times \text{Transmission efficiency}$$

The rationale for constructing the total-efficiency formula as a product rather than a weighted sum derives from the serial structure among the three stages. Information must pass sequentially through expression, aggregation, and transmission to complete its transformation from private cognition into a market-wide consensus price; this process is essentially a serial pipeline rather than three interchangeable parallel paths. In a serial system, if the efficiency of any one link approaches zero, the output of the entire pipeline approaches zero—precisely the "bottleneck effect" naturally captured by the mathematical properties of multiplication. A weighted-sum model (such as Total efficiency $= w_1 \cdot E_{\text{expression}} + w_2 \cdot E_{\text{aggregation}} + w_3 \cdot E_{\text{transmission}}$), by contrast, implicitly assumes that the stages can compensate for one another—that high expression efficiency can offset low transmission efficiency—but this runs counter to the physical reality of information transmission: no matter how efficiently informed traders express information, if the arbitrage channels are completely severed (transmission efficiency of zero), the information cannot reach the following markets. In this model, we bound the

value of each stage's efficiency to the interval $[0, 1]$, where 1 represents the theoretical frictionless ideal state (information is processed completely, instantly, and losslessly at that stage) and 0 represents the complete failure of that stage. The product formula is a first-order approximation of the three-stage process. In reality, nontrivial interactions and feedback exist among the three stages (for example, persistently low transmission efficiency will suppress informed traders' expression in the leading market, and high noise in the aggregation stage may induce them to delay or disperse their orders); a more complete model would need to introduce inter-stage feedback terms to capture these second-order effects. But the analysis in this chapter shows that the first-order product approximation already suffices to reveal the core fragility of price discovery in crypto markets.

This multiplicative relationship reveals a key fragility of price discovery in crypto asset markets: the weakest-link effect. Even if the market performs with extremely high efficiency at one or two stages, a serious failure at just one stage causes the overall price discovery function to collapse.

For example, during the extreme liquidation crisis of October 2025, informed traders could still submit orders efficiently (high expression efficiency), but as a flood of liquidation orders (mechanical noise) poured into the order book, market makers could not effectively separate signal from noise, and aggregation efficiency fell sharply; at the same time, a surge in Ethereum-mainnet gas prices delayed on-chain transfers, and with several exchanges suspending withdrawal processing due to system overload, cross-market margin-call delays impeded the arbitrage channels, and transmission efficiency deteriorated sharply. Ultimately, this caused the market-wide total price discovery efficiency to approach a standstill within a short time, with severe dislocations in prices across venues.

By decomposing price discovery into these three stages, researchers and regulators can more precisely locate the efficiency bottlenecks in market microstructure, providing a theoretical basis for optimizing trading mechanisms and designing more resilient DeFi protocols.

Figure 13-8 presents, as a flowchart, the complete transmission path of the three-stage price discovery model and the efficiency bottleneck at each stage.

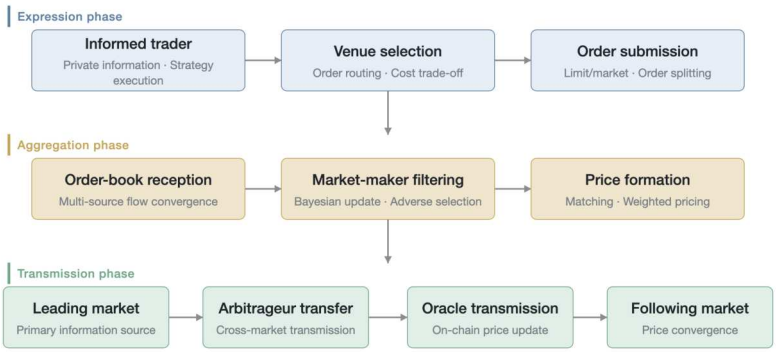


Figure 13-8. The complete transmission process of the three-stage price discovery model

As shown in Figure 13-8, information passes in turn through the expression stage of venue selection and order submission, the aggregation stage of order book matching and market-maker filtering, and the transmission stage of arbitrage and oracle diffusion, ultimately forming a market-wide consensus price.

The core of the aggregation stage lies in extracting the genuine information signal from order flow. Figure 13-9 decomposes the price impact triggered by a large trade into three components of different nature: the permanent information component, the transient liquidity component, and the non-informational mechanical component.

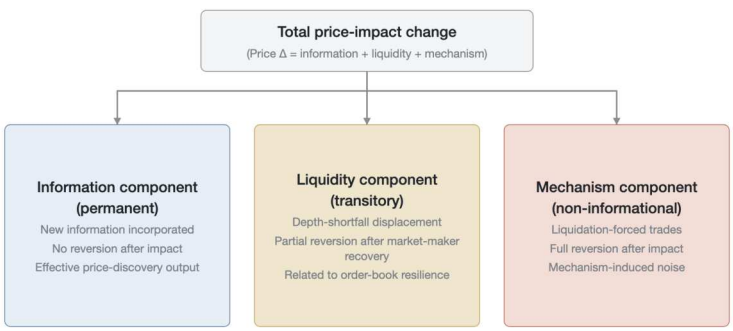


Figure 13-9. Decomposition of price impact into information, liquidity, and mechanical components (conceptual decomposition, no quantitative data)

Figure 13-9 clearly shows the differing behavior of the three components along the time dimension: the information component produces a permanent shift in the price level after the impact, and the price no longer returns to its original position—this is the evidence that new information has been absorbed by the market; the liquidity component exhibits the classic "impact-and-revert" pattern, reflecting a transient deviation caused

by momentary insufficient depth; and the mechanical component (such as the selling pressure triggered by forced liquidation) usually reverts completely once the impact fades, because these trades carry no information about the asset's fundamentals. This decomposition framework shows that the core task of the aggregation stage is to separate the information component effectively from the other two—only the permanent information component constitutes the effective output of price discovery, while the liquidity and mechanical components are noise to be filtered out.

Figure 13-10, by setting specific efficiency-parameter values, quantitatively illustrates the computed results of the multiplicative relationship among the three-stage efficiencies under two representative states: a normal market and a liquidation-intensive period.

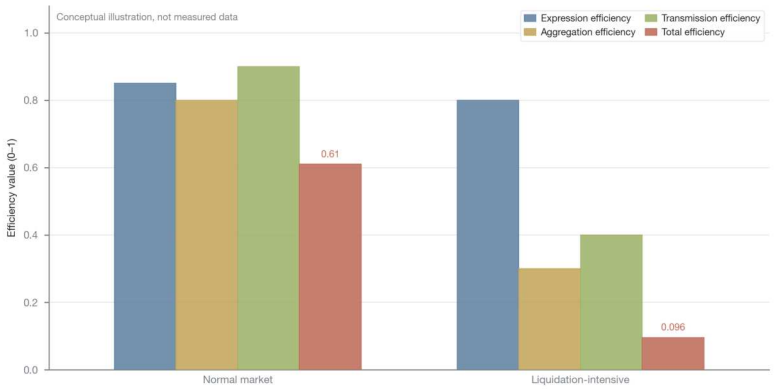


Figure 13-10. The multiplicative relationship of three-stage efficiency under different market states (conceptual illustration, not empirical data)

The computed results in Figure 13-10 reveal a key nonlinear feature: in a normal market, assuming expression efficiency of 0.85, aggregation efficiency of 0.80, and transmission efficiency of 0.90, the product of the three is about 0.61, meaning that about 39% of the information value is lost or delayed in transmission on the way from information generation to market-wide pricing. When the market enters a liquidation-intensive period, however, even if expression efficiency remains at 0.80, aggregation efficiency plunges to 0.30 due to mechanical noise and transmission efficiency falls to 0.40 due to impeded arbitrage channels, the total efficiency is only 0.096—less than one-sixth of the normal level. This feature, in which "the collapse of any single link can paralyze the entire pipeline," is precisely the empirical basis for constructing the total-efficiency formula as a product rather than a weighted sum.

13.3 The venue gravity model

In the three-stage model of Section 13.2, the core of the expression stage is informed traders' choice of venue. As noted in Section 13.1.5, the trading venues of crypto asset markets are highly fragmented. Empirical data show, however, that price discovery is not distributed evenly across these venues but is instead highly concentrated. This section presents the second core theoretical contribution of this chapter: the venue

gravity model, which explains why perpetual futures have become the absolute center of price discovery in crypto markets.

13.3.1 Informed traders' venue preferences

In traditional financial markets, the spot market is usually the cornerstone of price discovery, while the futures market plays a supporting role in risk management and price discovery. In crypto asset markets, however, this relationship has fundamentally reversed. According to CoinGlass's 2025 annual report, the crypto derivatives market recorded total trading volume of about $85.70 trillion for the full year [12]; of this, perpetual futures held the dominant position, with an estimated share of about 93% documented in industry research by Ruan and Streltsov and others [13]. Notional trading volume, however, has significant limitations as a measure of market activity. Because perpetual futures embed high leverage, notional volume is greatly inflated by the leverage multiple: a trader using 100x leverage who opens a position with $1,000 of principal generates $100,000 of notional volume, yet their actual risk capital is only $1,000. Moreover, notional volume can be distorted by wash trading (discussed in detail in Section 13.7.2). More analytically valuable measures therefore include open interest (OI), which reflects the scale of capital actually locked in the market and the commitment of positions, and leverage-adjusted, capital-weighted trading volume, which seeks to recover the true flow of risk capital behind trading volume. Even so, $85.70 trillion of notional volume, even after adjusting for leverage, still clearly indicates that the perpetual futures market has far surpassed the spot market in the absolute scale of capital activity.

The pioneering study by Alexander and colleagues (2020) shows that perpetual futures traded on unregulated exchanges are the strongest driver of Bitcoin price discovery [9]. A number of subsequent studies, including Ruan and Streltsov's (2024) analysis based on high-frequency order book data, further confirm the leading role of perpetual futures in price discovery [13].

Why do informed traders so favor perpetual futures? This is no accident but is determined by the endogenous properties of market microstructure. The venue gravity model holds that a trading venue's ability to attract information flow (its "gravity") depends on the combined performance across four key dimensions.

13.3.2 The four dimensions of gravity

We define the information gravity G_i of a trading venue i as a function of four dimensions:

$$G_i = f(L_i, D_i, A_i, C_i^{-1})$$

Specifically, leverage availability (L_i) satisfies informed traders' core demand for maximum risk exposure with minimum capital; the leverage of up to 10x–125x on perpetual futures lets them express private information with extremely low capital commitment, whereas the spot market lacks endogenous leverage and traditional futures markets have leverage capped by regulation. Liquidity depth (D_i) determines the price-impact cost of executing large orders; 2025 data show that the $\pm 1\%$ two-sided liquidity depth

for Bitcoin on Binance perpetual futures far exceeds that of any single spot market, and deep liquidity allows large information trades to be absorbed covertly and at low cost [12]. Anonymity (A_i) allows informed traders to defer the point at which their identity and intent are identified; DEX perpetual futures offer the highest pseudonymity, while CEX perpetual futures achieve a "hidden in the crowd" effect by blending order flow into a vast stream of retail trades. Yet the anonymity dimension contains a profound regulatory tension. On the one hand, anonymity lowers informed traders' cost of information leakage, making them more willing to express private information and thereby promoting price discovery; on the other hand, it also provides cover for insider trading and market manipulation. The FATF Travel Rule requires virtual asset service providers to transmit sender and receiver identity information for transfers above a certain amount, the EU's Markets in Crypto-Assets (MiCA) regulation imposes comprehensive customer-due-diligence obligations on crypto asset service providers, and the U.S. Bank Secrecy Act (BSA) extends compliance reach into the crypto domain through the suspicious activity report (SAR) regime. These regulatory frameworks are systematically narrowing the space for anonymous trading, which means that the anonymity dimension in the gravity model is not a static parameter but a time-varying factor being structurally weakened by institutional forces. Transaction cost efficiency (C_i^{-1}) must be understood from a "full cost" perspective rather than by focusing on explicit fees alone. The explicit advantages of perpetual futures—no rollover cost, no option time decay, and lower trading fees (makers often receive rebates)—are indeed significant, but the full cost of actual trading must also incorporate the following implicit dimensions: the holding cost of the funding rate, which during periods of crowded leverage can exceed 100% annualized and, for informed traders who need to hold positions for several days or more, can erode or even reverse trading profits; market-order slippage, a function of the ratio of order size to order book depth, which can significantly amplify execution costs during thin-liquidity periods; and the opportunity cost of margin capital being locked up, which likewise cannot be ignored, especially in a high-interest-rate environment. The advantage of perpetual futures in transaction cost efficiency therefore still holds in most situations, but their full cost is far higher than the impression given by observing fee rates alone.

As the following table shows, perpetual futures exhibit significant advantages along all four dimensions, forming a strong "gravitational field" that draws the vast majority of the market's information trading to them.

Table 13-2 compares the relative performance of four trading venues across the four dimensions of the gravity model.

Dimension	Perpetual futures	Spot market	Traditional futures	Options market
Leverage availability	10x–125x, endogenous to product design	No endogenous leverage, relies on external financing	5x–20x, constrained by regulation	Embedded leverage but limited by option premiums
Liquidity depth	BTC ±1% depth exceeds $50 million	Dispersed across multiple exchanges, low depth at any single one	Concentrated but limited by trading hours	Depth fragmented across strike prices and expiration dates
Anonymity	CEXs blend orders into massive retail order flow; DEXs offer pseudonymity	CEXs require KYC; DEXs are pseudonymous	Strictly regulated, low anonymity	Strictly regulated, low anonymity
Transaction cost efficiency	No rollover cost, low fees (makers often receive rebates)	Relatively higher fees, no leverage amplification	Rollover costs and expiry-roll friction	Time decay constitutes a holding cost

Table 13-2. Comparison of four trading venues across the four dimensions of the venue gravity model (liquidity depth data source: CoinGlass 2025 Annual Report [12], representing data at a specific snapshot point)

The "BTC ±1% depth exceeds $50 million" figure in Table 13-2 is a snapshot value at a specific observation point, whereas liquidity depth itself is highly time-varying. During market pullbacks, market makers contract their quotes, and depth can shrink dramatically to one-third of the normal level or even lower. More importantly, two-sided total depth (bid plus ask combined) masks the asymmetry between the buy and sell directions: in trending conditions, the counter-trend direction (for example, the buy-side depth during a downtrend) is usually significantly thinner than the with-trend direction, and it is precisely the counter-trend-side depth that determines the actual price impact when an informed trader builds a counter-trend position. When analyzing the liquidity constraints on price discovery, therefore, one should use directional depth rather than two-sided total depth as the benchmark.

13.3.3 Trading volume and the gravity advantage

The gravity advantage of perpetual futures is most directly reflected in trading volume. As Figure 13-11 shows, from 2020 to 2025 the trading volume of perpetual futures grew explosively. In 2025, global crypto derivatives trading volume reached $85.7 trillion; according to CryptoQuant's exchange-ratio data, the perpetual-to-spot volume ratio climbed to around 3.3, and at leading exchanges such as Binance this ratio at times reached as high as 5.1 [14].

The concentration of trading volume is not only the result of gravity but also a cause of its reinforcement. Liquidity exhibits network effects: the more informed traders choose perpetual futures, the more inclined market makers are to provide liquidity there; and

deeper liquidity in turn attracts still more traders, forming a positive feedback loop.

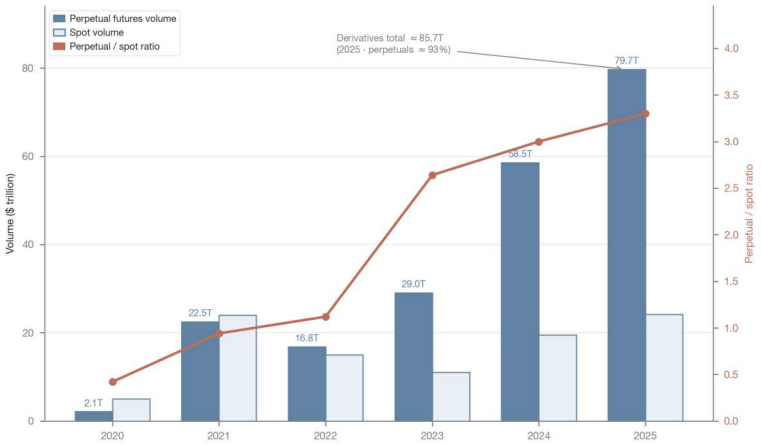


Figure 13-11. Annual trading volume growth of perpetual futures and spot markets (2020–2025)

Figure 13-11 shows that perpetual futures trading volume has grown exponentially, and its volume ratio relative to spot has climbed continuously, reflecting a structural shift of market liquidity and pricing power toward perpetual futures.

Figure 13-12 uses a pie chart to show how the total 2025 crypto derivatives trading volume of $85.7 trillion was allocated among perpetual futures, dated delivery futures, and options.

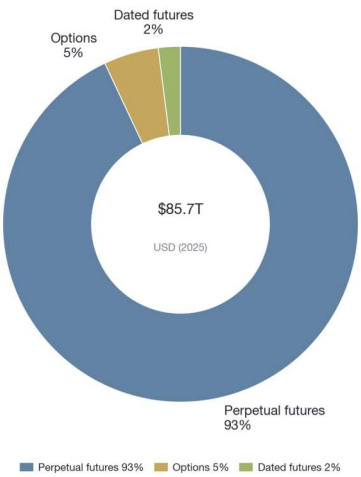


Figure 13-12. Product structure distribution of the crypto derivatives market in 2025

Figure 13-12 shows that perpetual futures accounted for about 93% of the derivatives market, with options and dated futures accounting for only a tiny fraction. This highly concentrated product structure indicates that the design features of no expiry and

continuous settlement confer a structural advantage on perpetual futures in trader preference. This landscape also means that the pricing mechanisms of perpetual futures (including the funding rate and the mark price) effectively serve as the price benchmark for the entire crypto derivatives ecosystem.

13.3.4 Funding rates and basis

Perpetual futures are called "perpetual" because they have no delivery date. To anchor the perpetual price to the spot index price, the market introduced the funding rate mechanism. This mechanism is not only a tool for price anchoring but also a highly valuable second-order information signal unique to perpetual futures markets.

When the perpetual price is above the spot price (a positive basis), the funding rate is positive and longs pay shorts; the reverse holds otherwise. As Figure 13-13 shows, this forms a negative feedback loop that incentivizes arbitrageurs to enter and close the price gap.

At the microstructure level, however, extreme values of the funding rate reveal the extreme tilt of market sentiment and the degree of leverage crowding. One widely discussed, representative observation holds that a nonlinear "U-shaped" relationship may exist between the funding rate and subsequent asset returns (see Figure 13-14). When the funding rate is at an extreme negative value (Q1, extremely crowded shorts), a short squeeze often follows and drives the price sharply higher; when the funding rate is at an extreme positive value (Q5, extremely crowded longs), a price plunge triggered by long liquidations often follows. The single-asset empirical evidence in the cited Presto Research [15] mainly finds that the contemporaneous funding rate has limited forward-looking predictive power for future returns; the quintile U-shaped pattern above is a representative illustration based on the logic of rate extremes, not a direct conclusion of that study. The funding rate itself has thus become an important information input for informed traders in constructing strategies.

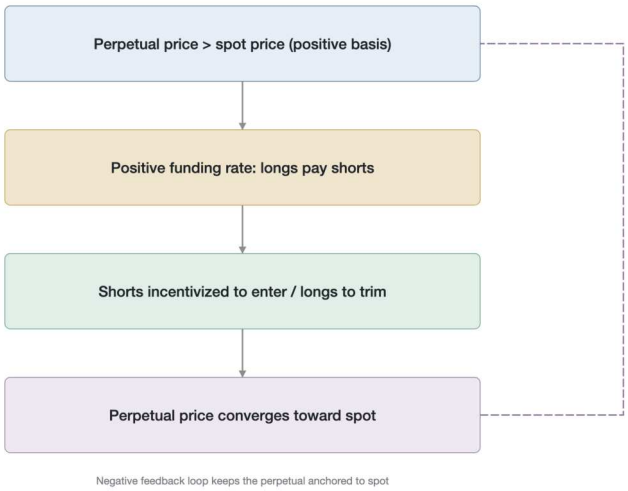


Figure 13-13. The price anchoring and negative feedback loop of the funding rate mechanism

Figure 13-13 shows that the funding rate not only drives the derivatives price back toward spot; its extreme values also constitute a second-order information signal measuring the market's leverage crowding and fragility.

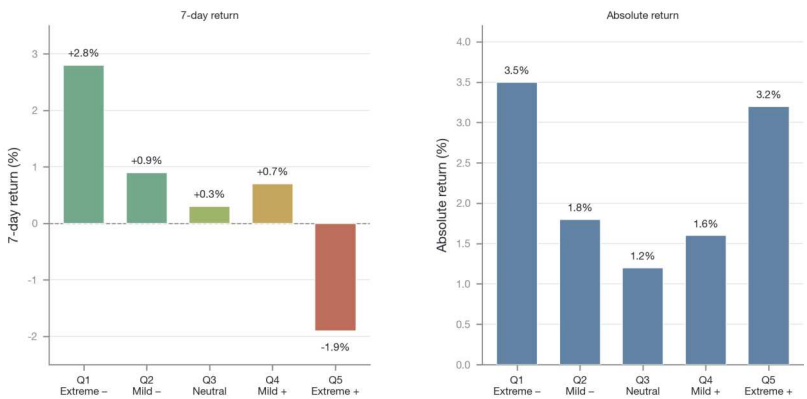


Figure 13-14. The nonlinear relationship between funding rate quintiles and subsequent 7-day returns (representative illustration: the quintile partition is a representative construction by the author based on the logic of rate extremes; the cited Presto Research [15] does not directly present this quintile conclusion)

Figure 13-14 illustrates, in representative terms, that both extreme negative rates and extreme positive rates correspond to higher absolute subsequent volatility, forming a pronounced U-shaped profile. This depiction suggests that when the funding rate deviates from its normal range, the market is accumulating directional risk, and the

subsequent mean reversion is often accompanied by violent volatility. The funding rate is therefore not only a mechanical tool for anchoring to spot but has also evolved into a second-order signal for measuring market fragility and forecasting short-term volatility.

The robustness of the U-shaped relationship above, however, requires several important qualifications. First, extreme funding rates tend to cluster during macro-regime shifts (such as an abrupt turn in Federal Reserve monetary policy, major geopolitical events, or regulatory shocks), so the predictive power of extreme rate values may derive in part from the mean reversion of a macro risk premium rather than from a pure crypto-market microstructure phenomenon. Stripping macro factors out of the rate signal is a necessary step for testing its independent predictive power, but the literature currently lacks such control analyses. Second, the funding-rate formulas used by different exchanges differ substantively (the specific settings of each exchange's interest-rate-minus-premium component, clamp boundaries, and final rate cap are given in Section 13.5.2), so the funding rates quoted by different exchanges at the same moment may differ significantly, and the sensitivity of the U-shaped relationship to the calculation methodology has not been systematically tested. Third, as the "funding rate hunter" strategy grows increasingly popular and crowded among quantitative funds and retail traders, the contrarian trading alpha based on extreme rates may be undergoing structural decay: as ever more traders try to exploit the same signal, its predictive power is weakened by strategy reflexivity.

13.3.5 The dynamics of the gravity model

Venue gravity is not static. An important corollary of the gravity model is state dependence: under different market states, the weights of the dimensions change, leading to a dynamic reconfiguration of the distribution of information gravity.

Using Hasbrouck's information share (IS) model to compute values, one can clearly observe that this dynamic reconfiguration exhibits several typical market states empirically. The IS estimates below are composite estimates based on a number of empirical studies; because different studies differ in sample period, data frequency, and choice of exchange set, they are presented as ranges. In the normal market state, CEX perpetual futures, by virtue of extremely high leverage and deep liquidity, hold about 50%–60% of the information share and are the core venue of price discovery. In the stressed market state (such as a macro shock), when sudden macro news (such as an unexpected Federal Reserve rate hike) is announced, informed traders urgently need the most efficient instrument to express information, and at this point the high-leverage, low-latency advantages of perpetual futures are further amplified, so their information share may climb to 60%–70%.

Analyzing the crisis state requires introducing a key typological distinction, because crises of different nature affect the leadership of price discovery in entirely different ways. The first type is the pure mechanical crisis, whose typical form is a liquidation cascade—a chain of forced liquidations triggered by technical factors (such as a single large account being liquidated, or insurance-fund depletion triggering auto-deleveraging (ADL)) with no input of major new information. In such a crisis, the perpetual futures market is flooded with market orders forcibly issued by the liquidation engine; these

orders carry no information about the asset's fundamentals but are pure mechanical noise. Market makers withdraw liquidity because inventory risk is too high, causing aggregation efficiency to collapse. At this point, the spot market, because it has no forced-liquidation mechanism, becomes by contrast a relatively pure information source, and its information share may overtake perpetual futures; based on a composite estimate from several studies, the spot market's IS share in this scenario can reach about 35%–45% or even higher, while the IS share of perpetual futures may fall to 30%–40%. The second type is the information-mechanical compound crisis, whose typical form is a major news event superimposed on a liquidation cascade—for example, in the Bybit hack the release of information about the security breach simultaneously triggered large-scale liquidations. In such a crisis, the market must process two entirely different kinds of price signal at once: an informational signal about the fundamental meaning of the event, and mechanical noise produced by liquidations. Because informed traders still prefer the derivatives market with the highest leverage efficiency to express their judgment about the event's fundamental meaning, the informational order flow in the leading derivatives market mixes with mechanical noise, but the information density of the former is sufficient to keep the derivatives market's IS share stable or even higher; based on a composite estimate from several studies, the leading derivatives market's IS share in this scenario can reach about 50%–60% or higher. This distinction explains a superficial paradox: whether the IS share of perpetual futures rises or falls in a crisis depends on the information structure of the crisis rather than on the magnitude of volatility alone.

State dependence indicates that price discovery leadership in crypto markets is not fixed at a single venue but switches dynamically between perpetual futures and the spot market according to the nature of the crisis.

Figure 13-15 compares the distribution of information share across trading venues under three market states: normal, stressed, and crisis.

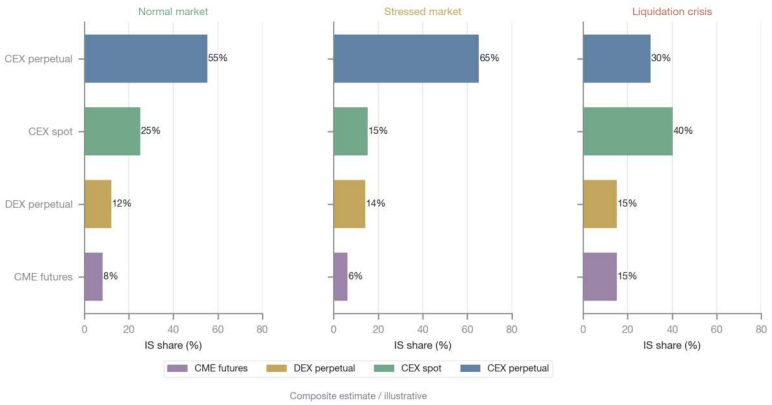


Figure 13-15. The distribution of information share across normal, stressed, and crisis market states (here "crisis" denotes a pure mechanical crisis / liquidation cascade; information shares are composite estimates / illustrative values)

Figure 13-15 shows that in normal and stressed markets, CEX perpetual futures dom-

inate the information share. In the crisis state, the distribution of price discovery leadership depends on the information structure of the crisis: in a pure mechanical crisis (liquidation cascade), the perpetual market is swamped by mechanical noise and the spot market passively takes over price discovery leadership; whereas in an information-mechanical compound crisis, informed traders' need to express the event's meaning keeps the leading derivatives market at the center of information.

13.3.6 The gravity divergence between CEX and DEX perpetual futures

Entering the 2024–2025 cycle, as "intent-centric" architectures and high-performance application chains (such as Hyperliquid) matured, decentralized perpetual futures began to pose a substantive challenge to the gravity of centralized exchanges.

As Figure 13-16 shows, by the end of 2025 the monthly trading volume of DEX perpetual futures had surpassed $1 trillion, and their share of the entire perpetual futures market had climbed to 22% [12]. However, similar to the trading-volume authenticity challenge facing CEXs (discussed in detail in Section 13.7.2), DEX volume likewise carries significant distortion risk. The fierce competition among Perp DEXs in 2025 spawned a large amount of trading activity of a non-price-discovery nature: airdrop farming drives traders to inflate volume in order to obtain token allocations, trading-mining mechanisms directly incentivize volume inflation through fee rebates, and bot farming manufactures fake activity on low-liquidity pairs through automated strategies. Although these activities contribute notional volume, they contain almost no price discovery information. The proportion of the 22% market share that genuinely represents informed traders expressing private information—that is, effective price discovery activity—may therefore be significantly lower than the notional figure implies.

Although CEXs (particularly Binance and OKX) still hold a significant advantage in absolute liquidity depth and the scale of institutional custody, DEXs hold an irreplaceable advantage along two gravity dimensions: "anonymity" and "censorship resistance." This differentiation in gravity dimensions has led to a stratification of informed traders: macro and institutional informed traders rely mainly on CEX perpetual futures, because they need extremely deep liquidity to absorb large orders and have strict compliance requirements. By contrast, on-chain-alpha and long-tail-asset informed traders increasingly favor DEX perpetual futures. For newly issued tokens or long-tail assets not yet listed on top CEXs, DEX perpetual futures are often the only leveraged shorting instrument, so for these specific assets DEX perpetual futures hold absolute pricing power.

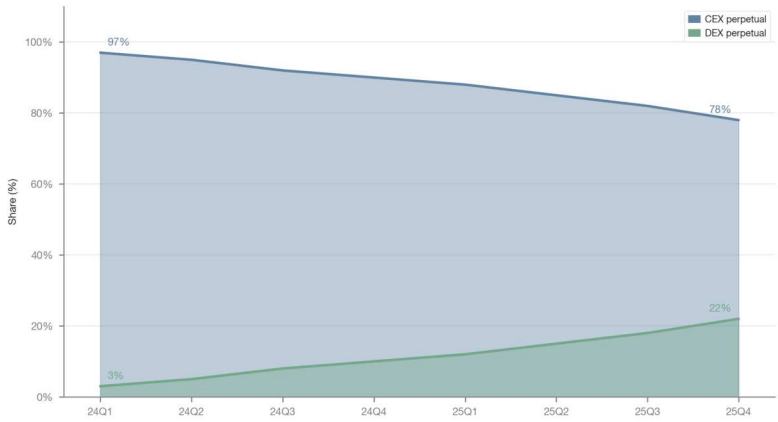


Figure 13-16. Comparison of CEX and DEX perpetual futures trading volume shares (2024–2025; shares measured on a notional-volume basis)

Figure 13-16 shows that the DEX perpetual futures market share rose steadily from 2024 to 2025, reflecting that high-performance application chains have begun to substantively divert CEX gravity in terms of liquidity and execution efficiency. This diversion has not occurred evenly: for mainstream assets such as BTC and ETH, CEXs still maintain an absolute liquidity advantage; but for long-tail assets and newly issued tokens, DEX perpetual futures have become the only leveraged shorting instrument and therefore hold pricing power for these specific assets. This differentiated distribution of gravity foreshadows a possible future price discovery landscape with a dual-track structure of "CEX dominance for mainstream assets, DEX dominance for long-tail assets."

Yet the "censorship resistance" gravity advantage of DEXs is itself facing a severe test against legal reality. The 2022 sanctioning of Tornado Cash by the U.S. Treasury's Office of Foreign Assets Control (OFAC) established for the first time the precedent that a government can add a smart contract address to a sanctions list; although this decision went back and forth under judicial review, its institutional signal has profoundly changed the compliance-risk expectations for decentralized protocols. For Perp DEXs, the key point is that "censorship resistance" does not equal legal immunity. Take Hyperliquid: the core of its high-performance architecture (a centralized sequencer) is both the technical basis for achieving sub-second finality and a natural chokepoint for sanctions compliance and a single point of failure for the system—regulators can achieve trade censorship by pressuring the entity operating the sequencer, as demonstrated by the emergency coordination among validators to shut down a market during the JELLY event of March 2025. More broadly, any DEX with an identifiable operating entity at the physical-infrastructure layer needs to draw a more careful distinction between "technical feasibility" and "legal enforceability" in its censorship-resistance claims.

13.4 The information flow topology

Following from the fragmented market structure described above, the formation of the "true price" is no longer an isolated process within a single market but a dynamic outcome of a multi-market game. This section presents the information flow topology, the third core theoretical contribution of this chapter (CC3), as a theoretical framework that systematically reveals the cross-market transmission mechanism of price discovery and explores the roles and standing of different types of markets in price discovery.

13.4.1 Constructing the topology

To describe systematically the price discovery process in a fragmented market, we construct the information flow topology. In this directed weighted graph $\mathcal{G} = (V, E, W)$, the set of nodes V represents the trading venues (such as Binance perpetual futures, Coinbase spot, Hyperliquid, and so on), the set of directed edges E represents the direction in which price signals are transmitted, and the weight function $W : E \to \mathbb{R}^+$ reflects the strength of information transmission (such as the elasticity of price-signal transmission per unit time). Here $\mathbb{R}^+$ takes strictly positive values—that is, if an edge exists, its weight must be greater than zero (the absence of a transmission relationship corresponds to the absence of an edge, not to a weight of zero).

Within this topological structure, information flow is not evenly distributed. High-liquidity, low-transaction-cost derivatives markets typically serve as "information source" nodes, reflecting new market information first; while spot markets with weaker liquidity or higher transaction costs serve as "receiving" nodes, passively following price changes. Arbitrageurs play the role of "connectors" in this network, closing price deviations through cross-market arbitrage strategies and thereby completing the physical transmission of information flow. When arbitrage friction (transaction fees, network latency, fund-transfer times, and so on) is low, the topology is strongly connected and prices tend to be consistent; when friction rises, connectivity declines and persistent price differences appear among nodes.

Figure 13-17 constructs an information flow topology of crypto asset markets centered on Binance perpetual futures.

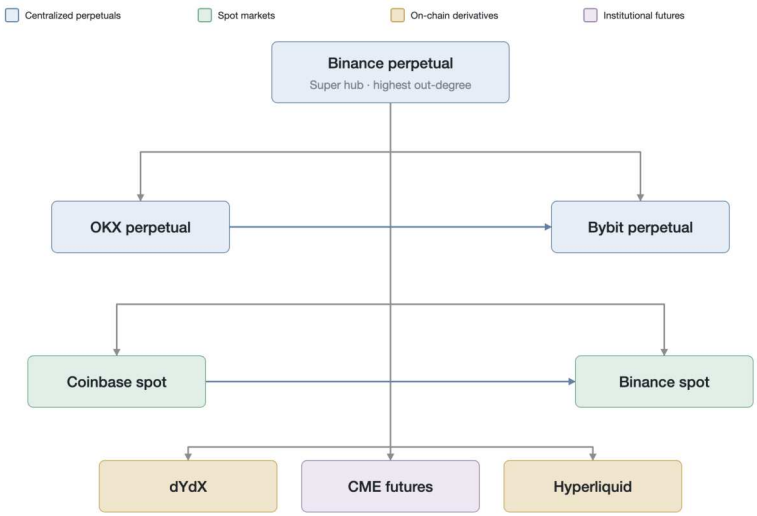


Figure 13-17. The information flow topology of crypto asset markets

Figure 13-17 shows the information transmission paths centered on Binance perpetual futures; directionally asymmetric information flows exist among CEX perpetuals, CEX spot, DEXs, and traditional financial markets, and the weights of the directed edges reflect the relative strength of transmission.

The topological structure depicted above reveals a core fact: in today's crypto market, Binance perpetual futures are the "super-hub" of the information flow network, with an out-degree far higher than that of any other node, meaning that the price signals produced by this market radiate unidirectionally to almost all other markets. At the same time, the index price and oracles, though they generate no trades themselves, constrain the price boundaries of the entire network in reverse by serving as the benchmark for liquidation and funding rate calculations.

13.4.2 Measurement tools

In empirical finance, the standard tools for measuring the contribution of different markets to price discovery comprise two classic models.

The Hasbrouck information share model is based on the residual covariance matrix of a vector error correction model and measures the proportion of the efficient price variance contributed by a given market [4]. Its core logic is that the market leading price discovery usually has higher volatility innovations—that is, that market reacts most sharply and rapidly to new information. The mathematical expression of the IS model can be summarized as:

$$IS_j = \frac{\psi_j^2 \sigma_j^2}{\sum_{k=1}^{n} \psi_k^2 \sigma_k^2}$$

where ψ_j is the weight of market j in the common trend (i.e., the common-factor loading derived from the moving-average representation of the vector error correction model),

and σ_j^2 is the variance of that market's innovation term. Note that here σ_j denotes the standard deviation of market j's innovation term, with a different meaning from the volatilities $\sigma_{\text{mech}}/\sigma_{\text{info}}$ used earlier. The expression above is the form after the innovation covariance has been diagonalized (orthogonalized); the general case requires a Cholesky decomposition, and the IS model is fairly sensitive to the Cholesky orthogonalization order of the error covariance matrix, with results depending on the ordering, so in practice one usually needs to compute upper and lower bounds to determine the final share.

The Gonzalo-Granger component share model, by contrast, focuses on the error-correction speed of each market's price toward the common trend [16]. The higher a market's CS share, the smaller its responsibility for correcting price deviations—that is, other markets must "move toward" it. Unlike the IS model, the CS model measures a market's capacity to lead permanent price shocks rather than its contribution to short-term volatility.

Table 13-3 compares the features of Hasbrouck IS, Gonzalo-Granger CS, and their joint use.

Measurement method	Core principle	Advantage	Limitation
Hasbrouck IS	Measures each market's contribution to the efficient price variance	Directly quantifies the proportion of information contribution	Sensitive to the Cholesky decomposition order; requires computing upper and lower bounds
Gonzalo-Granger CS	Measures each market's error-correction speed	Does not depend on orthogonalization; yields a unique result	Reflects only long-term adjustment, ignoring short-term information shocks
IS and CS combined	Integrates short-term volatility contribution and long-term adjustment capacity	Provides a more comprehensive picture of price discovery	Higher computational complexity; requires high-frequency synchronized data

Table 13-3. Comparison of price discovery measurement methods: Hasbrouck IS, Gonzalo-Granger CS, and their joint use (Data source: compiled by the author)

When assessing the price discovery efficiency of crypto markets, researchers usually use these two measurement tools in combination to obtain more comprehensive market-microstructure insight. Putniņš (2013) further proposed the information leadership share, which eliminates the interference of noise in the measurement by standardizing the ratio of IS to CS [17].

13.4.3 The empirical landscape

A large body of empirical research shows that in the current crypto asset price discovery environment, the perpetual futures markets of centralized exchanges hold a dominant position. Cosenza and Stalder (2024), using high-frequency data to analyze the price discovery process for Bitcoin in a fragmented market, found that unregulated exchanges

(especially Binance) are the primary source of price discovery, and that their perpetual futures and spot markets together contribute more than 50% of the information share [18]. The study by Entrop, Frijns, and Seruset (2020) further confirms that during periods of rising volatility and increasing market uncertainty, the contribution of the futures/perpetual market to price discovery strengthens significantly [19]. Dimpfl and Peter's (2021) cross-exchange analysis reveals a historic shift of price discovery leadership from Bitfinex to Binance [20]. The dominance of unregulated venues in price discovery itself faces regulatory-event risk. The $4.3 billion settlement Binance reached with the U.S. Department of Justice and the CFTC in November 2023 shows that enforcement actions targeting dominant venues can lead to a sudden reconfiguration of the information flow topology. Moreover, after the approval of the spot Bitcoin exchange-traded fund (ETF) in January 2024, the IS share of CME and regulated markets has grown significantly, and the "unregulated-venue-dominance" pattern is being changed by institutional forces.

Figure 13-18 compares the Hasbrouck IS and Gonzalo-Granger CS shares of the major trading venues.

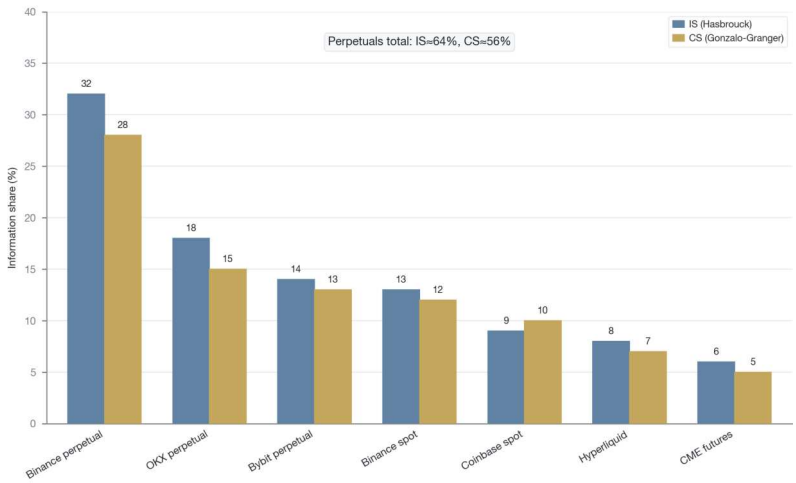


Figure 13-18. Comparison of Hasbrouck IS and Gonzalo-Granger CS shares across major trading venues (synthesizing multiple studies [18][19]; CME reflects a pre-2024-spot-ETF sample)

Figure 13-18 synthesizes the results of multiple empirical studies, showing that the three major CEX perpetual futures (Binance, OKX, Bybit) together contribute about 64% of the IS share and 56% of the CS share, confirming the dominant position of the derivatives market in price discovery; after including perpetual DEXs such as Hyperliquid, the combined IS of all perpetual futures is about 72%.

As the figure above shows, Binance perpetual futures are the "price leader," with an IS share of about 32% and a CS share of about 28%, both significantly higher than those of other venues. By comparison, traditional financial markets (such as CME futures) contribute only about 6%, which is related to factors such as their weekend closures, higher barriers to trading, and contract-size limits. However, CME's IS share is not proportional to its systemic importance: CME is the main channel for institutions to

allocate crypto exposure in a compliant manner, its settlement price is used in the net-asset-value calculation of various ETFs, and the gap effect produced by its trading-session closes and opens can itself be an information event. This 6% IS figure comes from research samples predating the approval of the spot Bitcoin ETF in 2024, since which the landscape has undergone a structural change: CME's institutional trading volume and information share have grown significantly, and the weight of regulated markets in price discovery is rising systematically.

High-performance DEXs, represented by Hyperliquid, are rising rapidly. Although their IS share is currently about 8%, their low latency, on-chain transparency, and permissionless access are attracting a growing number of market makers and arbitrageurs. As DEX infrastructure continues to improve, their weight in the price discovery topology is likely to rise further.

13.4.4 Three types of leaders

Based on the analysis of the information flow topology, we can divide the trading venues in crypto markets into three distinct types of price discovery leaders (Table 13-4). This classification system goes beyond the traditional "spot versus futures" binary and provides a more granular portrait of market roles.

Leader type	Core characteristics	Typical examples	Operating mechanism	IS contribution
Information-efficiency type	Extremely low trading friction, high leverage, dense high-frequency trading	Binance perpetuals, OKX perpetuals	Absorbs incremental information such as macro data and industry news the fastest, producing the initial price jump	~50%–65%
Arbitrage-hub type	Deep spot markets with fiat on/off-ramps	Binance spot, Coinbase spot	Serves as the benchmark leg for cross-instrument (spot-perpetual) and cross-exchange arbitrage, anchoring excessive deviations of derivatives	~20%–30%
Price-benchmark type	Indices aggregating multiple data sources, widely referenced	Oracles (e.g., Chainlink), CEX index prices	Generates no trades itself, but by serving as the benchmark for liquidation and funding rate calculations, constrains the price boundaries of trading markets in reverse	Indirect influence

Table 13-4. The three types of price discovery leaders in crypto markets and their operating mechanisms (Data source: compiled by the author)

These three types of leaders form a mutually constraining closed loop in the topology: information-efficiency markets produce price innovations, arbitrage-hub markets validate and transmit prices, and price-benchmark markets ensure the long-term anchoring of the entire system. This triangular checks-and-balances relationship functions well under normal market conditions, but in extreme conditions it can rupture as arbitrage friction rises sharply—precisely the issue the next subsection discusses.

13.4.5 Arbitrage friction and transmission delay

Information transmission in the topology is not instantaneous but is limited by arbitrage friction. Under normal market conditions, the arbitrage cost between CEXs is about 8–12 basis points and the price-convergence time is about 0.5–2 seconds; the arbitrage cost between a CEX and a DEX is higher, about 22–26 bps, with a convergence time of about 8–15 seconds. These cost estimates represent explicit transaction costs under ideal conditions. The "8 to 12 basis points" range is achievable only for institutional

arbitrageurs who have pre-positioned capital across multiple exchanges, and this implies a significant capital-locking cost. The actual full cost of arbitrage also includes cross-exchange withdrawal delays (10 minutes to several hours under normal conditions, and possibly suspended by exchanges under extreme conditions), basis risk during execution, and the opportunity cost of maintaining margin at multiple exchanges simultaneously. The effective full cost of arbitrage may therefore be several times the explicit cost. When the market encounters an extreme event, these frictions are amplified sharply, causing significant delays and discontinuities in price transmission.

The Bybit hack of February 21, 2025 provides a typical observational slice [1]. In that event, Bybit's Ethereum cold wallet was attacked by the North Korean hacking group Lazarus Group, with losses of about $1.4–1.5 billion in ETH and related derivative assets. According to Kaiko Research's analysis, the market did not begin to react on a large scale until about an hour after the hack occurred (i.e., after the CEO publicly confirmed it), and this delay itself reveals the persistence of information asymmetry: although blockchain data are public, the signal of a major event cannot be identified by the market instantly [3].

Figure 13-19 tracks the speed of price reactions across trading venues and the change in cross-market price differences during the Bybit hack.

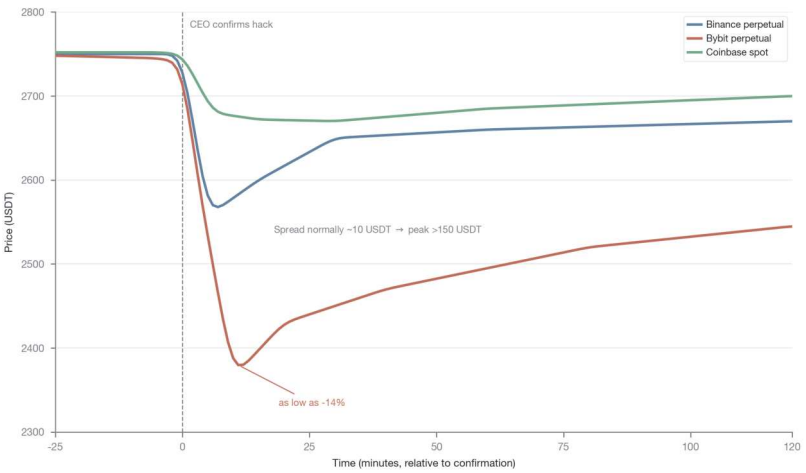


Figure 13-19. Cross-market price reactions and spread widening during the Bybit hack (price paths are illustrative; key values per Kaiko Research [3])

Figure 13-19 shows that after the event was confirmed, Binance perpetuals, Bybit perpetuals, and Coinbase spot differed significantly in reaction speed and depth of decline; the figure also marks the price difference between Binance perpetuals and Coinbase spot widening from the normal level of about 10 USDT to over 150 USDT at one point.

The market reaction after the event's confirmation exhibited three levels of characteristics. In terms of reaction speed, Binance perpetual futures reacted first, completing a repricing of about 7% within 5 minutes, whereas the Coinbase spot market exhibited a significant delay of about 8 seconds with a smaller decline (about 2%–3%), a strat-

ification fully consistent with the prediction of the information flow topology. At the same time, Bybit perpetual futures exhibited an excess decline far exceeding that of other markets (reaching about 14% at one point), which included the market's pricing of the credit risk to Bybit's platform solvency; this credit-risk premium is pure local noise and should not be regarded as information about ETH fundamentals. On the arbitrage dimension, a cross-exchange price difference of over 150 USDT would normally be closed quickly by arbitrageurs, but at the moment of crisis, as large numbers of market makers withdrew liquidity (Bybit's 1% market depth fell from $68 million to $28 million within hours of the event, a decline of 59%), and with restricted fund transfers and excessive one-sided risk, arbitrage friction tended toward infinity and the information-transmission chain ruptured [3].

According to Kaiko Research's statistics, in the first hour after the CEO confirmed the hack, the cumulative volume difference for the ETH-USDT pair alone reached negative $52 million, including a single large sell order of about $500,000 [3].

13.4.6 The state dependence of the topology

The information flow topology is not a static structure but is highly state-dependent. As the market environment evolves from a normal state to a stressed state and then to a crisis state, the node weights and connectivity of the topology undergo a systematic dynamic reconfiguration.

Figure 13-20 shows the dynamic evolution of the information flow topology from a normal state to a crisis state.



Figure 13-20. The evolution of the information flow topology from normal to stressed to crisis states (here "crisis" denotes an information-mechanical compound crisis; venue IS values are illustrative)

Figure 13-20 shows that as the market evolves from a normal state to a crisis state, the weight of price discovery gradually concentrates in the leading platforms, and the IS share of peripheral nodes shrinks dramatically. This concentration dynamic reminds us that static IS/CS measurements must be interpreted together with the market state;

otherwise one would underestimate the distortion of price signals caused by the structural collapse of liquidity under extreme conditions.

In the normal state, market volatility is calm and liquidity is ample. The topology exhibits a multi-center feature, with CEX spot and leading DEXs able to maintain a certain price discovery share (for example, Coinbase spot at about 9% IS and Hyperliquid at about 8%), arbitrage channels flowing freely, and cross-market price differences kept within a normal range.

In the stressed state, macro data releases or local negative news drive volatility higher. Market makers contract their quote widths and lower their quoting depth, and liquidity concentrates in Binance perpetual futures, which have the lowest trading cost. At this point, the IS share of Binance perpetuals rises from about 32% in the normal state to about 42%, while the shares of DEXs and second-tier CEXs decline accordingly. The "centralization" of the topology deepens.

In the crisis state, the mode of reconfiguration of the topology depends on the information structure of the crisis, an analysis consistent with the typological framework of Section 13.3.5. Following the crisis typology of Section 13.3.5, in an information-mechanical compound crisis (a major news event superimposed on a liquidation cascade, such as the Bybit hack), the cross-market arbitrage mechanism fails on a large scale, and the connectivity of the topology declines sharply; yet price discovery instead becomes "highly centralized": informed traders still prefer the leading derivatives market with the highest leverage efficiency to express their judgment about the event's fundamental meaning, the leading derivatives node becomes the most important information source, and other nodes are largely reduced to "price takers" or fall into liquidity exhaustion. In a pure mechanical crisis (a liquidation cascade triggered by high leverage, with no input of major new information), by contrast, the price signal of perpetual futures is polluted by large amounts of mechanical noise, its IS share instead declines, and leadership temporarily shifts to the spot market, which is not affected by the forced-liquidation mechanism (see the analysis in Section 13.3.5). This distinction means that the degradation of the topology in a crisis state does not follow a single pattern: an information-mechanical compound crisis causes the topology to concentrate extremely on the leading derivatives node, whereas a pure mechanical crisis may cause a temporary shift of information-flow leadership from the derivatives market to the spot market.

Beyond the three market states above, regulatory shock constitutes a fourth type of trigger for topological reconfiguration. When regulators take enforcement action against a dominant venue (such as restricting a particular exchange's operations in a given jurisdiction), the information flow topology may undergo an abrupt change within days, traders are forced to migrate to alternative venues, and lead-lag relationships may fully reverse in the short term. This regulation-driven topological reconfiguration is entirely different in nature from the market-volatility-driven state transitions: the latter are endogenous and continuous and usually recover as volatility subsides; the former is exogenous and discrete, and its impact may be a permanent structural change rather than a temporary deviation.

This state dependence indicates that when assessing the price discovery efficiency of

crypto assets, one must consider the macro and micro environment the market is in. Static IS/CS measurements often mask the risk of a structural collapse of liquidity under extreme conditions. For regulators and risk managers, what matters is not how high price discovery efficiency is in the normal state but how severe the degradation of the topology is in the crisis state, because it is precisely at these moments that the distortion of price signals causes the greatest economic losses.

13.5 The three pathways of mechanical noise

In the previous section, we established the dominant position of perpetual futures in price discovery. This leadership, however, is not without cost. The distinctive microstructure design of perpetual futures markets (high leverage, automatic liquidation, and the funding rate mechanism), while raising capital efficiency and the speed of information absorption, also endogenously produces a non-information-driven price volatility—that is, mechanical noise.

This section presents the fourth core theoretical contribution of this chapter (CC4): perpetual futures are not merely the engine of price discovery but also a source of disturbance to the price system of the entire crypto market. Through three specific pathways—liquidation cascades, funding rate extremes, and the mark price EMA delay—it analyzes how this mechanical noise is generated, amplified, and transmitted outward.

13.5.1 Liquidation-driven price displacement

Crypto derivatives markets widely adopt deterministic automatic liquidation mechanisms, but their concrete implementation is far more complex than the simplified narrative of "forced liquidation at market." When an account's maintenance margin ratio falls below a specific threshold, the risk engines of leading exchanges (such as Binance and OKX) do not dump the entire position at once but adopt a tiered-liquidation strategy: they divide the position into several tiers according to its size and close it batch by batch starting from the outermost tier, with each batch executed at a different price level, in the hope of reducing the market impact while controlling risk. In terms of execution, the liquidation engine usually issues aggressive limit orders rather than pure market orders—that is, it completes the liquidation with limit orders posted near the current best opposing price, rather than unconditionally accepting any price. Moreover, when counterparty liquidity is severely insufficient during liquidation and the insurance fund's consumption reaches a preset threshold, some positions are not executed on the open market but enter the auto-deleveraging (ADL) process, in which the system directly designates counterparties among profitable traders to complete the closeout. Although tiered liquidation and the limit-order mechanism can effectively mitigate market impact in a normal volatility environment, in situations of extreme liquidity exhaustion (order book depth plunging to a few percent of the normal level), aggressive limit orders cannot obtain sufficient fills against thin quotes, the successive batches of tiered liquidation are triggered in quick succession within an extremely short time, and their actual effect degrades to a behavior pattern close to market liquidation. This mechanism design, while protecting the exchange from bad-debt risk, therefore

still creates enormous "mechanical selling pressure" under extreme conditions, driven not by any new information about the asset's fundamentals but by the deterministic execution of system rules.

When the market encounters an initial downward shock, the first batch of long liquidations triggered produces additional selling, driving the price down further and thereby triggering the liquidation of more highly leveraged positions. This positive feedback loop is called a liquidation cascade. During the cascade, the violent price decline departs entirely from the guidance of fundamental information and is driven purely by system rules and liquidity exhaustion. This liquidation-driven, non-informational volatility—that is, the "mechanical volatility" defined in Section 11.8.3 of Chapter 11—is denoted in this chapter by the symbol σ_{mech} to distinguish it from information-driven volatility σ_{info}; its full formalization (the liquidation reflexivity equation, the divergence criterion $G > 1$, and the four-stage dynamics) appears in Chapter 11, Sections 11.3 and 11.5. Here G is the liquidation gain coefficient defined in Chapter 11 and is unrelated to this chapter's topology $\mathcal{G}$ or venue information gravity G_i.

The crypto market flash crash of October 10–11, 2025 is an extreme manifestation of this mechanism [21]. Over the course of a single weekend, forced liquidations of over $19 billion occurred market-wide (on CoinGlass's basis, approximately $19.1–19.4 billion), a scale roughly 19 times that of the March 12, 2020 event (see Section 13.7.1). The average intraday maximum drawdown of about 1,600 tokens approached 60%, and hundreds of small-cap tokens lost 70%–90% of their quoted value on at least one trading venue.

Figure 13-21 simulates the dynamic evolution of price, visible liquidation notional, and open interest during a typical liquidation cascade event.

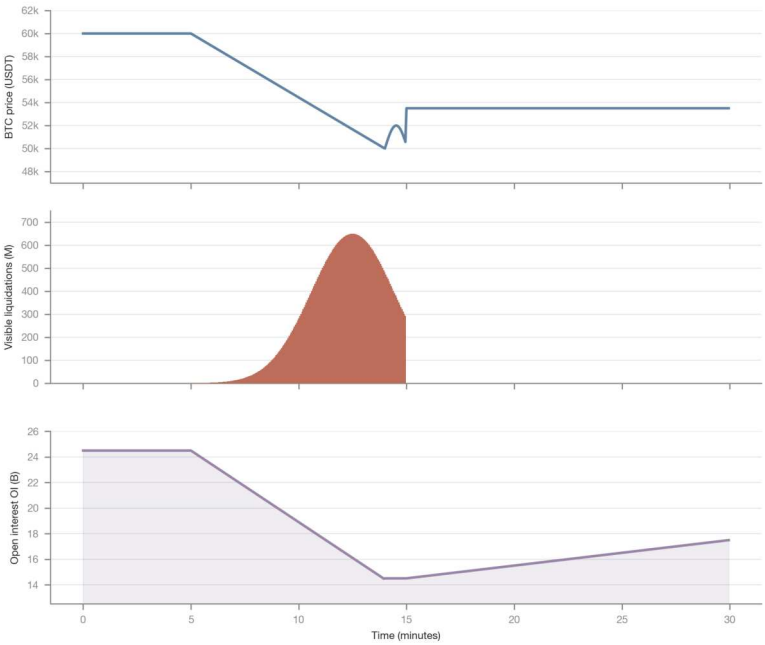


Figure 13-21. The dynamic evolution of price, visible liquidation notional, and open interest during a liquidation cascade (conceptual illustration: simulated data, not empirical)

Figure 13-21 simulates a typical liquidation cascade: after the initial trigger, the price probes downward nonlinearly, the visible liquidation notional grows exponentially, and open interest shows a cliff-like drop of about 40%.

In the "peak pressure" phase of the liquidation cascade, the depth of the price decline far exceeds the reasonable range of the initial information shock. This overshoot caused by forced liquidation is the most typical mechanical noise. The destructive power of a liquidation cascade stems from three mutually reinforcing factors. At the level of deterministic execution, the automatic liquidation engine executes unconditionally according to preset rules, without regard to market liquidity conditions or price reasonableness—as Greco noted in his analysis of the October 2025 event, "deterministic automation is predictable, and predictability is exploitable" [21]. On the liquidity dimension, during the cascade market makers withdraw their quotes as risk rises sharply, causing order book depth to plunge, so that the liquidation sell orders face a market whose liquidity is shrinking sharply, and every forced liquidation causes a larger price impact. Cross-margin contagion further amplifies the crisis: under a unified account system, the depreciation of a single collateral asset can trigger the simultaneous liquidation of multiple unrelated positions—as in the October 2025 event, when the local liquidity exhaustion of Ethena's USDe on Binance's internal order book (falling to about $0.65 at one point, a discount of about 35%) triggered chain liquidations across all products through the cross-margin mechanism of the unified account [21]. This discount was mainly a pricing distortion of Binance's single low-liquidity USDe pair—over the same period the deviation on Curve was under 100 basis points, Bybit was only about $0.92, and the Ethena protocol maintained over-collateralization and normal redemptions, so

the 35% does not reflect USDe's true solvency risk (see the related analysis in Chapter 11). Unified margin does not alter the fundamental correlations among assets in a normal market (the correlation among BTC, ETH, and altcoins is usually between 0.3 and 0.6), but under liquidation contagion it creates an endogenous tail-correlation jump: the depreciation of a single collateral asset causes unrelated positions to be liquidated simultaneously, so that the realized correlation can spike to nearly 1.0. This "tail-correlation regime shift" poses a fundamental challenge to models such as Value at Risk (VaR) that rely on historical correlations for risk measurement. The "margin spiral" framework proposed by Brunnermeier and Pedersen (2009) provides a theoretical basis for understanding this endogenous correlation mechanism [22].

At the tail of the liquidation cascade, the auto-deleveraging (ADL) mechanism constitutes an independent and far-reaching noise-transmission pathway. The trigger condition for ADL is usually tied directly to the state of the exchange's insurance-fund consumption: when the liquidation engine cannot complete a closeout on the open market at the bankruptcy price or better, and the insurance fund's balance falls below a preset threshold, the system initiates the ADL process, no longer relying on open-order-book matching but instead forcibly selecting counterparties among profitable traders to take on the liquidated position. The selection of counterparties follows a specific priority-ranking logic: most leading exchanges (such as Binance and Bybit) use the product of profit rate and leverage multiple as the ranking metric—that is, the accounts with the richest profits and the most aggressive leverage use are selected first. The original intent of this design is to prioritize reducing the positions with the most concentrated risk in the system, but it triggers a highly destructive chain reaction. The typical business model of a market maker is to hold hedged positions in multiple markets simultaneously—for example, going short in the perpetual futures market to hedge a long exposure in the spot market. When the ADL mechanism forcibly liquidates their profitable short leg in perpetual futures, the originally risk-neutral portfolio instantly becomes a one-directional naked exposure. Facing this sudden risk exposure, the market maker's rational response is to immediately withdraw all remaining quotes from the market to avoid further losses in an extreme-volatility environment. This withdrawal manifests at the micro level as a further collapse of order book depth, and at the macro level as a chain exit of liquidity supply: the retreat of one market maker intensifies price volatility, which prompts more market makers to trigger their own risk-control thresholds and exit one after another. ADL is therefore not only the last line of defense of the liquidation cascade; it is itself a noise amplifier: by destroying market makers' hedge structures, ADL turns a position-level risk event into a market-wide liquidity crisis.

The analysis of the liquidation cascade above focuses on its mechanical layer: insufficient margin triggers the liquidation engine, liquidation orders consume liquidity, and the price decline triggers more liquidations—each link of this positive feedback loop is the automatic execution of deterministic rules. In reality, however, the liquidation cascade is also overlaid with an equally important, and possibly even more forceful, behavioral amplification layer. Data platforms such as CoinGlass and Hyblock Capital broadcast market-wide liquidation data in real time, and the token, direction, and amount of every large forced liquidation are pushed to traders' screens and social-media feeds

within seconds of its occurrence. This unprecedented transparency of the liquidation process creates a reflexive behavioral feedback mechanism: when traders observe that large-scale liquidations are occurring, they close positions preemptively to avoid risk on the expectation that "the cascade is not yet over," or actively build short positions to profit from the expected further decline—behavior that may be termed "liquidation front-running." Because the dense liquidation price levels displayed by liquidation heatmaps provide market participants with common knowledge about the price levels at which the next wave of forced liquidations may be triggered, traders' expectations and actions tend to become self-fulfilling: when enough participants make the same directional trading decision based on the same information, this itself generates pressure that moves the price toward the liquidation-dense zone. From the perspective of price discovery, the existence of this behavioral layer means that the total price displacement in a liquidation cascade cannot simply be classified as pure mechanical noise σ_{mech}. The mechanical component (the forced liquidations executed by the liquidation engine according to rules) indeed carries no information about the asset's fundamentals and belongs to the mechanical component defined in Section 13.2.2, and it should revert completely after the price impact fades. The behavioral-expectation component, however, is more complex: traders' active responses to liquidation data may partly reflect their genuine assessment of the market's fragility structure (leverage distribution, liquidity depth, insurance-fund adequacy), and this assessment itself carries valid information about the state of the market's systemic risk. To treat all price displacement during a liquidation cascade as noise to be filtered out may therefore lead to an underestimation of the market's actual risk level. A more careful analytical framework should decompose the price change during a cascade into a mechanical component and a behavioral-expectation component, and recognize the ambiguity of the latter's informational nature.

13.5.2 The distortionary effects of funding rate extremes

The mechanism by which perpetual futures anchor to the spot price through the funding rate has already been developed in Section 13.3.4. In one-directional conditions, however, the funding rate often moves toward extremes, thereby producing serious distortions in traders' position behavior and in asset prices [23].

In an extremely bullish market, crowded long sentiment drives the funding rate to surge. Taking the standard mechanism of settlement every 8 hours as an example, when a single rate reaches 0.1%, the annualized cost of holding a long position exceeds 100%. Three mechanism details must be noted when interpreting funding rate extremes. First, most exchanges set a cap on the final rate (such as Binance's $\pm 0.75\%$ per 8 hours) and set an additional clamp boundary on the "interest-rate-minus-premium-index" difference (such as Binance's $\pm 0.05\%$); the 0.1% in the example is far from touching that cap. Second, the standard 8-hour settlement frequency is diverging: Hyperliquid already settles hourly, and some protocols are experimenting with continuous funding mechanisms; a higher settlement frequency lowers the absolute value of any single rate but raises the temporal resolution of the rate signal. Third, the actual rate consists of two parts—an interest-rate component (fixed or semi-fixed) and a premium index (a dynamic component reflecting the basis between perpetual futures and spot); dis-

cussing only the extreme total rate without decomposing its constituents omits the independent information content of the two components. This high holding cost produces two effects: first, it forces some capital-inefficient longs to close out, generating selling pressure unrelated to fundamentals; second, it attracts arbitrageurs to buy in the spot market and go short in the contract market (i.e., "spot-perpetual" basis arbitrage), artificially increasing spot buying and contract selling.

This funding rate extremization recurs in historical data. Figure 13-22 selects data from a typical bull-market phase of BTC perpetual futures, overlaying the 8-hour change sequence of the funding rate with the price path and the annualized holding cost.

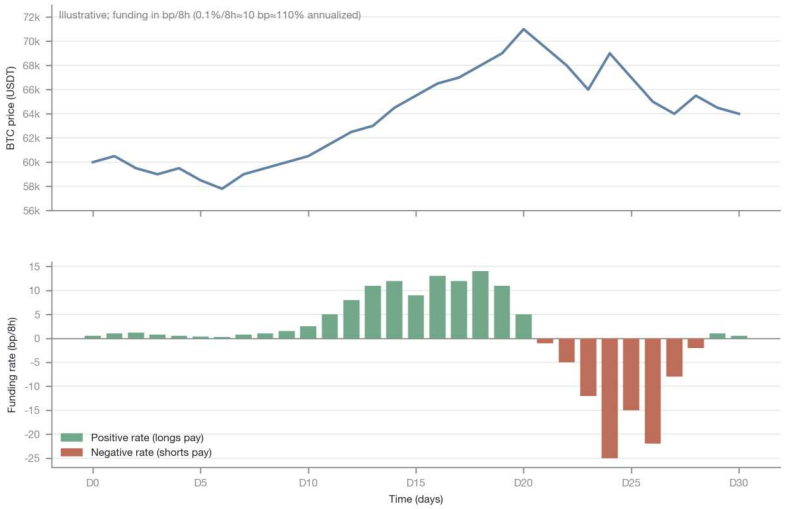


Figure 13-22. The distortionary effect of funding rate extremes on the BTC price and holding cost (representative illustration; rates measured in basis points per 8 hours, 0.1% per 8 hours ≈ 10 basis points ≈ approximately 110% annualized)

Figure 13-22 clearly shows the two distortionary effects of funding rate extremization. In the crowded-long phase (the high-rate interval marked in the figure), a single 8-hour rate of 0.1% translates into an annualized cost of over 100%, and this staggering holding cost causes all participants except the highest-conviction longs to bleed continuously—market makers and capital-inefficient trend-followers are forced to close out and exit, and the resulting selling pressure is driven entirely by the rate mechanism and is unrelated to BTC's fundamental valuation. At the same time, the extreme positive rate attracts large numbers of "spot-perpetual" basis arbitrageurs to enter: they buy BTC in the spot market and go short in the perpetual futures market to collect the high funding rate subsidy, and this purely arbitrage-motivated trading further artificially alters the supply-and-demand structure of the spot and contract markets.

The reverse extreme is equally destructive. In a panic-selling phase, the funding rate can turn deeply negative (for example, on February 28, 2026, the annualized BTC funding rate briefly fell to about −6% (on an annualized/aggregate basis, not a value within a single-settlement clamp boundary)), meaning that shorts must pay a high fee to longs [24]. Such negative rates attract contrarian arbitrageurs to build long positions,

but their motive is not a bullish view on fundamentals but rather to collect the rate subsidy. Once the rate returns to normal, these "rate hunters" close out en masse, generating price volatility unrelated to information.

The mechanical noise produced by funding rate extremization has an important feature: 8-hour periodicity. Because the funding rate at leading exchanges is settled every 8 hours (00:00, 08:00, and 16:00 UTC), before and after the settlement moment, traders' position-adjustment behavior produces regular price disturbances. This periodic noise is clearly visible in high-frequency data and has been widely exploited by quantitative trading firms. The strategy faces significant execution constraints: the actual trading window is extremely narrow (usually effective only within 1 to 5 minutes before and after settlement), broad awareness of the strategy has led to high crowding, and the mean-reversion positions built during periods of extreme rates face directional tail risk, because an extreme rate is sometimes a signal of a persistent trend rather than a brief deviation.

13.5.3 The EMA delay of the mark price

To prevent market manipulation and price wicks from triggering unnecessary liquidations, the vast majority of CEXs adopt a mark price mechanism [25]. The mark price is usually calculated based on the following formula:

$$P_{\text{mark}} = P_{\text{index}} + \text{EMA}(\text{basis})$$

where P_{index} is the weighted index price across multiple spot exchanges, and EMA(basis) is the exponential moving average of the difference between the perpetual futures' last traded price and the index price; here basis is defined as the perpetual futures' last traded price minus the index price (a positive value indicates a premium). The decay factor of the EMA determines the length of the smoothing window: the longer the window, the stronger the smoothing effect, but also the greater the delay. This formula is a conceptual, simplified expression, and the actual implementations of different exchanges differ significantly (and at the contract level are all fixed-point arithmetic, involving rounding to the tick). Binance uses a three-value median method, that is, $P_{\text{mark}} = \text{Median}(P_1, P_2, P_3)$, where P_1 is the last traded price, $P_2 = P_{\text{index}} + \text{moving-average basis}$, and $P_3 = P_{\text{index}}$; Bybit uses $P_{\text{mark}} = P_{\text{index}} \times (1 + \text{decayed EMA basis rate})$. The EMA window parameters also differ significantly; for example, Binance uses a simple moving average of the 30-second basis for Price2 (1 minute before September 2025) rather than an exponentially weighted one. These implementation differences directly affect the specific path and magnitude of the EMA-delay noise, causing different exchanges to produce different degrees of mark-price deviation in the same market event.

Although EMA smoothing effectively filters out instantaneous high-frequency noise, it necessarily introduces a phase delay mathematically. When the market moves rapidly in one direction, the reaction of the mark price lags significantly behind the true last traded price and the spot index price.

Figure 13-23 compares the delay and cumulative deviation of the mark price relative to the spot index under different EMA windows.

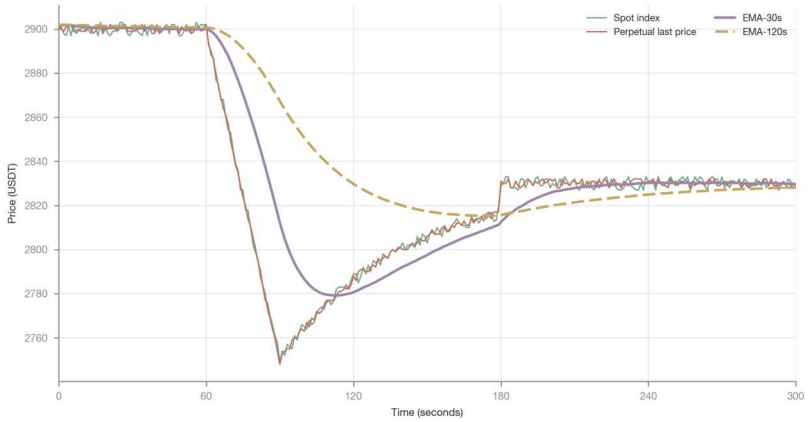


Figure 13-23. The delay and cumulative deviation of the mark price under different EMA windows (illustrative price path demonstrating EMA phase delay, not empirical market data)

Figure 13-23 compares the performance of the spot index price, the perpetual futures' last traded price, and the mark price under different EMA windows (30 seconds, 120 seconds) in a rapidly falling market, intuitively demonstrating the phase delay introduced by EMA smoothing: the longer the window (such as 120 seconds), the more pronounced the lag of the mark price and the larger its cumulative deviation relative to the spot index.

EMA smoothing produces two contradictory effects. As for delayed liquidation, when the true price rapidly breaks below the liquidation-trigger line, the mark price may still be in the safe zone because of the EMA's smoothing, which means that positions that should have been liquidated persist temporarily and accumulate enormous latent liquidation pressure at the low price level; once the mark price finally catches up with the true price, the deferred liquidation instructions erupt all at once, forming an even more violent secondary shock. As for cumulative deviation, in a persistent one-directional move the deviation between the EMA mark price and the true price keeps widening: a longer EMA window (such as 120 seconds), while providing stronger resistance to manipulation, also means a larger cumulative deviation. This deviation not only affects the timing of liquidation but also distorts the calculation of the funding rate (because the rate is usually based on the deviation between the mark price and the index price), thereby feeding the noise of pathway three back into pathway two and forming cross-contamination.

13.5.4 The superposition effects of the three pathways

The three pathways above do not exist in isolation; in actual markets they often interweave and superimpose in resonance. Because these noise sources arise from deterministic system rules (the funding rate settled every 8 hours, the fixed-parameter EMA algorithm, the explicit maintenance margin ratio), they exhibit high periodicity and predictability in the time series.

Understanding the noise characteristics of a single pathway is important, but in reality

the three pathways are often active simultaneously, interweaving and superimposing. Figure 13-24, through a time-series simulation, unfolds the three noise pathways in layers over a multi-day window and superimposes them into a total-noise curve in the bottom panel, revealing how the composite noise peaks cluster around the funding rate settlement moments.

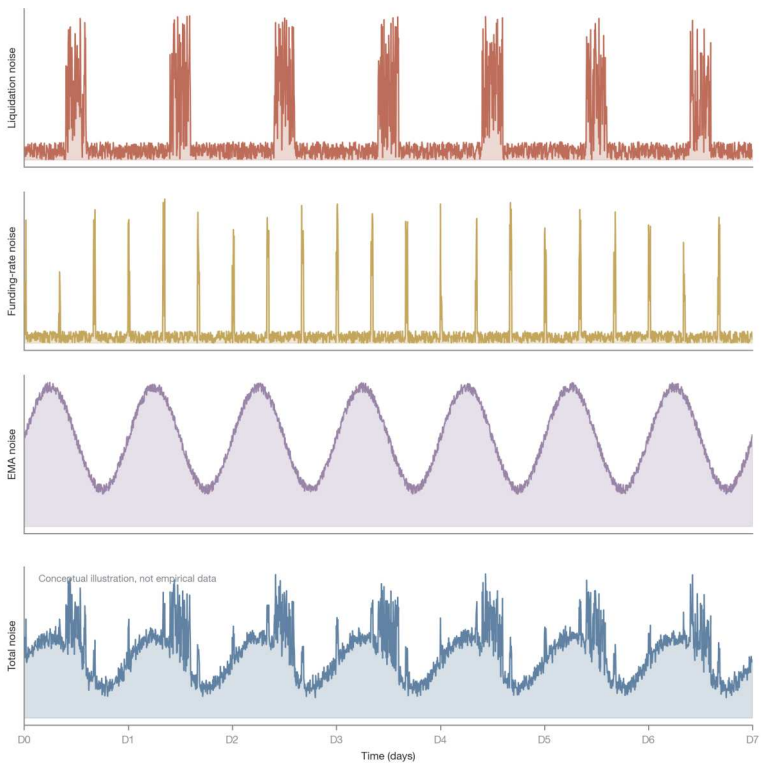


Figure 13-24. Time-series superposition and periodic patterns of the three mechanical noise pathways (conceptual illustration: simulated data, not empirical)

The four-layer panel structure of Figure 13-24 reveals the markedly different temporal-frequency characteristics of each pathway. The first panel shows the sporadic, pulse-like feature of liquidation-driven noise: its intraday distribution is not even but clusters around the alternation of the Asian and European/U.S. sessions, because cross-time-zone liquidity gaps tend to expose the fragility of highly leveraged positions at these moments. The second panel presents the highly regular 8-hour periodicity of funding-rate-extremization noise—about 15 to 30 minutes before each settlement at 00:00, 08:00, and 16:00 UTC, position-adjustment behavior forms an identifiable price pulse. The third panel shows the persistent, low-frequency deviation of mark-price EMA-delay noise, accumulating monotonically during trending conditions. The bottom superposition panel is the most critical: when a liquidation pulse happens to overlap with a funding settlement window (as in the peak region marked in the figure), the resonance of the three noise sources makes the total noise magnitude far exceed the simple sum of any single pathway.

As the simulated results in the figure above show, the noise of the three pathways exhibits different frequency characteristics along the time axis:

These frequency differences have direct significance for the design of quantitative trading strategies. Table 13-5 systematically summarizes the attribute differences of the three noise pathways and their superposition effect along three dimensions: temporal-frequency characteristics, trigger conditions, and ex ante predictability. This classification provides a practical analytical framework for identifying noise-dominated periods versus signal-dominated periods.

Noise pathway	Temporal-frequency characteristics	Trigger condition	Predictability
Pathway 1: Liquidation cascade	Sporadic, but with intraday clustering	Price breaks through a liquidation-dense zone	Moderate (can be anticipated via OI distribution and leverage ratios)
Pathway 2: Funding rate extremes	Strict 8-hour periodicity	Severe long-short imbalance	High (both settlement times and rates are observable)
Pathway 3: EMA delay	Persistent low-frequency deviation, accumulating during trends	Rapid one-directional moves	High (EMA parameters are public; deviation is computable)
Three-pathway superposition	Composite periodicity, peaking around funding settlement windows	A combination of the above conditions	Moderate-to-high (the overall pattern is identifiable)

Table 13-5. Comparison of the attributes of the three mechanical noise pathways and their superposition effect (Data source: compiled by the author)

High-frequency trading firms and quantitative market makers exploit precisely this predictable mechanical noise to extract profits through statistical arbitrage strategies. For example, in the minutes before funding settlement, market makers can anticipate the direction of position adjustment and position themselves in advance; near a liquidation-dense zone, algorithmic traders can predict the trigger point of a cascade by monitoring the distribution of open interest. A distinction must be drawn: the "moderate predictability" in Table 13-5 refers to the quality of the ex ante directional risk signal, not to the profitability of actual trading. The data provided by liquidation heatmaps (such as CoinGlass and Hyblock) are mostly probabilistic estimates based on position distribution rather than precise data on exchanges' actual liquidation thresholds. Strategies that trade around a cascade face serious execution challenges, and the strategy itself may amplify the cascade through a reflexivity mechanism. This "noise-hunting" behavior itself constitutes part of the market microstructure, partly mitigating the magnitude of mechanical noise, but also further increasing the market's complexity.

13.5.5 Cross-market transmission of noise

Bringing together the analysis above, we must recognize a key fact: the mechanical noise produced by perpetual futures is not confined within the derivatives market. Through cross-market arbitrage and oracle price-feed mechanisms, this noise is amplified and transmitted to the entire crypto asset ecosystem, forming a systemic "contamination."

Figure 13-25 presents, as a flowchart, the complete chain by which noise inside perpetual futures is transmitted to the market at large through the arbitrage and oracle mechanisms.

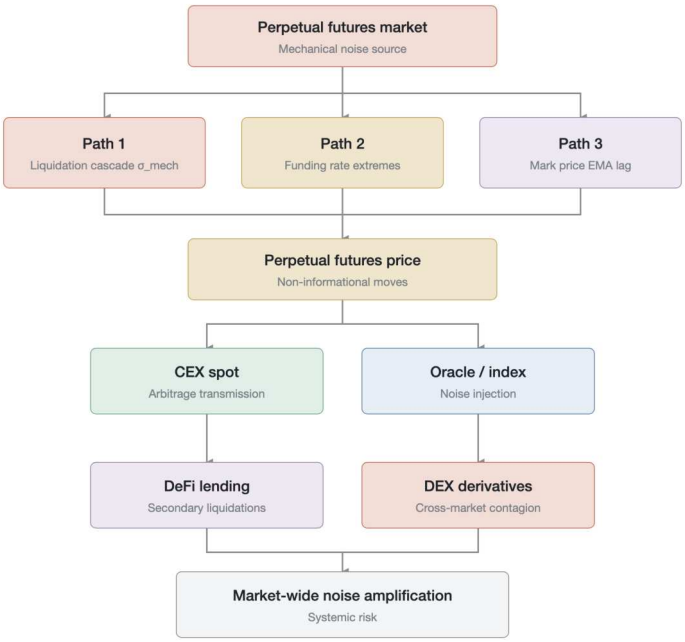


Figure 13-25. The cross-market transmission chain of mechanical noise in perpetual futures

Figure 13-25 shows how the three types of non-informational volatility inside perpetual futures trigger secondary liquidations in DeFi lending and DEX derivatives through the arbitrage and oracle mechanisms, ultimately amplifying into market-wide systemic noise.

The complete chain of noise transmission begins inside perpetual futures. The three pathways—liquidation cascades, funding rate extremes, and EMA delay—first produce non-informational price volatility within the perpetual futures market. Subsequently, after observing the violent price decline in perpetual futures, cross-market arbitrageurs execute corresponding sell operations in the spot market, transmitting the derivatives market's noise to the spot market. Because arbitrageurs cannot distinguish information-driven price changes from mechanical-noise-driven price changes, the two are transmitted indiscriminately. The noise-polluted CEX spot price then enters the on-chain price

index through the aggregation mechanism of oracles such as Chainlink; although oracles usually adopt manipulation-resistant designs such as the median or weighted average, when the spot prices of multiple CEXs are simultaneously polluted by noise, the aggregated index price likewise deviates from true fundamental value.

The polluted oracle price feeds back into DeFi lending protocols (such as Aave and Compound), causing the valuation of on-chain collateral to fall; when the valuation drops below the liquidation threshold, the automatic liquidation mechanism of the DeFi protocol is triggered, producing forced on-chain selling—this is a "secondary cascade" triggered by perpetual futures noise at an entirely different market layer [21]. The on-chain selling pressure produced by DeFi liquidation further depresses DEX and CEX spot prices, feeds back into the oracle and perpetual futures index prices, and completes the closed loop of positive feedback.

Each link of the transmission chain has an intermediate buffering mechanism, so that the transmission of noise is not a perfectly lossless channel. Oracles such as Chainlink have a deviation threshold (such as 0.5%) and a heartbeat interval (as long as 1 hour for some assets), triggering an on-chain update only when the price change exceeds the threshold; the health-factor buffer of lending protocols such as Aave (health factor > 1) is sufficient to absorb the impact of small price fluctuations. The "complete chain" above is therefore fully activated only in the extreme case where the speed and magnitude of the price change simultaneously break through all buffer thresholds. In qualitative terms, the noise transmission ratio in the arbitrage-transmission link is about 1:1, the oracle-aggregation link produces a decay of about 0.6 to 0.8 due to median aggregation and the deviation threshold, and the DeFi-liquidation link may produce an amplification of 1.2 to 2.0 due to the leverage effect within the protocol. In addition, the oracle link itself faces distinctive operational risk: the 2022 Mango Markets incident showed that an attacker can deliberately distort an oracle price feed by manipulating a thin-liquidity spot market, thereby triggering cascade liquidations in DeFi protocols—so the noise-transmission chain can not only be passively activated but also actively weaponized.

As the October 2025 flash crash revealed, the local liquidity exhaustion of USDe on Binance's internal order book (mechanical noise in a single market) ultimately evolved, through the cross-margin mechanism of the unified account and cross-market arbitrage transmission, into a market-wide liquidation disaster of over $19 billion [21]. This event fully demonstrated that, under today's highly interconnected microstructure, perpetual futures have become the "public noise source" of the entire crypto financial system.

This recognition has profound implications for policy and risk management. For regulators, focusing on the risk of a single market is not enough; they must examine the noise-spillover effect of the perpetual futures market from a systemic perspective. For exchanges, introducing mechanisms such as tiered liquidation, multi-source oracle aggregation, and minimum-liquidity filters is a key engineering measure for severing the noise-transmission chain. For investors, understanding the existence and transmission paths of mechanical noise helps distinguish "signal" from "noise" in extreme conditions and avoid making wrong investment decisions amid non-informational price volatility.

13.6 The ceiling-floor model

The price discovery efficiency of perpetual futures is not a static constant but a dynamic variable that fluctuates violently between a theoretical ceiling and an insurmountable floor. This section presents the "ceiling-floor model," the fifth core theoretical contribution of this chapter (CC5), as a unified analytical framework for assessing price discovery in crypto derivatives markets. The core insight of this model is that perpetual futures are not simply "more efficient" or "less efficient" than traditional markets; rather, they have entirely distinct efficiency characteristics: extremely high efficiency in normal markets, but the potential for systemic failure in extreme markets. Understanding these "distinct efficiency characteristics" is of greater theoretical and practical value than assigning a static efficiency score.

13.6.1 The ceiling

The ceiling of price discovery efficiency represents the theoretical upper bound of a market's ability to convert information into price in its optimal state. This upper bound is not determined by a single factor but is jointly shaped by three core dimensions: the number and information quality of informed traders, the rate at which information is generated, and the quality of the market microstructure.

In crypto markets, the height of the ceiling is directly related to the richness of the information environment. As on-chain analytics tools (such as Nansen, Arkham, and Dune Analytics) proliferate, the threshold for information interpretation falls sharply, enabling more participants to become informed traders. More informed traders mean more private information and on-chain observable information injected into the market, thereby raising the theoretical upper bound of price discovery. In addition, the participation of high-frequency market makers and extremely deep liquidity pools ensure that order flow can be absorbed with minimal friction, further raising the quality of the microstructure. In the venue gravity model of Section 13.3, the systematic advantages of perpetual futures across the four dimensions of leverage availability, liquidity depth, anonymity, and transaction cost are the microstructural basis for their ceiling being higher than the spot market's.

The ceiling can be summarized by the following functional relationship:

$$\text{Ceiling} = f(\text{number of informed traders},\ \text{information quality},\ \text{market microstructure quality})$$

Along all three dimensions, the perpetual futures market exhibits structural advantages. The availability of leverage allows informed traders to express a stronger information signal with less capital; deep liquidity ensures that large orders do not produce excessive price impact; and 24/7 continuous trading ensures that information can be priced in instantly at any moment. Together, these features push the ceiling of perpetual futures to the highest level among all crypto trading venues.

13.6.2 The floor

Opposite the ceiling, the floor of price discovery efficiency represents the lower bound of a market's performance in its worst state. In the perpetual futures market, this floor is set by mechanical noise and is insurmountable in nature. As long as perpetual futures rely on leverage and a forced-liquidation mechanism, mechanical noise cannot fall to zero.

The depth of the floor is jointly determined by the three noise-injection pathways analyzed in Section 13.5:

$$\text{Floor} = f(\sigma_{\text{mech}},\ \text{funding rate distortion},\ \text{mark price delay})$$

Among these, liquidation-driven mechanical volatility (σ_{mech}) is the most destructive noise source. In extreme markets, these three pathways act simultaneously, injecting into prices large amounts of selling or buying pressure that contain no fundamental information at all. In today's mainstream perpetual futures design, the leverage function inherently depends on some form of forced risk-clearing mechanism, so mechanical noise is an endogenous feature that the perpetual futures market must accept. Alternative designs are exploring different trade-off paths; for example, Curve's LLAMMA soft-liquidation mechanism allows a position to smoothly transform its collateral form (gradually converting from a volatile asset into a stablecoin) as it approaches the liquidation threshold, rather than triggering hard-liquidation-style forced selling. Such designs can significantly reduce the instantaneous impact of a liquidation event on the order book, but they cannot fully eliminate mechanical noise: even the collateral-form conversion in a soft liquidation still produces observable market order flow and exerts non-informational pressure on the price. This logic reveals a profound design paradox: the very same mechanism (leverage) that gives perpetual futures the highest ceiling also sets their highest floor.

Although the floor cannot be eliminated, it can be lowered by engineering means (that is, by reducing the destructive power of mechanical noise). For example, optimizing cliff-edge liquidation into gradual liquidation or introducing a vault-takeover mechanism (see Section 11.4) can significantly reduce the impact of a single liquidation on the order book; raising the settlement frequency of the funding rate (such as from once every 8 hours to once per hour or continuous settlement; see Section 10.5) can reduce the price distortion from rate arbitrage; and optimizing the calculation logic of the mark price to reduce reliance on a single spot price source (see Chapter 14) can also lower the floor to some extent. Each optimization can lower the floor, but none can bring it to zero.

13.6.3 The state dependence of efficiency

In reality, the price discovery efficiency of perpetual futures exhibits pronounced state-dependent fluctuation between the ceiling and the floor, which can be divided into three typical phases.

In a normal market, mechanical noise is dormant, and the efficiency curve runs close to the ceiling. At this time, informed traders' information is expressed and aggregated efficiently, and perpetual futures exhibit price discovery leadership far exceeding that of the spot market. The information flow topology of Section 13.4 shows that in this state,

the information share of Binance perpetual futures is usually about 30%–40%, and its perpetual and spot markets combined can exceed 50%, making it a widely recognized center of price discovery.

In a liquidation-intensive period, however, as the chain reaction of leveraged liquidations is triggered, mechanical noise surges exponentially. At this time, the efficiency curve plunges sharply, approaching the floor set by the mechanism. Price changes no longer reflect the updating of fundamental information but are driven entirely by order book depth and the selling force of the liquidation engine. In this phase, perpetual futures not only lose their price discovery efficacy but become a source of contamination that outputs mispricing signals to the entire crypto ecosystem, and the cross-market noise-transmission paths analyzed in Section 13.5 are fully activated at this moment.

In the recovery period after a crisis, the rebound of efficiency exhibits a pronounced asymmetry. The decline of efficiency usually occurs instantaneously (accompanying a liquidation cascade), but the rebound is a slow process. This is because the loss of liquidity tends to be sudden, whereas market makers reassessing risk, replenishing inventory, and restoring normal quotes take considerable time. This asymmetry is fully consistent with the liquidity-resilience feature analyzed in Section 8.5: liquidity is lost far faster than it recovers. From the perspective of behavioral finance, the asymmetry of recovery stems not only from the mechanical process of market makers replenishing inventory but also from the pervasive loss of trust after a crisis. The core finding of prospect theory shows that the degree of persistent risk aversion after experiencing a loss systematically exceeds what rational Bayesian updating would predict, and trust, as a psychological variable, naturally recovers more slowly than balance-sheet variables such as inventory. This means that the actual efficiency-recovery time may be significantly longer than the prediction given by a pure microstructure model based on inventory-adjustment logic. In practice, this means that just after a liquidation crisis subsides, market participants cannot assume that price discovery efficiency has recovered immediately; they must wait for microstructure indicators such as the bid-ask spread and order book depth to improve substantively before trusting the perpetual price as an accurate reflection of fundamentals.

Figure 13-26 visualizes the ceiling-floor model, marking the moments in four historical crises when efficiency plunged from the ceiling to the floor.

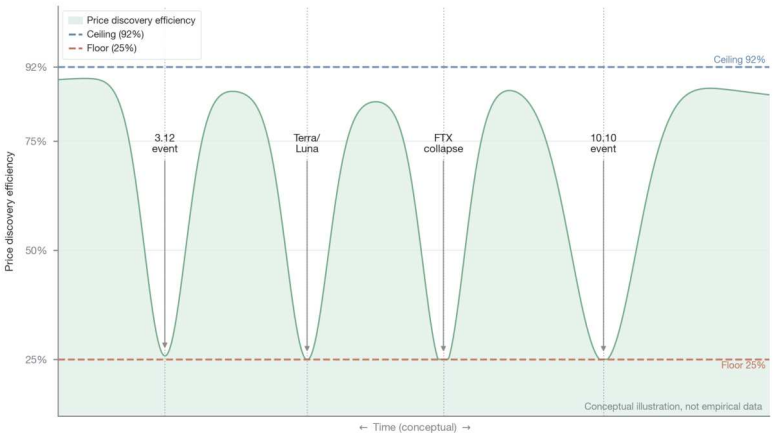


Figure 13-26. The ceiling-floor model of perpetual futures price discovery efficiency (conceptual illustration, not empirical data)

Figure 13-26 shows that the price discovery efficiency of perpetual futures runs close to the ceiling in normal markets, plunges to near the floor in a liquidation crisis, and exhibits an asymmetric recovery characteristic of "fast decline, slow rebound." The four historical crisis events marked in the figure clearly show the moments of efficiency plunge.

13.6.4 How to raise the ceiling and lower the floor

Having understood the determinants of the ceiling and the floor, market designers face two relatively independent engineering problems: raising the ceiling and lowering the floor (Table 13-6). These two directions can and should be pursued simultaneously, because they determine, respectively, the market's competitiveness in good times and its resilience in bad times.

The path to raising the ceiling focuses mainly on optimizing the information environment and market structure. This is a question of "how to make signals clearer and transmission faster." Specific paths include developing more advanced on-chain data-parsing tools to reduce information asymmetry (such as the application of AI agents in information processing discussed in Section 4.5); introducing high-frequency arbitrage algorithms to accelerate cross-market information transmission and reduce the transmission-stage friction analyzed in Section 13.4; attracting informed traders and high-quality market makers through a more reasonable fee structure (such as a maker rebate mechanism); and building deeper liquidity pools so that large information-driven trades can be absorbed by the market at a lower impact cost.

The path to lowering the floor is entirely a matter of protocol mechanism design. This is an engineering challenge of "how to minimize mechanical noise while retaining the leverage function." Beyond the gradual liquidation and high-frequency funding settlement mentioned earlier, optimizing the auto-deleveraging trigger logic to reduce improper interference with the profitable side, expanding the buffering capacity of the insurance fund to absorb liquidation impacts, and implementing a more refined cross-asset margin-

assessment model to avoid the "weakest-link" effect of unified margin are all effective means of preventing efficiency from completely collapsing under extreme conditions.

Engineering direction	Core question	Key pathways	Related sections
Raise the ceiling	How to make signals clearer and transmission faster	Proliferation of on-chain analytics tools, AI-agent information processing, market-maker incentive optimization, reduction of cross-market arbitrage friction	4.5, 13.3, 13.4
Lower the floor	How to minimize mechanical noise while retaining leverage	Gradual liquidation, high-frequency funding settlement, mark price optimization, improved ADL logic, insurance fund expansion	10.5, 11.4, 14

Table 13-6. Engineering pathways for raising the price discovery ceiling and lowering the floor (Data source: compiled by the author)

13.6.5 A comparison of three markets

Applying the ceiling-floor model to different market types clearly reveals their essential differences in price discovery (Table 13-7).

Market type	Ceiling	Floor	Range width	Efficiency profile
Traditional futures (e.g., CME)	Medium-high: mature information environment; since CME BTC futures launched in 2017, they have traded nearly 23 hours on weekdays (Sunday to Friday, 17:00–16:00 CT, including a 1-hour break), but still cannot trade on weekends and holidays; institutional participation increased markedly after the 2024 spot ETF approval	Low: liquidation cascades are tightly constrained by the periodic "leverage reset" at expiry, the central counterparty default waterfall, and price limits and circuit breakers	Narrow	Relatively stable efficiency; "not the best, but never too bad"
Crypto spot (e.g., Coinbase)	Medium: good information environment and 24/7 trading, but the lack of leverage limits informed traders' "expressive power"	Extremely low: no forced-liquidation mechanism, almost no mechanical noise	Medium	"No mechanical noise, but insufficient information expression"
Perpetual futures (e.g., Binance)	Highest: high leverage + deep liquidity + low barriers + 24/7, the fullest expression of informed trading	Highest: liquidation cascades + rate distortion + mark price delay, the largest mechanical noise	Widest	"Best at its best, worst at its worst, most volatile"

Table 13-7. Comparison of three market types under the ceiling-floor model of price discovery (Data source: compiled by the author)

The ceiling of traditional futures, although benefiting from a mature information environment and the depth of institutional participants (especially after the approval

of the spot Bitcoin ETF in January 2024, which significantly raised CME's institutional trading volume and information share), is still limited by trading interruptions on weekends and holidays (CME BTC futures trade on weekdays from Sunday to Friday, 17:00–16:00 CT, nearly 23 hours, not around the clock without interruption). Its floor is effectively constrained jointly by the central counterparty's risk-management system, the "periodic leverage reset" at expiry, and price limits and circuit breakers. Price limits and circuit breakers are institutional buffers unique to traditional futures markets: when a price change reaches a preset threshold, trading is forcibly suspended or restricted within a certain band, thereby providing market participants a window to reassess information and replenish liquidity, and effectively interrupting the positive feedback loop of panic selling. By contrast, crypto perpetual futures, while trading 24/7 without interruption, completely lack such circuit-breaker mechanisms—precisely one important institutional reason their floor is higher than that of traditional futures. For this reason, the efficiency-fluctuation range of traditional futures is relatively narrow, exhibiting a stable but non-optimal character.

The ceiling of crypto spot is limited by the lack of leverage; informed traders can only express their views at 1:1 capital efficiency, which within the Kyle model framework implies a lower intensity of information expression. But its floor is nearly zero, because without a forced-liquidation mechanism there is no mechanical noise. The efficiency profile of crypto spot is "stable but insufficient."

The ceiling of perpetual futures is the highest, because the four-dimensional gravity of leverage, liquidity, anonymity, and low cost (Section 13.3) makes informed traders' information expression the fullest. But its floor is also the highest, because liquidation cascades, rate distortion, and mark-price delay (Section 13.5) constitute the strongest source of mechanical noise. This combination of "highest ceiling + highest floor + widest range" means that perpetual futures are a widely recognized price discovery leader in normal markets but may exhibit the most severe price distortion in extreme markets.

Perpetual futures are not "more efficient" or "less efficient" than traditional markets—they are "more efficient in normal markets, possibly less efficient in extreme markets, with a range of efficiency fluctuation far greater than that of traditional markets." This is a "distinct efficiency profile," not a simple "better" or "worse."

Figure 13-27 uses a bar chart to compare the ceiling height, floor depth, and efficiency-range width of the three markets: traditional futures, crypto spot, and perpetual futures.

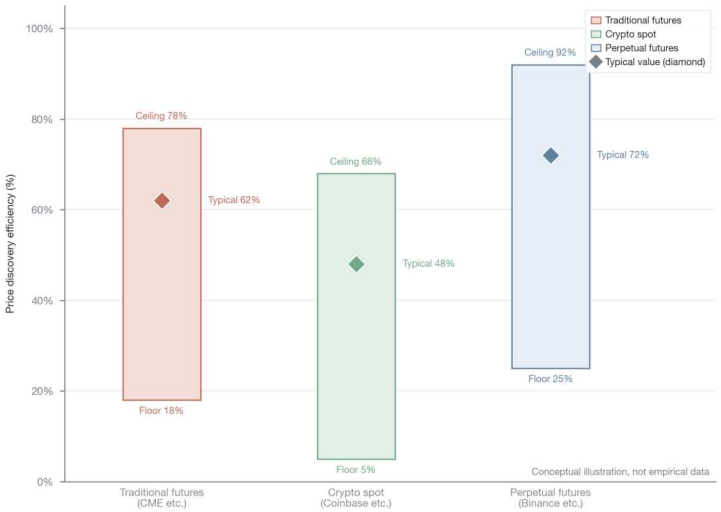


Figure 13-27. Ceiling-floor comparison of traditional futures, crypto spot, and perpetual futures (conceptual illustration, not empirical data; value conventions consistent with Figures 13-26 and 13-38)

The comparison in Figure 13-27 vividly confirms the markedly different efficiency profiles of the three markets. The distance between the ceiling and floor of traditional futures is the narrowest, reflecting how their mature institutional constraints (price limits, central counterparty clearing, periodic expiry) lock efficiency fluctuation into a small range—at the cost of a ceiling that is not as high as that of perpetual futures. The floor of crypto spot nearly touches the zero line (because there is no forced-liquidation mechanism), but its ceiling is also clearly lower, because the lack of leverage limits informed traders' capacity for information expression. The visual image of perpetual futures is the most striking: the ceiling is the highest and the floor is also the highest, and the range between them is the widest of the three. This graphical comparison captures the core paradox of price discovery in perpetual futures—the same mechanism (leverage) both pushes the ceiling to the highest and raises the floor to the highest.

13.7 Obstacles to price discovery

After analyzing the three injection pathways of mechanical noise in Section 13.5, this section broadens the view to the wider obstacles to price discovery. Four forces—the reflexivity of liquidation cascades, the signal pollution of market manipulation, the amplification effect of behavioral biases, and the transmission delay of information friction—appear simultaneously and reinforce one another under extreme conditions, leading to a systemic failure of price discovery. If the ceiling-floor model of Section 13.6 describes the "boundaries" of efficiency, this section dissects the specific "forces" that push efficiency toward the floor.

13.7.1 Liquidation cascades and price overshoot

The liquidation cascade is the most destructive microstructure phenomenon in crypto derivatives markets. It is not merely a risk-clearing process but a classic reflexive feedback loop that can drive prices far from fundamentals and produce enormous overshoot.

The basic mechanism of this feedback loop has already been defined in Section 13.5.1: an initial downward shock triggers forced liquidations, the liquidation sell orders depress the price, and this in turn triggers the liquidation of still deeper positions. The reflexivity perspective of Section 13.7.1 further highlights the amplifying role of market-maker behavior in this loop: facing concentrated non-informational selling pressure, market makers, to avoid adverse-selection risk and inventory risk, usually choose to withdraw orders or widen the bid-ask spread sharply, causing the buy-side depth of the order book to plunge; a new wave of liquidation sell orders is then executed against an even thinner order book, producing a more severe price impact than the previous round, until the leverage is fully cleared or external intervention occurs.

Figure 13-28 presents, as a flowchart, the positive feedback loop mechanism of the liquidation cascade and its amplification path.

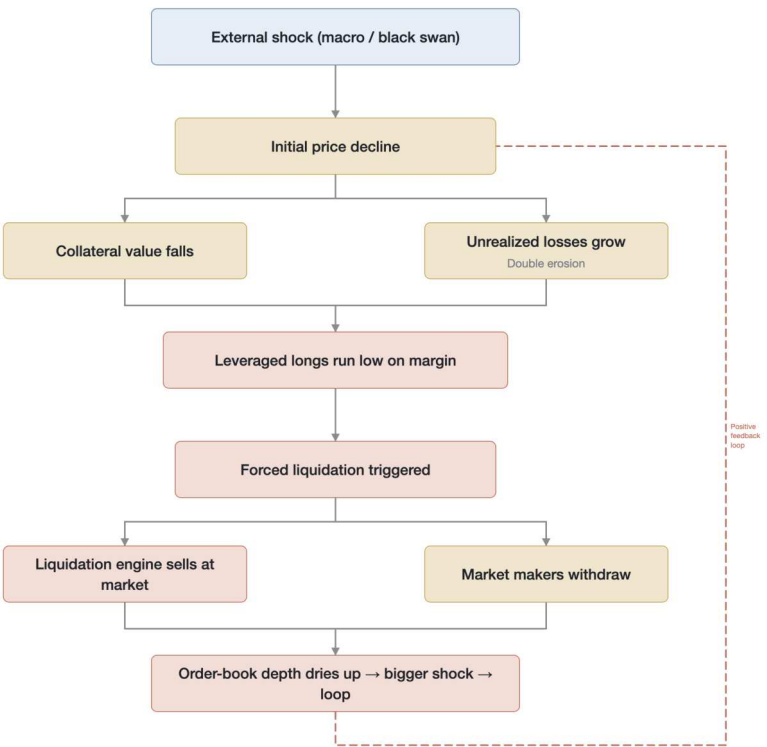


Figure 13-28. The reflexive positive feedback loop of a liquidation cascade

Figure 13-28 shows the complete positive feedback chain from external shock, double erosion, insufficient margin, forced liquidation, and market-maker withdrawal to the

exhaustion of order book depth, as well as the superimposed amplification of this loop by panic contagion.

Two important amplification mechanisms are also nested within the liquidation cascade. The double-erosion effect is one key link: under the unified-margin model, a price decline not only increases unrealized losses but also simultaneously depreciates the crypto asset serving as collateral, accelerating the depletion of the margin balance from two directions at once. FTI Consulting noted in its 2025 report that unified margin is "highly efficient in calm markets" but that "under stress, the same design ties the portfolio to its weakest asset" [26]. On this basis, panic contagion further amplifies the crisis: liquidation data (such as "$X hundred million liquidated in the past hour") spread rapidly through social media, triggering panic selling by retail investors, and this sentiment-driven selling pressure superimposes on the mechanical selling pressure, further intensifying the liquidity crisis.

From a historical-data standpoint, the scale and destructive power of liquidation cascades exhibit an alarming escalating trend. Table 13-8 summarizes the key parameters of four landmark liquidation events in crypto markets:

Event	Date	24h liquidations (approx.)	BTC max drawdown (approx.)	Core trigger	Key structural feature
March 12 event	March 2020	$1 billion	~50%	COVID panic + BitMEX structural flaw	Single-exchange dominance, order book depth to zero
Terra/Luna	May 2022	$2 billion	~22%	Algorithmic stablecoin depegging	Cross-protocol contagion, compounded by DeFi liquidations
FTX bankruptcy	November 2022	$2.8 billion	~25%	Exchange solvency crisis + fraud	Trust crisis, withdrawal run
October 10 event	October 2025	Over $19 billion	~14.5% (spot; perpetuals ~16%–17%)	Macro shock + unified-margin risk	ADL mechanism, USDe depegging, infrastructure freeze

Table 13-8. Comparison of key parameters across four landmark liquidation events in crypto markets (Data source: compiled by the author, based on public market data)

Figure 13-29 uses a bar chart and a line chart to compare the 24-hour visible liquidation notional and the BTC maximum drawdown across four events: the March 12, 2020 event, the Terra/Luna collapse of 2022, the FTX bankruptcy of 2022, and the October 10, 2025 event.

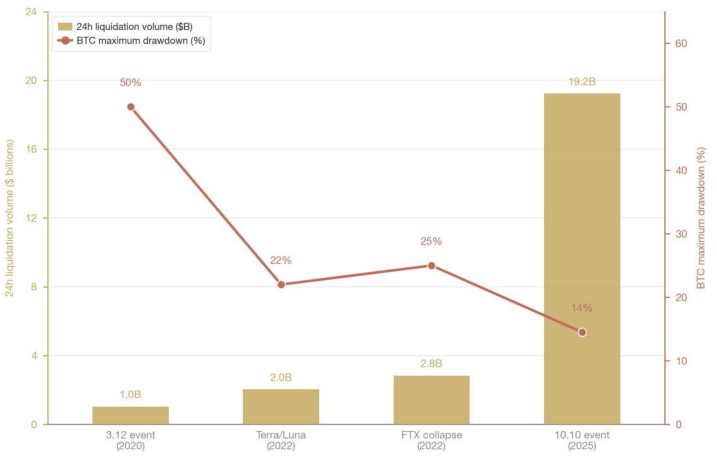


Figure 13-29. Comparison of the scale and impact of four major liquidation cascade events (Data source: [27][26])

Figure 13-29 shows that from the March 12 event of 2020 to the October 10 event of 2025, the 24-hour visible liquidation notional grew about 19-fold; at the same time, the BTC maximum drawdown on the whole did not rise but fell—the accumulation of tail risk is reflected mainly in the sharp swelling of liquidation scale rather than in a widening of the price decline.

Whether the 19-fold growth in visible liquidation notional is linearly related to the growth of the market's total open interest (OI) deserves closer examination. From 2020 to 2025, the total OI of the crypto derivatives market grew from under $10 billion to on the order of $100 billion, an increase of about 10-fold, while over the same period the visible liquidation notional grew nearly 19-fold, meaning that the liquidation risk per unit of OI is rising rather than falling. Behind this superlinear trend is the systematic climb in the ratio of the market's total leverage to market capitalization: the proliferation of the unified-margin model allows the same principal to leverage a larger notional exposure, and cross-asset margin sharing further amplifies the effective leverage level. As the scale of perpetual futures continues to expand, a key question emerges: can the risk-absorption capacity of exchange insurance funds and ADL mechanisms keep pace with this superlinear growth? The scale of insurance funds grows mainly through the linear accumulation of liquidation surplus, whereas the tail risk they must bear swells at a nonlinear rate, which suggests the possible existence of a critical threshold: beyond it, existing risk-mitigation mechanisms will be unable to absorb liquidation shocks effectively, and the intensity of systemic crises will jump up in a step change.

In the March 12 event of 2020, the order book structure of BitMEX collapsed completely. According to Multicoin Capital's microstructure autopsy, only about $20 million of buy orders remained on the BitMEX order book at the time, yet they had to absorb over $200 million of long-liquidation selling pressure [27]. The Bitcoin price broke below $4,000 within minutes, and this price level reflected not fundamental information but purely the product of a market-structure failure. BitMEX's structural flaw lay in its

accepting only BTC as margin collateral, which meant that all long positions naturally bore the double-erosion effect. Had BitMEX not briefly gone offline for "system maintenance," interrupting the operation of the liquidation engine, the price would very likely have reached levels far below fundamentals.

By the October 10 event of 2025, the visible liquidation notional reached an unprecedented height. According to preliminary estimates from multiple parties, over $19 billion of leveraged crypto positions were forcibly liquidated in about a day, roughly 19 times the March 12 event (Greco [21] and FTI Consulting [26] gave similar estimates based on different methodologies, but because exchanges use inconsistent liquidation-data disclosure standards, this figure should be regarded as an order-of-magnitude estimate rather than a precise statistic). The BTC price plunged during the event from an intraday high of about $122,500 to about $104,800 in spot (a maximum drawdown of about 14.5%; perpetual futures, due to depth exhaustion, wicked lower at some venues, to about $102,000), then partially rebounded to about $113,000 within hours, but full recovery to pre-event levels took several weeks. This event introduced several new structural risk dimensions. The unified-margin system tied the fate of a portfolio to its weakest-performing asset; exchange interfaces froze due to the surge in traffic, so that traders could not add margin or manage risk; and the auto-deleveraging mechanism forcibly closed counterparties' profitable positions, instantly turning a perfectly hedged portfolio into a naked exposure. As FTI Consulting noted, the ADL mechanism essentially introduces "rule risk": a trader's strategy may be logically flawless, but at the most critical moment the exchange's survival rules do not allow that strategy to persist [26].

In these extreme events, the imbalance of the order book reached a pronounced degree. FTI Consulting's report shows that on the day of the October 10 event, the top-of-book depth for BTC shrank by more than 90%, and the bid-ask spread widened from single-digit basis points to double-digit percentages [26]. The symmetrically distributed bid and ask depth of a normal market became extremely asymmetric during the crisis: buy-side depth nearly vanished, while the sell side piled up with vast liquidation sell orders and panic selling. At this point, the "information machine" of the aggregation stage defined in Section 13.2 was already paralyzed: the order book no longer aggregated information but merely mechanically processed the liquidation engine's selling instructions. The typical behavior pattern of market makers in extreme conditions can be summarized as "retreat first, assess later": when volatility surges and liquidation selling pressure pours in in concentration, a market maker's optimal strategy is to immediately withdraw all quotes to avoid inventory risk, rather than attempting to keep providing liquidity amid the chaos. This collective withdrawal is a rational individual decision, yet it leads to a catastrophic collective outcome: the order book loses its liquidity supply at precisely the moment liquidity is most needed. After the crisis subsides, the market makers' return process follows an entirely different dynamic. The typical indicators by which market makers assess a "safe" return include a retreat in realized volatility (usually needing to fall below one-third of the crisis peak), a significant decline in liquidation volume (indicating that the leverage clearing is essentially complete), and a stabilization of OI (indicating that the risk of a new round of passive liquidation has fallen sharply). As these conditions are gradually

met, market makers rebuild positions tentatively, starting from extremely wide quotes with small orders. The speed of spread narrowing is therefore far slower than that of its widening—this "minute-scale evaporation, day-scale recovery" temporal asymmetry and its behavioral-finance roots were developed in Section 13.6.3.

The damage of the liquidation cascade to price discovery is ultimately reflected in the distortion of the order book's microstructure. Figure 13-30 uses depth snapshots to compare the bid and ask distribution of BTC-USDT perpetual futures in a normal market versus at the peak of a liquidation crisis.

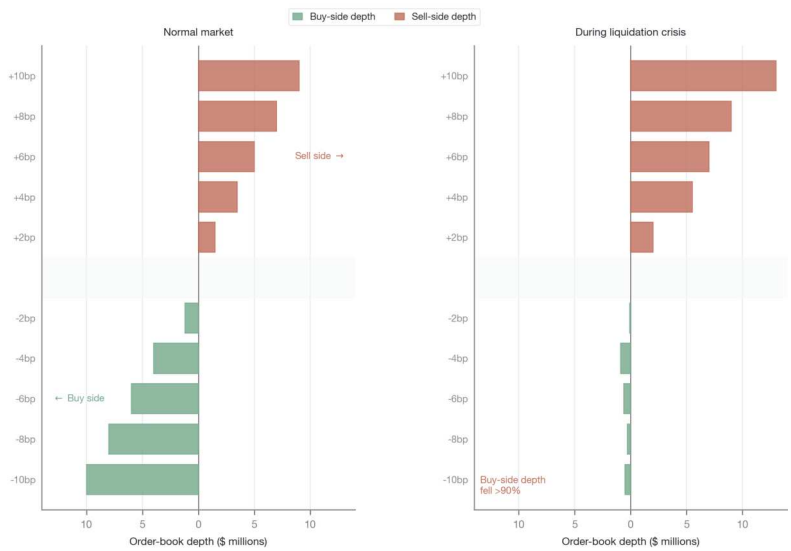


Figure 13-30. Order book structure comparison between normal markets and a liquidation crisis (Data source: [26]; level-by-level depth is a representative reconstruction of FTI's "top-of-book depth reduced by more than 90%" conclusion)

The comparison in Figure 13-30 reveals a shocking microstructural fact. In a normal market, the two sides of the order book exhibit an approximately symmetric depth distribution, with the bid and ask sides each maintaining tens of millions of dollars of orders near the mid-price. At the peak of a liquidation crisis, however, this symmetry is completely shattered: buy-side depth shrinks by more than 90%, leaving only a few tentative low-price orders scattered far from the mid-price, while the sell side piles up with large orders forcibly issued by the liquidation engine and small-to-medium orders from retail panic selling, forming an extremely asymmetric structure of "a mountain of sell orders and paper-thin buy orders." In this order book form, every new liquidation sell order causes an unusually large price impact, because it essentially faces a vacuum with almost no absorptive capacity.

In discussing liquidation cascades, one must also elevate the degradation of exchange infrastructure from an "incidental malfunction" to a systemic operational-risk factor deserving scrutiny. The exchange-interface freezes, API-latency spikes, and withdrawal suspensions that appeared during the October 10 event were not isolated technical accidents but the systemic result of exchanges' capacity planning long prioritizing average

load over extreme load. An exchange's infrastructure-investment decision is essentially an economic trade-off: over-provisioning servers and bandwidth to handle extreme peaks means bearing high idle costs during the vast majority of normal trading hours, so exchanges generally tend to plan capacity at some multiple of average traffic rather than at extreme peaks. Yet a deep incentive misalignment exists in this decision: exchanges directly earn liquidation fees from the liquidation process (usually a certain proportion of the position's notional value), and when their infrastructure degrades during extreme volatility, users lose the ability to add margin or actively reduce positions at precisely the moment they most need to manage risk, and thus passively slide toward liquidation. In other words, a structural conflict of interest exists between the losses an exchange bears from infrastructure failure (reputational risk and user attrition) and the revenue it earns from users' unavoidable liquidations. This operational risk directly affects the setting of the "floor" in the ceiling-floor model of Section 13.6: infrastructure degradation itself becomes an independent force pushing price discovery efficiency toward the floor, and its destructive power is especially pronounced when superimposed on a liquidation cascade.

13.7.2 Wash trading and spoofing

If the liquidation cascade is a mechanism-level destruction of price discovery, then market manipulation is a direct pollution of the price discovery signal. In a microstructure where order flow serves as the primary information carrier (see Section 7.6 and Chapter 8), wash trading and spoofing seriously interfere with the judgments of market makers and informed traders, fundamentally weakening the signal-extraction capacity of the aggregation stage.

Wash trading refers to a single entity acting simultaneously as both buyer and seller, artificially manufacturing a false impression of trading volume and liquidity. This behavior not only misleads investors' assessment of an asset's activity level but also directly undermines the validity of microstructure metrics such as order flow imbalance: if a considerable proportion of trading volume is self-generated, then volume-based information signals are severely diluted. Regarding the proportion of wash trading in crypto markets, different research institutions have given vastly different estimates based on different methodologies:

Figure 13-31 summarizes the estimates of the wash trading proportion in crypto markets from different research institutions.

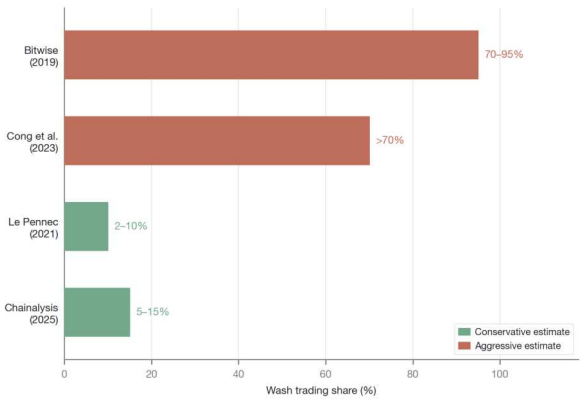


Figure 13-31. Estimates of the wash trading proportion in crypto markets across different studies (Data source: [28][29][30][31]; the conservative ranges of Chainalysis 5%–15% and Le Pennec 2.1%–10% are compiled by the author and pending verification)

Figure 13-31 shows that, owing to differences in detection methods and sampled exchanges, the estimates of the wash trading proportion given by various research institutions range from 2% to 95%, and this vast range itself illustrates the complexity of the problem. Even using the most conservative estimate of 2% to 10%, the pollution of the price discovery aggregation stage by wash trading cannot be ignored, because it systematically dilutes the signal strength of order flow information metrics such as OFI, biasing market makers' judgment of adverse-selection risk. As on-chain DEX trading becomes more transparent and regulatory technology advances, the detectability of wash trading keeps improving, but in the closed environment of CEXs this problem will persist for a long time.

In its 2019 report submitted to the SEC, Bitwise, based on trading-volume-pattern analysis, noted that as much as 95% of the Bitcoin trading volume on unregulated exchanges was fabricated [28]. Cong and colleagues (2023), in research published in *Management Science*, used statistical detection and Benford's law to estimate that the wash trading proportion at secondary exchanges exceeds 70% [29]. Le Pennec and colleagues (2021), through a refined on-chain detection method, gave a lower estimate of 2.1% to 10% [30]. Chainalysis and other institutions based on on-chain tracing gave a relatively conservative estimate of 5% to 15% [31]. The difference in estimates itself illustrates the complexity of the problem and the difficulty of measurement, but even using the most conservative data, the pollution of the price discovery aggregation stage by such false signals cannot be ignored.

Spoofing, by contrast, involves a manipulator posting large limit orders deep in the order book to manufacture false supply-and-demand pressure, inducing other algorithms or traders to trade in the direction the manipulator wants, and then rapidly canceling the orders before they fill. This strategy directly attacks the market maker's "learning mechanism": market makers infer adverse-selection risk from order book depth and order-flow patterns (see Chapter 8), and spoof orders cause this inference to deviate systematically from the true state, leading market makers to provide quotes that deviate from true value. In traditional financial markets, spoofing has been explicitly

designated as illegal by the Dodd-Frank Act, but in crypto markets the monitoring and punishment mechanisms have long been relatively weak. This picture, however, is undergoing a substantive shift. The U.S. Commodity Futures Trading Commission (CFTC) has in recent years brought multiple enforcement actions against crypto market manipulation, covering a variety of violations including wash trading, spoofing, and market manipulation. A landmark but repeatedly reversed case is the Mango Markets manipulation case: Eisenberg, by manipulating the price of MNGO on the Mango Markets platform in October 2022, extracted over $110 million from the protocol and was convicted by a jury in April 2024; but in May 2025, the U.S. District Court for the Southern District of New York vacated all criminal convictions under Rule 29 of the Federal Rules of Criminal Procedure on grounds of improper venue and insufficient elements of wire fraud, the prosecution has filed an appeal, and its final legal characterization remains uncertain as of this writing. This reversal itself reveals precisely the legal uncertainty of criminally prosecuting on-chain manipulation on questions of venue and the elements of fraud. At the same time, the CFTC's assertion of jurisdiction over perpetual futures is moving from theory to practice, and in several enforcement actions it has treated perpetual futures as "swaps" within its regulatory purview. Even so, cross-jurisdictional enforcement coordination still faces major challenges, manipulation by offshore exchanges remains difficult to reach at the level of actual enforcement, and the interference of spoofing with price discovery will remain a persistent problem for a considerable period.

13.7.3 Behavioral biases and narrative-driven trading

Price discovery in crypto markets is subject not only to mechanism defects and deliberate manipulation but also deeply to investor behavioral biases. Owing to the lack of a traditional valuation anchor (such as discounted cash flow or a price-to-earnings ratio), the prices of crypto assets are highly susceptible to being driven by sentiment and narrative, which in the three-stage model of Section 13.2 acts mainly on the expression stage, where large amounts of non-informational, sentiment-driven orders pour into the market and swamp the signals of informed traders.

A 2024 survey by Kraken of 1,248 crypto asset holders revealed pronounced behavioral characteristics: as many as 84% of respondents admitted to having made investment decisions driven by "fear of missing out" (FOMO), and 81% had traded out of "fear, uncertainty, and doubt" (FUD) [32]. This highly emotional herd behavior injects a large amount of non-fundamental noise in the expression stage, making it harder for market makers to separate "signal" from "noise" in the order flow in the aggregation stage. This Kraken survey, however, has an obvious selection bias: the respondents were all existing crypto asset holders who voluntarily participated in the survey through an exchange platform, and their behavioral characteristics should not be directly generalized to the overall population of market participants.

Even so, the broader academic literature on crypto behavioral finance corroborates, along multiple dimensions, the pronounced amplification effect of behavioral biases in this market. As for the disposition effect, a number of empirical studies show that crypto-market investors exhibit an even stronger tendency to "sell winners and hold

losers" than traditional stock-market investors, and this effect is further amplified in the 24/7 continuous-trading environment: investors can check their account P&L and make trading decisions at any moment, lacking the "cooling-off period" provided by the closing hours of traditional markets, which significantly raises the frequency of emotion-driven trading. As for social-media-driven attention trading, the information-dissemination ecosystem of crypto markets exhibits a distinctive structure: real-time P&L sharing (PnL screenshots) in Telegram groups and Discord servers creates strong social-comparison pressure, copy-trading platforms directly turn individual trading decisions into signal sources that can be replicated in bulk, and narrative dissemination by KOLs through Twitter/X (often called "Crypto Twitter" or "CT") constitutes an information-cascade channel from a few influencers to a mass retail audience. Together, these mechanisms constitute a herd-behavior amplifier unique to crypto markets: compared with traditional markets, where information spreads mainly indirectly through analyst reports and institutional-holdings disclosures, information dissemination in crypto markets is more immediate, more intuitive, and characterized by strong social reinforcement.

At the theoretical level, the Adaptive Markets Hypothesis (AMH) proposed by Lo (2004) provides a more apt framework for understanding these phenomena than the traditional efficient market hypothesis (EMH) [33]. AMH holds that market efficiency is not a constant state but a process that fluctuates dynamically with changes in participant composition, competitive intensity, and the institutional environment. In crypto markets, the composition of participants undergoes dramatic cyclical shifts: during bull markets, large numbers of inexperienced retail investors flood in, and their behavior is dominated by narrative-chasing and momentum trading; during bear markets, after retail investors exit, the market is dominated by more experienced participants, and information efficiency rises accordingly. This cyclical change in participant composition means that the price discovery efficiency of crypto markets is itself a time-varying quantity that may sit at different positions on the efficiency spectrum in different market phases. This theoretical insight resonates directly with the ceiling-floor model of Section 13.6: the efficiency range between the ceiling and the floor is not static but expands or contracts dynamically as the ecosystem of market participants evolves.

In addition, narrative-driven trading logic is especially pronounced in crypto markets. Whether "DeFi Summer," "the rise of NFTs," or "AI + Crypto," a powerful narrative can rapidly attract enormous inflows of capital, even if the relevant projects lack real applications or sustainable cash flows. In this mode, the driver of price discovery is alienated from "assessing fundamental information" into "guessing the breadth of narrative dissemination," which is similar to Keynes's "beauty contest" theory: investors are not assessing the true value of an asset but guessing what other investors will consider valuable.

Narrative-driven reflexivity is significantly amplified in crypto markets. A bullish narrative pushes the price up, and the price rise in turn "confirms" and reinforces the narrative, attracting more followers and pushing the price up further—a classic positive feedback loop. When the narrative collapses, however, the negative feedback loop is equally violent: the price decline "falsifies" the narrative, followers exit in panic, and the price plunges further. The lack of a traditional "anchoring value" (such as net asset

value or stable cash flows) for crypto assets makes the amplitude of this reflexive loop far greater than in traditional asset markets, further intensifying price overshoot.

13.7.4 The superposition of four types of obstacles

The four types of obstacles above—liquidation cascades (mechanism), market manipulation (pollution), behavioral biases (sentiment), and information friction (infrastructure delay)—do not exist in isolation. Under extreme conditions, they often interweave and superimpose layer upon layer, leading to a systemic failure of the price discovery mechanism.

Figure 13-32 maps the four types of obstacles—liquidation cascades, market manipulation, behavioral biases, and information friction—onto the three stages of expression, aggregation, and transmission, marking the link each type of obstacle most directly damages and its cross-stage superimposed amplification.

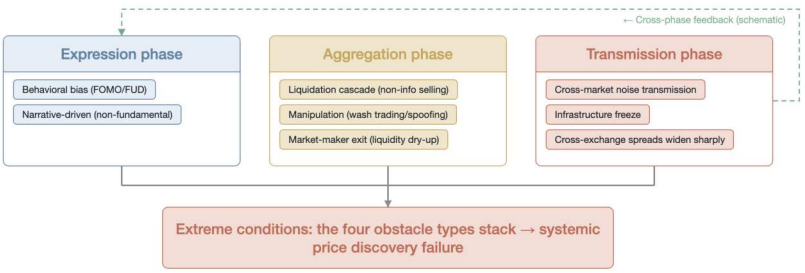


Figure 13-32. Mapping of the four types of price discovery obstacles onto the three-stage model

Figure 13-32 reveals that the four types of obstacles are not evenly distributed across the three-stage model: behavioral biases and narrative-driven trading act mainly on the expression stage, where large amounts of non-informational sentiment orders pour into the market and dilute the concentration of informed traders' signals; liquidation cascades and market manipulation (wash trading, spoofing) concentrate their damage on the aggregation stage—the former swamping market makers' signal-extraction capacity with mechanical selling pressure, the latter directly polluting market makers' Bayesian-inference process with false order flow; and information friction (network congestion, withdrawal freezes, infrastructure degradation) severs the arbitrage channels of the transmission stage, preventing the prices of different venues from converging. More importantly, the figure uses dashed lines to mark the cross-stage feedback links among the obstacles: for example, the price overshoot produced by a liquidation cascade in the aggregation stage spreads through social media and triggers more panic orders in the expression stage, forming a closed positive feedback loop.

Taking the October 10 event of 2025 as an example, one can clearly trace how the multiple obstacles reinforced one another at each stage. The crisis began with the enormous uncertainty in interpreting an external macro shock (a 100% tariff threat); market participants held serious disagreements about the meaning of this information, and in the

crypto-market environment of 24/7 trading without circuit breakers, this uncertainty was instantly priced in as violent price volatility. The initial decline rapidly triggered the forced liquidation of highly leveraged positions, the double-erosion effect of unified margin accelerated the depletion of margin, and the liquidation engine's mechanical selling overwhelmed market makers' quoting capacity, so that liquidity evaporated sharply; at the same time, exchange interfaces froze due to the surge in traffic, traders could not add margin or manage risk, and infrastructure friction and the liquidation cascade formed a severe superposition.

Panic sentiment spread rapidly through social media, and retail panic selling and mechanical liquidation converged into large-scale concentrated selling. Under this extreme stress, the cross-market transmission mechanism broke down: arbitrageurs could not close price differences because of network congestion and capital constraints, causing USDe to trade at a discount of as much as 35% on Binance while remaining near $1 on other exchanges and in DeFi pools [26]. This extreme pricing deviation had nothing to do with fundamentals and was entirely a product of venue microstructure, yet it in turn triggered, through Binance's margin system, more liquidations that should not have occurred, forming a vicious loop in which noise was amplified rather than attenuated in transmission.

This extreme USDe discount event exposed a deeper systemic risk in the unified-margin system: the contagion effect of the credit risk of the margin asset. When USDe traded at a 35% discount on Binance, all positions using USDe as margin suddenly faced insufficient margin—even if the price movement of the underlying assets of those positions had not yet reached the liquidation line, the shrinkage of the margin asset's own value had already eroded the margin adequacy from another dimension. This mechanism is essentially isomorphic to the double-erosion effect discussed earlier, but its contagion range is broader: the double-erosion effect affects positions using the underlying asset itself as collateral, whereas margin-asset credit risk affects all heterogeneous position portfolios that use that asset as margin. Under the architecture of the unified-margin system, the "price stability" of the margin asset is an implicit core assumption: margin is usually valued at face value or at a fixed discount close to face value, rather than being marked to market in real time. Once this assumption is broken, the reliability of the entire margin-calculation system is fundamentally called into question. Even more severe, if multiple stablecoins simultaneously trade at significant discounts in an extreme event—for example, when systemic credit panic spreads across the entire stablecoin ecosystem—then positions using different stablecoins as margin may face insufficient margin at the same time, and the margin system will face the risk of cascading collapse. This analysis links directly to the discussion of insurance funds in Chapter 12: if an insurance fund's reserve assets themselves include stablecoins, then in a scenario of systemic stablecoin discounting, the insurance fund's actual usable value will shrink in tandem, losing its buffering capacity at precisely the moment it is most needed.

In this cascade process, where multiple obstacles reinforce one another, price discovery is no longer a matter of being "inefficient"; it has completely lost its function. The price of the perpetual futures market no longer reflects any information about the asset's future value but merely reflects the degree of liquidity exhaustion and mechanism fragility of that exchange at that moment. In the language of the ceiling-floor model, efficiency has

at this point reached the floor, and may even briefly break through the floor, because the superposition effect of multiple obstacles exceeds the expected destructive power of any single mechanical-noise pathway. This is precisely the extreme manifestation of the "widest range" of perpetual futures in the model of Section 13.6: when the four types of obstacles are activated simultaneously, the fall of efficiency from ceiling to floor can be completed within minutes.

13.8 Price discovery on on-chain order book DEXs

High-performance on-chain central limit order book decentralized exchanges, represented by Hyperliquid and dYdX v4, are reshaping the price discovery environment of digital asset markets at a structural level. This evolution is not a simple migration of trading venues but a fundamental transformation of the information-processing mechanism. On-chain transparency brings dual effects—information symmetrization and the maximal extractable value (MEV) problem—while the three types of native information channels unique to crypto markets provide a wholly new dimension for price discovery that does not exist in traditional finance. The maturity of decentralized sequencers and MEV-protection mechanisms will ultimately determine whether the price discovery function can migrate to the on-chain environment on a larger scale.

13.8.1 The changing price discovery environment

For a long time, centralized exchanges dominated the price discovery process of crypto markets. Yet the CEX architecture has an inherent problem of opacity. In a CEX, the entire price discovery process is completed within centralized servers, and external observers can only passively receive the final price and order book snapshots pushed via API, unable to observe the complete life cycle from order submission and matching to execution. This black-box state requires users to bear the counterparty risk of asset custody and makes it impossible to verify cryptographically whether the exchange is using its information advantage to front-run or manipulate internally. The collapse of FTX (November 2022) fully exposed the fragility of this trust model: user assets were misappropriated, the internal trading desk enjoyed an asymmetric information advantage, and all of this was revealed only after the fact.

The rise of on-chain CLOB-DEXs is precisely a structural response to this problem of opacity. By placing order book matching and settlement logic entirely on a blockchain or a customized application chain, the life cycle of a trade becomes fully verifiable. From order submission, sequencing, and matching to final settlement, every operation is recorded on an immutable distributed ledger. Users' assets are always held in smart contracts under their own control (self-custody), fundamentally eliminating the risk of a centralized custodian misappropriating assets. Self-custody, however, while eliminating one class of risk, introduces another: the smart contract itself may harbor security hazards such as code vulnerabilities, maliciously exploited upgrade mechanisms, or governance attacks, and the resulting asset losses are often technically irreversible. Even more noteworthy, DEXs are not truly immune to CEX-like centralized intervention in extreme events. The Hyperliquid JELLY event of March 2025 provides a vivid case: a whale trader, by maliciously manipulating the perpetual futures position of

the JELLY token, placed the protocol's HLP (Hyperliquid Liquidity Provider) vault at risk of enormous losses. Facing this emergency, the Hyperliquid team coordinated validators to shut down that specific market by emergency vote and forcibly settled all open positions at a designated price, with the HLP vault bearing the socialized losses. This response is essentially no different from the risk-control intervention of a centralized exchange. Similarly, dYdX has, in extreme market events, adjusted the risk parameters of specific markets through governance votes. These cases show that the "decentralization" and "censorship resistance" that DEXs tout may, when systemic risk threatens the protocol's survival, be overridden by centralized decisions at the governance layer. A more careful formulation is therefore: on-chain DEXs change the nature and distribution of risk rather than unilaterally eliminating it. This shift means that market participants move from relying on the self-discipline of centralized institutions to playing an open game with all participants under transparent but not risk-free rules.

Figure 13-33 compares the structural differences between CEXs and on-chain CLOB-DEXs across dimensions such as order processing, asset custody, and verifiability.

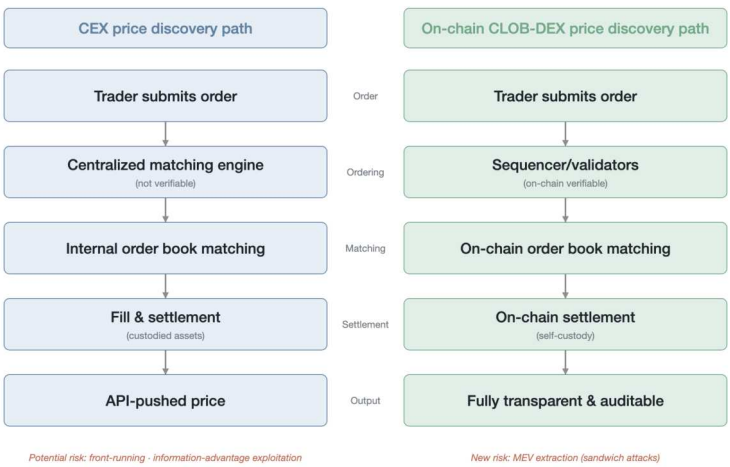


Figure 13-33. Comparison of price discovery mechanisms between CEXs and on-chain CLOB-DEXs

Figure 13-33 uses side-by-side structural diagrams to compare the process differences between CEXs and on-chain CLOB-DEXs across five links: order submission, sequencing, matching, settlement, and asset custody. In the CEX architecture, all links are completed in the black box of centralized servers, and external observers can see only the final output of the execution price and order book snapshots; whereas in the on-chain CLOB-DEX architecture, the execution record of every link is written to an immutable blockchain ledger, making the complete life cycle from order to execution cryptographically verifiable. The difference between the two architectures on the asset-custody dimension is especially striking: CEXs require users to transfer assets into addresses controlled by the exchange (the custodial model), whereas in a DEX users'

assets are always held in smart contracts under their own control (the self-custodial model).

This paradigm shift is clearly borne out in trading data. According to a report released by CoinGecko, the trading volume of decentralized perpetual futures grew explosively in 2025. For the full year 2025, the top ten perpetual futures exchanges (including both CEXs and DEXs) cumulatively processed $92.9 trillion in trading volume, up 64.6% year over year. Of this, DEX perpetual futures trading volume surged 346%, reaching $6.7 trillion [34]. Even more striking, the ratio of DEX perpetual futures trading volume to CEX perpetual futures trading volume climbed all the way from 2.1% in January 2023, reaching a record high of 11.7% in November 2025, achieving 14 consecutive months of growth [34]. By the end of 2025, the DEX perpetual futures market had stably occupied over 10% of market share, marking that on-chain DEXs have evolved from a marginal "alternative choice" into a core price discovery venue that cannot be ignored.

Figure 13-34 uses a monthly line chart to track the trajectory of DEX perpetual futures trading volume as a share of CEX perpetual futures trading volume over the full year 2025, showing that this ratio accelerated upward in the second half of the year.

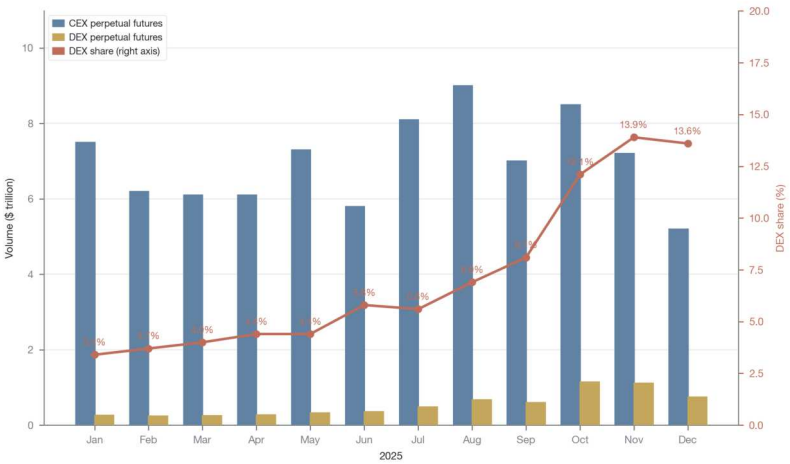


Figure 13-34. The evolution of on-chain DEX perpetual futures trading volume as a share of CEX volume (2025; the share is the ratio of monthly DEX-perpetual to CEX-perpetual trading volume, with a monthly peak of about 13.9% in November; the 11.7% cited in Section 13.8.3 is the value for the same month under a different convention, the difference arising from the monthly ratio versus a cumulative convention and whether the denominator includes DEX; monthly volume and share per CoinGecko [34], market landscape also see [35])

This growth was not evenly distributed. DEX perpetual futures trading volume in the second half of 2025 accounted for about 65% of the full-year total, exhibiting a clear acceleration, with single-month trading volume in the fourth quarter repeatedly surpassing $1 trillion. This trend indicates that, as the performance of on-chain infrastructure continues to improve and user experience continues to be optimized, informed traders are migrating their price discovery activity from CEXs to DEXs at an ever-faster

pace. The deeper driver of this migration is not only the improvement in technical performance but also the structural shift in market preference toward self-custody and transparent settlement after the FTX incident.

13.8.2 The dual effects of transparency: information symmetrization and the MEV problem

The absolute transparency brought by on-chain order books has profound and complex dual effects on the price discovery process. On the one hand, transparency achieves unprecedented information symmetrization. In the on-chain environment, all operations are auditable, fundamentally eliminating the possibility of the exchange itself front-running or covertly manipulating. This transparency forces market makers to provide better quotes, because any quote that deviates significantly from fair value would be quickly captured and exploited by on-chain arbitrageurs. In addition, on-chain-observable large positions and liquidation thresholds change the game structure among market participants. For example, on platforms such as Hyperliquid, all users can calculate the liquidation price of a whale account from on-chain data—information strictly kept secret in a CEX becomes a public game variable in a DEX. This symmetrization of information, in theory, helps reduce adverse-selection costs, thereby narrowing the bid-ask spread and improving the overall liquidity quality of the market.

On the other hand, transparency also gives rise to a severe maximal extractable value problem. Because trade intent is publicly visible before being packed into a block, MEV searchers can identify large trades by monitoring the mempool or the sequencer queue and execute ahead of them by paying a higher priority fee or colluding with block builders. The most typical form is the sandwich attack: the searcher inserts their own buy order ahead of the target's large buy order to push up the price, and after the target order fills at the higher price, immediately sells for a profit. According to EigenPhi's data, although the amount extracted by sandwich attacks on Ethereum fell from about $10 million per month at the end of 2024 to about $2.5 million in October 2025, the attack frequency remained at a high level of 60,000 to 90,000 times per month [36]. This mechanism directly harms the interests of the original trader and increases the hidden friction cost of trading.

Yet from the pure efficiency perspective of price discovery, the impact of MEV is not entirely negative. The study by Capponi, Jia, and Yu (2025) reveals a counterintuitive conclusion: DEX trades that pay high priority fees actually contain more private information and contribute more to price discovery [37]. To ensure that their time-sensitive orders carrying information are executed quickly, informed traders often adopt a "jump-bidding" strategy, deterring competitors by setting an extremely high initial fee. In this sense, the MEV mechanism actually accelerates the incorporation of information into price, because it provides a market-based pricing mechanism for the time value of information. But this gain in efficiency comes at the cost of market fairness, externalizing execution costs onto ordinary users. The net effect of MEV on overall market quality therefore remains a contested frontier topic in current microstructure research.

13.8.3 The three native information channels of crypto markets

In traditional financial markets, price discovery relies mainly on public information such as macroeconomic data and corporate earnings, along with the private research of institutional investors. In crypto markets, however, the on-chain environment has bred three native information channels found nowhere else, which intervene in the three-stage price discovery model (expression, aggregation, and transmission) in different ways, providing a wholly new information dimension for asset pricing.

These three native information channels each have distinctive information attributes and temporal characteristics, and their intervention points in the three-stage model also differ. Figure 13-35 summarizes the information sources, typical signal types, and main stages of action for each of the three channels: on-chain data, mempool data, and derivatives data.

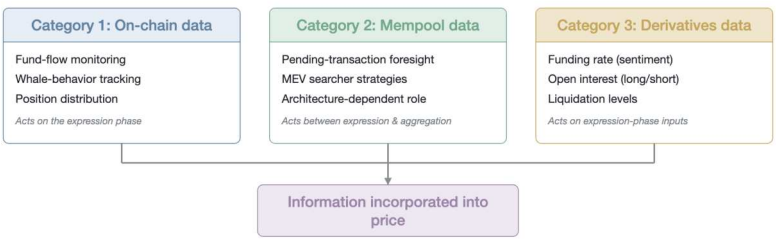


Figure 13-35. The three native information channels of crypto markets and their roles in price discovery (conceptual classification illustration, no quantitative data)

The first is the on-chain data channel, which acts directly on the "expression stage" of price discovery as a foundational information source. Such data include monitoring stablecoin flows between exchanges and on-chain wallets, the sudden activation of long-dormant whale wallets, and large asset transfers. For example, when hundreds of millions of dollars of USDT are transferred from an exchange hot wallet to cold storage, it often signals that a large holder is accumulating a spot position, a signal that releases directional information to traders with on-chain analytical capability before the actual buying takes place. In addition, analysis of a token's holding distribution and concentration can also provide deep insight into market structure. All these on-chain behaviors, before being converted into actual order flow, already constitute a "pre-expression" information signal.

The second is the mempool data channel, which acts mainly on the transitional zone between the "expression stage" and the "aggregation stage." Transactions awaiting confirmation in the mempool provide a brief foresight into future market states. Informed MEV searchers use this foresight to design front-running strategies. From an information-theoretic perspective, although this behavior is predatory, it objectively accelerates the transmission of the supply-and-demand information embedded in the mempool to the final execution price. The role of the mempool differs significantly

across DEX architectures. Designs using a centralized sequencer (such as Hyperliquid) usually do not expose a public mempool, thereby reducing the risk of sandwich attacks, but this brings with it the problem of centralization of sequencing power. In a fully decentralized architecture (such as dYdX v4), by contrast, validators can observe pending transactions, which creates conditions for validator-level MEV extraction.

The third is the derivatives data channel, which serves as a key information input to the "expression stage," feeding back into and influencing traders' decisions. This includes the funding rate, changes in open interest, and the distribution of liquidation levels. As described in the preceding sections, the funding rate is not only the mechanism that keeps the perpetual futures price anchored to spot but is itself a second-order information indicator reflecting market sentiment and leverage crowding. In on-chain DEXs, the liquidation thresholds of large positions are fully transparent, which makes "liquidation sniping" a trading strategy based on deterministic information, whereas in a CEX such a strategy can rely only on vague probabilistic guesses. A sudden surge or plunge in open interest directly reveals that market participants are building or unwinding directional exposure.

The common feature of these three native information channels is that they are, in technical principle, open to all participants, greatly lowering the absolute threshold of information access. Yet owing to the complexity of information processing and the intensity of competition, the window of informational advantage is compressed extremely, shrinking from the days or weeks of traditional finance to seconds or even a few blocks in crypto markets. The focus of competition has shifted entirely from "who can obtain information" to "who can process and execute information faster and more accurately"—the concrete manifestation, in the domain of price discovery, of the "transparency paradox" explored in Chapter 1.

13.8.4 A comparison of two architectures

The on-chain derivatives track currently exhibits a coexistence of multiple technical architectures. Among them, Hyperliquid and dYdX v4 represent two entirely different design philosophies, and their differences in consensus mechanism, sequencer design, and verifiability directly shape the microstructure characteristics of price discovery in their respective markets. Table 13-9 systematically compares the key architectural parameters of the two.

Dimension	Hyperliquid	dYdX v4
Underlying architecture	Dedicated Layer 1 (HyperCore + HyperEVM)	Cosmos SDK application chain
Consensus mechanism	HyperBFT (custom-optimized)	CometBFT (evolved from Tendermint)
Transaction finality	Median ~0.2 seconds, P99 < 0.5 seconds	~1–2 seconds
Throughput	~200,000 TPS	~10,000 TPS
Sequencer design	Centralized sequencer + on-chain verification	Decentralized validator-rotation sequencing
Gas fees	Zero gas fees	Gas fees required
MEV risk profile	Sequencer-level (concentrated but auditable)	Validator-level (dispersed but hard to coordinate defenses against)
Cross-chain interoperability	HyperEVM bridging	Native Inter-Blockchain Communication (IBC) interoperability
Verifiability	Matching results verifiable on-chain	Entire process verifiable on-chain

Table 13-9. Comparison of Hyperliquid and dYdX v4 architectural features (Data source: compiled by the author from each platform's public documentation)

Hyperliquid has chosen a "performance-first" infrastructure route. It has built a dedicated Layer 1 blockchain and adopted a custom HyperBFT consensus algorithm optimized specifically for end-to-end latency. This architecture enables Hyperliquid to achieve transaction finality with a median of about 0.2 seconds and throughput of up to 200,000 transactions per second (TPS) [38]. In terms of sequencer design, Hyperliquid currently uses a centralized sequencer to ensure extremely low latency and a zero-gas-fee experience, but all matching and settlement results are fully verifiable on-chain. This hybrid model, however, requires a precise distinction regarding the boundary of "verifiability." What is verifiable on-chain is matching correctness—that is, whether, given a set of already-sequenced orders, the matching engine executes correctly according to price-priority and time-priority rules. But ordering fairness—that is, whether orders are queued fairly on a first-come-first-served basis after arriving at the sequencer—is not cryptographically verifiable under the current architecture. A centralized sequencer can entirely extract value by selectively delaying or prioritizing specific orders without violating any matching rule, and such behavior cannot be identified after the fact in the on-chain record. The aforementioned JELLY event (March 2025) exposed, from another dimension, the decision-power-concentration risk of the centralized-sequencer architecture: when a whale account's malicious manipulation threatened the protocol's systemic safety, Hyperliquid needed only a few validators (about 5 controlled by the foundation) to form a quorum to unilaterally shut down a specific market and forcibly settle all positions within minutes. This fact shows that although Hyperliquid's validator network in theory provides decentralized verification, its actual emergency decision power is highly concentrated in a few nodes. From a systemic-risk perspective, as of the end of 2025 about 43% of DEX perpetual futures trading volume flowed through this single protocol and its centralized sequencer, and this degree of concentration itself constitutes a non-negligible systemic-risk exposure.

Despite the structural concerns above, this hybrid model of "centralized sequencing + decentralized verification" has achieved notable results on liquidity metrics. In BTC perpetual futures, Hyperliquid can maintain a bid-ask spread of about $1, with cumulative sell-side depth of about 140 BTC on its order book, even surpassing the depth level of about 80 BTC at some top CEXs over the same period. Several important qualifications must be attached to this liquidity comparison, however. First, the HLP (Hyperliquid Liquidity Provider) vault is a market-making capital pool operated by the Hyperliquid platform itself rather than an independent third-party market maker, and the liquidity it provides is different in nature from that provided by the diversified, competitive network of market makers on a CEX. The share of the HLP vault's contribution to the aforementioned 140 BTC of depth is a key parameter for assessing the sustainability and stress resilience of this liquidity, but it currently lacks transparent public disclosure. Second, the comparison above is based on an order book snapshot at a single point in time and fails to reflect time-weighted average depth or to measure depth resilience during high-volatility periods—yet it is precisely at these critical moments that the true quality of liquidity matters most. Third, a significant difference may exist on a DEX between displayed depth and executable depth: in the on-chain environment, orders can be rapidly withdrawn within a single block, so the depth displayed at a given moment may not truly be executable. Finally, the incentive mechanism of the HYPE token attracts a large amount of incentivized liquidity through market-making rewards; the providers of such liquidity are driven mainly by token incentives and may withdraw rapidly when incentives are reduced or removed, so it should not be equated with sustainable, organic liquidity driven by endogenous market-making profits.

Figure 13-36 compares the order book liquidity depth of Hyperliquid and Binance in BTC perpetual futures.

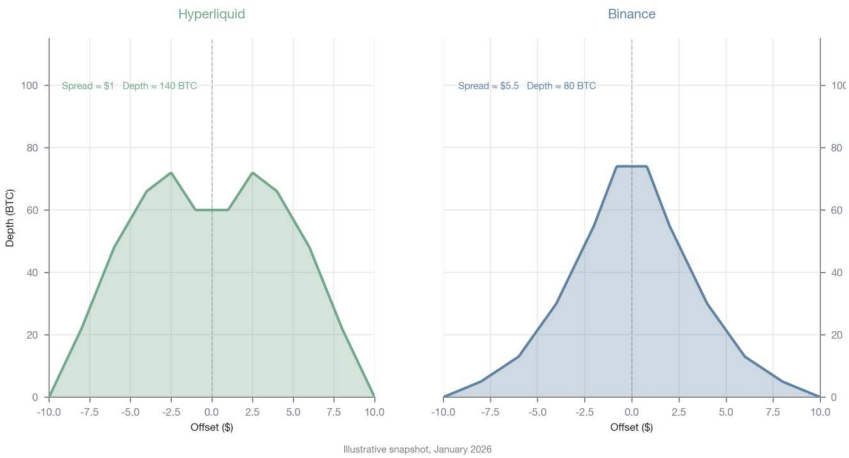


Figure 13-36. Illustrative comparison of Hyperliquid and Binance BTC perpetual futures liquidity depth (January 2026 snapshot; the HLP vault's market-making strategy causes displayed depth to diverge from executable depth; depth data not independently verified)

Figure 13-36 uses cumulative order book depth curves to compare the bid and ask dis-

tribution of Hyperliquid and Binance in BTC perpetual futures at the same point in time. The horizontal axis is the percentage deviation from the mid-price, and the vertical axis is the cumulative BTC order quantity. Within the very narrow $\pm 0.1\%$ spread range, Binance, by virtue of the fierce competition among multiple high-frequency market makers, maintains higher top-of-book liquidity; but in the wider range of $\pm 0.5\%$ to $\pm 1\%$, Hyperliquid's cumulative sell-side depth (about 140 BTC) instead exceeds Binance's level over the same period (about 80 BTC). This seemingly counterintuitive phenomenon is related to the HLP vault's market-making strategy—its order-placement strategy tends to deploy greater depth at positions slightly farther from the mid-price to reduce adverse-selection risk.

By contrast, dYdX v4 has chosen a fully decentralized application-chain route based on Cosmos SDK and CometBFT consensus. Its core feature is that a decentralized validator network takes turns responsible for order sequencing and block production. This design maximally guarantees censorship resistance and decentralization attributes, eliminating the systemic risk that a single sequencer might bring. Decentralized consensus, however, inevitably brings a compromise in performance: its transaction finality is usually between 1 and 2 seconds, and its throughput is also relatively limited. In addition, validator-level MEV extraction becomes a problem requiring focused defense in the dYdX v4 architecture, because the validator producing the current block can observe pending transactions and reorder them.

From the microstructure perspective of price discovery, the differences between these two architectures produce entirely different market characteristics. Hyperliquid's ultra-low latency makes it closer to the operating environment of traditional high-frequency trading, where market makers can update quotes more frequently and thereby provide tighter bid-ask spreads and deeper liquidity. This environment is naturally favorable for informed traders to express information quickly, making prices respond faster to new information. The presence of a centralized sequencer, however, means that ordering fairness depends on the operator's reputation rather than a cryptographic guarantee, which in theory constitutes a trust premise. dYdX v4's decentralized sequencing, while sacrificing some latency, eliminates the single-point trust dependency through its fully decentralized architecture and holds an advantage in long-term institutional robustness.

The competition between these two architectures for market share vividly illustrates informed traders' dynamic choice of price discovery venue. In the first half of 2025, Hyperliquid, by virtue of its CEX-comparable performance experience and innovative HIP-3 permissionless-listing mechanism, at one point held as much as 75% to 80% market share in the Perp DEX space [39]. As competitors such as Lighter (using zero-knowledge-proof zkRollup technology, zero fees) and Aster (backed by Binance Labs) entered in the second half of the year, however, Hyperliquid's market share experienced violent fluctuations. In September 2025, its share fell at one point to about 38%, while Lighter and Aster held 16.8% and 14.9%, respectively [40]. By year-end, as the market landscape reconsolidated, Hyperliquid's share recovered to about 43%. This fierce venue competition indicates that liquidity in the on-chain DEX space is highly elastic, and the shift of price discovery leadership is faster and more dynamic than in the CEX space.

Figure 13-37 uses a dual-axis chart to show the monthly trading volume composition of the major Perp DEXs in 2025 (Hyperliquid, dYdX, Lighter, Aster, Jupiter, and others) (stacked bars on the left axis) and the trajectory of Hyperliquid's market share (line on the right axis), depicting the evolution from "one dominant player" to "a contest among many."

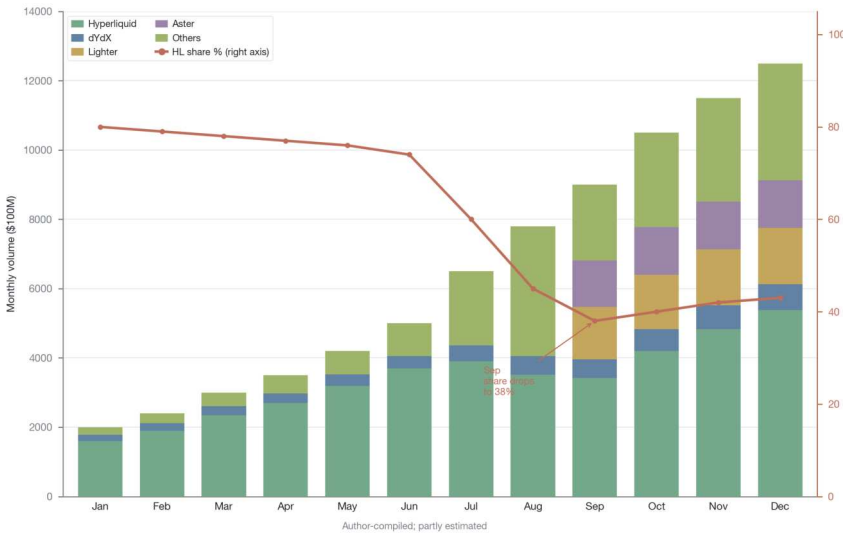


Figure 13-37. Monthly Perp DEX trading volume and the evolution of Hyperliquid's market share (2025; compiled by the author, partly estimated; the month-by-month composition is pending verification against data snapshots)

The data in Figure 13-37 exhibit two salient features. First, the total monthly trading volume of Perp DEXs grew from about $200 billion in January 2025 to over $1.2 trillion in December, a full-year increase of about fivefold, with the growth curve clearly steepening in the second half of the year. Second, Hyperliquid's market share experienced violent fluctuations, from a near-monopoly of about 75%–80% in the first half of the year, to a plunge to about 38% in September as Lighter and Aster entered, and back to a recovery to about 43% by year-end. This rapid redistribution of share is extremely rare in the CEX market—Binance's dominant position in CEX perpetual futures has remained highly stable over the past several years. The low friction of liquidity migration in the on-chain DEX space (no need for cross-exchange transfers, no withdrawal delays) enables informed traders to reallocate their trading activity across different platforms at extremely low switching cost; this feature endows the DEX ecosystem with stronger competitive vitality but also means that any single platform's price discovery leadership is more fragile.

13.8.5 Sequencers and MEV protection

For on-chain DEXs to take on the price discovery function of crypto markets at a larger scale, they must, while maintaining high performance, solve the two core pain points of sequencer centralization and malicious MEV extraction. Current technical evolution focuses mainly on three directions, and these innovations will profoundly change the

microstructure mechanism by which information is incorporated into price.

The first direction is the proliferation of decentralized sequencers. Dispersing the power of transaction sequencing from a single entity to multiple participant nodes is a key path to eliminating single points of failure and internal front-running risk. The industry is currently exploring several technical routes. For example, the fair sequencing service proposed by Chainlink uses a decentralized oracle network to determine the fair order of transactions; its core idea is to reach consensus on transaction arrival times through multiple independent nodes, thereby preventing any single node from manipulating the ordering [41]. Threshold-encryption-based schemes, by contrast, require transactions to remain encrypted before being sequenced, and to be decrypted only through the cooperation of distributed key holders after being included in a block, thereby fundamentally depriving the sequencer of the ability to use transaction content to front-run. This "sequence first, decrypt later" paradigm provides a theoretically optimal solution for building cryptographic-level ordering fairness, but its actual deployment faces severe engineering challenges. Current threshold-encryption schemes require multiple key holders to perform cooperative computation in the decryption link, a process that usually adds several hundred milliseconds of end-to-end latency. For Hyperliquid, which pursues transaction finality of about 0.2 seconds, this additional latency would almost mean a doubling of the finality time, materially affecting the information-expression efficiency of high-frequency market-making strategies and informed traders. As of the end of 2025, no mainstream perpetual futures DEX had actually deployed a threshold-encryption scheme in a production environment. It is therefore necessary to distinguish clearly: threshold encryption remains a research direction with theoretical appeal rather than a production-ready solution, and a considerable distance still separates theoretical argument from large-scale engineering deployment.

The second direction is system-level MEV-protection mechanisms. To protect ordinary users without fully sacrificing the price-discovery-acceleration effect that MEV brings, various private transaction pools and MEV-protection networks have emerged. These mechanisms allow users' orders to bypass the public mempool and be sent directly to trusted block builders, thereby avoiding capture by sandwich attacks. Further exploration includes "MEV redistribution" mechanisms, which aim to return part of the MEV profit extracted by searchers to the affected traders or liquidity providers, seeking a new balance between efficiency and fairness. From the perspective of price discovery, the ideal effect of such mechanisms is to retain the part of MEV that benefits information transmission (i.e., accelerating the execution of informed orders) while eliminating its purely predatory part (i.e., the exploitation of uninformed traders by sandwich attacks).

The third direction is the application of zero-knowledge-proof constraints in matching engines. Architectures represented by Lighter.xyz move the complex order book matching and liquidation logic to off-chain execution and submit only the generated zero-knowledge proofs to the Ethereum mainnet for verification [42]. In this model, the behavior of the sequencer and matching engine is strictly constrained by cryptography, and any sequencing that deviates from the predetermined rules (such as price priority and time priority) cannot generate a valid proof, thereby guaranteeing the fairness of the matching process at the mathematical level. The profound significance of this

scheme is that it places the most core "aggregation stage" of the price discovery process under cryptographic protection, enabling on-chain DEXs to simultaneously achieve high performance (off-chain execution) and strong trust (on-chain verification). Lighter, with $1.3 trillion in trading volume, ranked among the top ten global perpetual futures exchanges in 2025 [34], preliminarily validating the feasibility of this architecture in actual market competition.

Taken together, these three directions are not mutually exclusive but can act synergistically. An ideal future architecture might simultaneously adopt decentralized sequencers to disperse sequencing power, threshold encryption to protect transaction privacy, and zero-knowledge proofs to constrain matching behavior. When these technologies mature and are deployed at scale, on-chain DEXs will have the potential to surpass CEXs simultaneously along the three dimensions of transparency, fairness, and performance, thereby truly becoming the dominant venue for price discovery in crypto markets.

13.9 Chapter summary

Across the chapter as a whole, the perpetual futures market simultaneously plays a dual role in digital asset price discovery: accelerating information and injecting noise. This inherent tension constitutes the core thread for understanding the information efficiency of crypto markets.

Under normal market conditions, perpetual futures are an efficient price discovery instrument. Based on the venue gravity model (Section 13.3), their systematic advantages across the four dimensions of leverage, liquidity depth, anonymity, and transaction cost make them the venue of choice for informed traders to express private information, and they exhibit extremely high efficiency in the "expression stage" of the three-stage model. Empirical data repeatedly confirm (Section 13.4) that in the normal market state, perpetual futures dominate both the Hasbrouck information share and the Gonzalo-Granger component share. The funding rate mechanism (Sections 13.3.4 and 13.5.2) ensures the long-term anchoring of the perpetual price to spot through continuous arbitrage incentives, so that the price information discovered in perpetual futures is transmitted effectively to the market at large.

Yet the market mechanism of perpetual futures simultaneously produces significant interference with the price signal, injecting non-informational mechanical noise through three pathways (Section 13.5): the microstructural impact of forced liquidation by the liquidation engine (σ_{mech}), the distortion of positions by funding rate extremes, and the delay caused by mark-price EMA smoothing. Under extreme conditions (Section 13.6), the three pathways are active simultaneously and superimpose on one another, producing enormous mechanical buying and selling pressure that causes prices to deviate systematically from fundamentals and rapidly contaminates spot and other derivatives markets through oracle price feeds and cross-market arbitrage. In the crisis state (Section 13.7), the direction of reconfiguration of the information flow topology depends on the information structure of the crisis (Sections 13.3.5 and 13.4.6): in a pure mechanical crisis (a liquidation cascade with no major new information), perpetual futures are swamped by mechanical noise, and the spot market may passively take over lead-

ership; in an information-mechanical compound crisis (major news superimposed on a liquidation cascade), informed traders' need to express information keeps the leading derivatives market at the center of information. In either case, blindly trusting the perpetual price of a single market brings enormous risk-control risk.

Through the ceiling-floor model, we can comprehensively evaluate this dual nature. Compared with the spot market and the traditional futures market, the perpetual futures market exhibits the distinctive characteristic of "the highest ceiling, the highest floor, and the widest efficiency range." The ceiling of perpetual futures is the highest, meaning that they are the most efficient price discovery venue in normal times; but their floor is also the highest (that is, their efficiency loss in the worst case is the most severe), and the range between the ceiling and the floor is the widest, meaning that the volatility of their price discovery efficiency is far greater than that of the spot and traditional futures markets.

Figure 13-38 distills the core results of this chapter into a comprehensive comparison chart, with the three market types on the horizontal axis and price discovery efficiency on the vertical axis, marking each market's ceiling, floor, typical operating range, and the landmark events of efficiency plunge in historical crises.

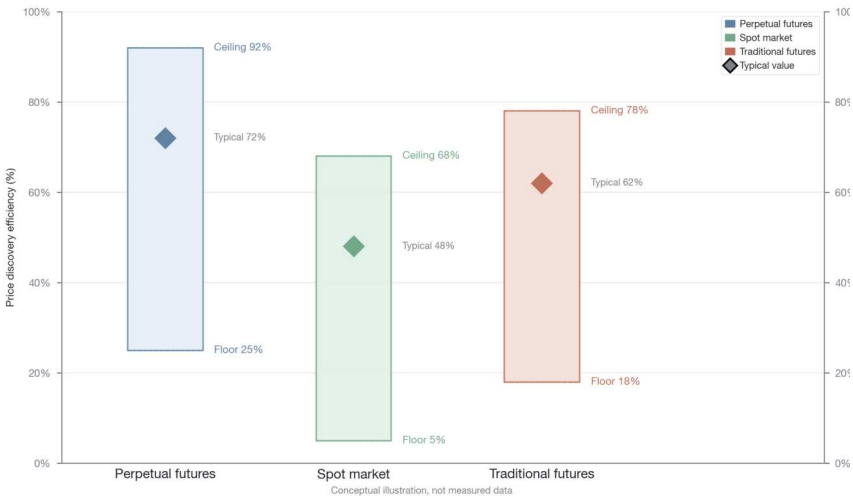


Figure 13-38. A comprehensive ceiling-floor assessment of price discovery efficiency across three market types (conceptual illustration, not empirical data; value conventions consistent with Figures 13-26 and 13-27)

Figure 13-38 condenses the core conclusions of the whole chapter into an intuitive visual language. Traditional futures exhibit a "narrow-band steady-state" efficiency profile—a medium ceiling and a low floor, with a range between them that is narrow and stable, like a river of modest amplitude. Crypto spot exhibits a "low ceiling, low floor" profile—information expression is limited by the lack of leverage, but for the same reason it is almost undisturbed by mechanical noise, like a calm but limited pool of water. Perpetual futures are entirely different: their efficiency range is like rapids in a canyon, with a highest water level (the ceiling) far exceeding that of the other two but a lowest water level (the floor) also higher than the floors of the other markets,

and the amplitude of the rise and fall is astonishing. These three markedly different efficiency profiles together constitute a complete portrait of the current price discovery ecosystem of crypto markets.

The price discovery efficiency of perpetual futures is therefore not simply "superior" or "inferior" to that of other markets but a highly state-dependent dynamic characteristic. In assessing the price discovery contribution of perpetual futures, one must simultaneously consider the current market state (volatility level, leverage crowding, liquidity depth) and the specific time scale.

The advantages and disadvantages of perpetual futures stem from the same root—leverage. The high-leverage mechanism, while accelerating information expression, also sows the seeds of liquidation cascades. Leverage amplifies informed traders' capacity for information expression, enabling a small amount of capital to produce a significant price impact, which is the core reason perpetual futures have become the price discovery leader. But it is also leverage that, when the market direction reverses, triggers the chain reaction of forced liquidation, producing a mechanical price impact unrelated to fundamental information. In the same way, 24/7 continuous trading lets information be priced continuously, but it also removes the buffers and shock absorbers of traditional financial markets, such as circuit-breaker halts, so that noise propagates without temporal interruption. Market participants cannot enjoy the extremely high expression efficiency of perpetual futures while being fully exempt from the interference of their mechanical noise. This inseparability is the essential feature of the microstructure of perpetual futures markets.

Having clarified the inherent laws and defects of price discovery as a natural process, the next chapter (Chapter 14) turns the perspective to the level of institutional intervention. Since the price signal of perpetual futures inevitably mixes in mechanical noise, how can market architects filter this noise, guard against market manipulation, and provide a reliable price benchmark for the entire crypto derivatives ecosystem by constructing infrastructure such as the index price, the mark price, and oracles? This is precisely the logical progression from "price discovery" to "price benchmark": the former describes how information is spontaneously incorporated into price, while the latter explores how institutions actively purify and anchor price.

References

[1] Chainalysis. (2025). *Bybit hack: Leveraging transparency for collaboration in the wake of record-breaking theft*. https://www.chainalysis.com/blog/bybit-exchange-hack-february-2025-crypto-security-dprk/

[2] Liu, A. (2025, February 21). Bybit's $1.4 billion ETH hack: Market impact and future implications. *Forbes*. https://www.forbes.com/sites/aliceliu/2025/02/21/bybits-14-billion-eth-hack-market-impact-and-future-implications/

[3] Kaiko Research. (2025). *Bybit hack by the numbers*. https://research.kaiko.com/insights/bybit-hack-by-the-numbers

[4] Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *The Journal of Finance*, *50*(4), 1175–1199. https://doi.org/10.111

1/j.1540-6261.1995.tb04054.x

[5] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

[6] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100. https://doi.org/10.1016/0304-405x(85)90044-3

[7] Silantyev, E. (2019). Order flow analysis of cryptocurrency markets. *Digital Finance*, *1*, 191–218. https://doi.org/10.1007/s42521-019-00007-w

[8] Easley, D., López de Prado, M. M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *The Review of Financial Studies*, *25*(5), 1457–1493. https://doi.org/10.1093/rfs/hhs053

[9] Alexander, C., Choi, J., Park, H., & Sohn, S. (2020). Price discovery in Bitcoin: The impact of unregulated markets. *Journal of Financial Stability*, *50*, 100776. https://doi.org/10.1016/j.jfs.2020.100776

[10] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, *135*(2), 293–319. https://doi.org/10.1016/j.jfineco.2019.07.001

[11] Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35–55.

[12] CoinGlass. (2025). *CoinGlass 2025 crypto derivatives market annual report*. https://www.coinglass.com/learn/2025-annual-report-en

[13] Ruan, Q., & Streltsov, A. (2024). *Perpetual futures contracts and cryptocurrency market quality*. Cornell SC Johnson College of Business.

[14] CryptoQuant. (2026, March). Binance futures/spot ratio hits 1.5-year high. https://cryptoquant.com/insights/quicktake/69b158571d8c7d419887fc10-Binance-FuturesSpot-Ratio-Hits-15-Year-High

[15] Jung, J. H. (2024). *Can funding rate predict price change?* Presto Research. https://www.prestolabs.io/research/can-funding-rate-predict-price-change

[16] Gonzalo, J., & Granger, C. W. J. (1995). Estimation of common long-memory components in cointegrated systems. *Journal of Business & Economic Statistics*, *13*(1), 27–35. https://doi.org/10.1080/07350015.1995.10524576

[17] Putniņš, T. J. (2013). What do price discovery metrics really measure? *Journal of Empirical Finance*, *23*, 68–83.

[18] Cosenza, R., & Stalder, S. (2024). *Where is the price of Bitcoin determined? Price discovery in a fragmented market* (SSRN Working Paper). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5070964 https://doi.org/10.2139/ssrn.5070964

[19] Entrop, O., Frijns, B., & Seruset, M. (2020). The determinants of price discovery on Bitcoin markets. *Journal of Futures Markets*, *40*(5), 816–837. https://doi.org/10.1002/fut.22101

[20] Dimpfl, T., & Peter, F. J. (2021). Nothing but noise? Price discovery across cryptocurrency exchanges. *Journal of Financial Markets*, *54*, 100584. https://doi.org/

10.1016/j.finmar.2020.100584

[21] Greco, A. (2025, October). October 10–11, 2025: Anatomy of the largest crypto flash crash. *Medium*. https://medium.com/@Alessandro_Greco/october-10-11-2025-anatomy-of-the-largest-crypto-flash-crash-and-what-it-teaches-engineers-risk-49ea96b6d1c2

[22] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[23] Gornall, W., Rinaldi, M., & Xiao, Y. (2024). *Perpetual futures and basis risk: Evidence from cryptocurrency* (AEA Conference Paper). https://www.aeaweb.org/conference/2026/program/paper/ByyFEfr4

[24] CryptoNews. (2026). Bitcoin funding rates just flashed one of the bleakest signals. https://cryptonews.net/news/bitcoin/32527016/

[25] Bybit Help Center. (2025). *Mark price calculation*. https://www.bybit.com/en/help-center/article/Mark-Price-Calculation-Perpetual-Contracts

[26] FTI Consulting. (2025). *The crypto crash of October 2025: When leverage met liquidity*. https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity

[27] Multicoin Capital. (2020, March 17). *March 12: The day crypto market structure broke (Part 1)*. https://multicoin.capital/2020/03/17/march-12-the-day-crypto-market-structure-broke/

[28] Bitwise Asset Management. (2019). *Presentation to the U.S. Securities and Exchange Commission: Analysis of Bitcoin spot trading volume*.

[29] Cong, L. W., Li, X., Tang, K., & Yang, Y. (2023). Crypto wash trading. *Management Science*, *69*(11), 6427–6454. https://doi.org/10.1287/mnsc.2021.02709

[30] Le Pennec, G., Fiedler, I., & Ante, L. (2021). Wash trading at cryptocurrency exchanges. *Finance Research Letters*, *43*, 101982.

[31] Chainalysis. (2025). *Market manipulation: Suspected wash trading in cryptocurrency markets*.

[32] Kraken. (2024). *Crypto FOMO survey: 84% of crypto investors report making decisions driven by fear of missing out*.

[33] Lo, A. W. (2004). The adaptive markets hypothesis. *The Journal of Portfolio Management*, *30*(5), 15–29. https://doi.org/10.3905/jpm.2004.442611

[34] CoinGecko. (2026). *Is the future of crypto perpetual? The meteoric rise of perp DEXs*. https://www.coingecko.com/learn/rise-of-perpetuals-and-perp-dexs

[35] PinkBrains [@PinkBrains_io]. (2025, December). The big picture of perp DEXs in 2025 [Post]. X. https://x.com/PinkBrains_io/status/2030880808257667244

[36] Cointelegraph. (2025, December 4). Exclusive data from EigenPhi reveals that sandwich attacks on Ethereum have waned. https://www.tradingview.com/news/cointelegraph:fa12ba092094b

[37] Capponi, A., Jia, W., & Yu, S. (2025). *Price discovery on decentralized exchanges* (SSRN Working Paper). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4236993 https://doi.org/10.2139/ssrn.4236993

[38] Dropstab. (2025). *How does Hyperliquid's L1 redefine on-chain trading?* https://dropstab.com/research/crypto/what-is-hyperliquid

[39] Binance Square. (2025, September 20). In-depth analysis of perp DEX. https://www.binance.com/en/square/post/29943921066041

[40] Bitget. (2025, September 28). Aster, Hyperliquid, Lighter, EdgeX, and Jupiter Perpetuals. https://www.bitget.com/news/detail/12560604991584

[41] Chainlink. (2024). *Threshold encryption: Security and privacy for Web3.* https://chain.link/article/threshold-encryption-security-privacy-web3

[42] Lighter. (2026). *Order book matching and liquidations with transparent and verifiable zero-knowledge proofs* [White paper]. https://assets.lighter.xyz/whitepaper.pdf

The Governance of Price Benchmarks — Indices, Mark Prices, and Oracles

This chapter examines how price benchmarks shape market microstructure, serving not merely as technical tools but as implicit governance mechanisms. Consider a long position in Ethereum perpetual futures with a notional value exceeding $2 million and 50× leverage. The system set the position's liquidation price at $3,502.00. When the market crashes abruptly, the trading terminal displays five different prices at once: a last traded price of $3,501.50, a mid price of $3,505.00, an index price of $3,505.20, a mark price of $3,505.15, and an oracle price of $3,506.00. The last traded price indicates that the position should be liquidated; the mark price shows that it remains safe; and the oracle price implies an ample margin buffer. In the first instant of the crash, the basis component within the exponential moving average (EMA) has not yet reflected the widening negative basis, which explains why the mark price still hugs the index price rather than deviating downward. The maximum spread among these five numbers is only $4.50, yet in a high-leverage environment this micro-level difference directly determines the survival of a large financial position. The mark price, however, determines only the timing of the liquidation trigger, not the execution price of the liquidation. Under extreme conditions, the liquidation engine closes the position via market orders against an order book whose depth has severely contracted, and the actual execution price may fall far below the mark price at the moment of triggering. This gap between the mark price and the execution price—the execution slippage—often far exceeds the construction error of the mark price itself during systemic sell-offs, and the resulting loss is ultimately borne by the insurance fund and the auto-deleveraging (ADL) mechanism.

During the extreme market conditions of October 2025, the last traded price of Bitcoin on a major perpetual futures platform flash-crashed from approximately $122,500 to approximately $116,400 within 3 minutes, then rebounded to approximately $120,000 within 2 minutes [1]. Over this 5-minute window, the mark price—computed by EMA smoothing over a multi-exchange price index—declined slowly to a trough of approximately $120,300 before beginning to recover. A significant gap of approximately $3,900 separated the trough of the mark price from the trough of the last traded price. This meant that all long positions with liquidation thresholds above approximately $120,300 were spared forced closure. The EMA smoothing mechanism effectively protected these

traders during this transient price spike. Had the mark price been configured with a shorter time window, or had it tracked the last traded price directly, it could have triggered an additional liquidation cascade exceeding $1 billion [1].

Within a 5-minute interval, the choice of a single smoothing parameter determined the fate of billions of dollars in open interest. This parameter is by no means a neutral technical setting; it is a governance decision embedded deep in the exchange's algorithmic architecture and carrying major economic consequences. Traditional finance offers an important precedent for the inherent risks in benchmark design. The scandal surrounding the London Interbank Offered Rate (LIBOR), exposed in 2012, showed that a seemingly objective interest-rate benchmark underpinning roughly $350 trillion in financial contracts had been systematically manipulated by the very institutions responsible for calculating it [2]. In cryptocurrency markets, the average daily trading volume of crypto derivatives has reached the level of hundreds of billions of dollars, and the liquidation logic governing this vast volume depends entirely on the mark price and the index-construction methodology. The fragility of these benchmarks is not merely theoretical: the Mango Markets attack of October 2022 demonstrated how an attacker could extract approximately $116 million within roughly half an hour by manipulating the oracle price in a low-liquidity market [3]. Moreover, during the extreme market volatility of May 19, 2021, position liquidations in the crypto derivatives market exceeded $8 billion within 24 hours, which shows how central benchmarks are to triggering liquidation cascades [4].

This chapter deconstructs the architecture and governance of price benchmarks in cryptocurrency markets. The analysis first establishes a unified framework that decomposes any price benchmark into its computational and institutional specifications. It then examines the governance implications of index construction, the inherent trilemma in mark price design, and the structural bottleneck introduced by oracle networks. By investigating how these *engineered prices* shift from passive measurement tools into active governance variables, the chapter reveals the microstructure consequences of benchmark design. Building on a detailed analysis of failure modes, attack surfaces, and historical liquidation cascades, the chapter ultimately proposes six design principles for building robust price-benchmark governance.

14.1 The price lineage and dual specifications

Price is not a single concept but a continuous lineage extending from micro-level execution records to macro-level governance benchmarks. The evolution from a *transaction price* to a *governance price* is not merely a simple mathematical transformation; it is a process of systematic information refinement, noise filtering, and institutional authorization. Along this lineage, each additional layer of processing increases the price's robustness and institutional authority while decreasing its immediacy and informational granularity. Figure 14-1 presents, from top to bottom, the complete processing chain from the last traded price to the oracle price, annotating the information-filtering mechanism at each layer.



Figure 14-1. Price lineage and the information-refining chain

Any robust price benchmark must be defined simultaneously by its computational specification (which prescribes the mathematical method of calculation) and its institutional specification (which establishes the authority for management and audit)—a dual analytical framework developed in detail at the end of this section.

14.1.1 The transaction-price layer

The last traded price represents the exact price at which the most recent trade actually occurred in a given market. Its core attribute is immediate authenticity, reflecting an undeniable, completed exchange between assets. This price serves as the frontier of price discovery, providing the most immediate signal of market activity [5]. Its event-driven nature, however, means that it updates only when a trade executes and is prone to stalling during periods of insufficient liquidity. Furthermore, the last traded price is highly susceptible to idiosyncratic noise, such as erroneous orders, transient liquidity gaps, or aggressive market orders against a thin order book. For this reason, major cryptocurrency exchanges explicitly prohibit using the last traded price as the basis for triggering liquidation, so that traders are not needlessly liquidated by transient market anomalies [6].

The mid price is computed as the arithmetic mean of the best bid and best ask on the current order book. Unlike the event-driven last traded price, the mid price is state-driven, updating continuously as the order book evolves. This continuous updating provides a smoother representation of market state than the last traded price. In Hyperliquid's robust price index architecture, for example, the mid price is one of the three core inputs that determine the mark price [7]. Despite its continuity, the mid

price remains constrained by its reliance on a single market. During periods of extreme volatility, as market makers withdraw liquidity, the bid-ask spread can widen sharply, causing the mid price to lose representativeness and potentially jump to irrational levels. A simple arithmetic mid price can deviate from true fair value when the resting order sizes at the top of the book are severely asymmetric. Some exchanges and market makers therefore adopt a size-weighted microprice as a more accurate estimate of fair value, but this refinement comes at the cost of greater computational complexity and data dependence. Both the last traded price and the mid price share the features of originating in a single trading venue, high immediacy, and substantial noise. They represent the raw, unprocessed facts of the market and serve as the foundational inputs for all subsequent benchmark construction.

14.1.2 The benchmark-price layer

The index price forms the foundation of the benchmark-price layer, using spatial aggregation to construct a more representative fair value. This process aggregates trading data from multiple spot exchanges that are geographically and operationally independent. This aggregation rarely uses a simple arithmetic mean. When computing the index price, for example, an exchange typically takes a weighted average of the constituent prices and automatically caps any single-exchange price that deviates too far from the median, mitigating the impact of local anomalies [8]. The core advantage of the index price is its manipulation resistance: an attacker attempting to manipulate the index must simultaneously distort the prices of multiple high-liquidity spot markets, a task whose cost and complexity far exceed those of manipulating a single derivatives market [9].

The mark price introduces temporal smoothing on top of the index price's spatial aggregation, making it the core governance price for derivatives risk management. Nearly all major centralized and decentralized exchanges treat the mark price as the sole determinant for computing unrealized profit and loss (unrealized P&L), assessing margin adequacy, and triggering liquidation [10] [8]. The basic logic is consistent across platforms: they anchor to an externally, spatially aggregated price and apply a smoothing mechanism such as an EMA or a median filter to absorb the short-term volatility inherent in derivatives markets [7]. The mark price is not merely a price but a deterministic trigger that prescribes the specific conditions under which a trader's position will be forcibly closed.

14.1.3 The on-chain price layer

The oracle price represents the third layer of the price lineage; it performs the function of transmitting aggregated price information from off-chain centralized exchanges (CEXs) to on-chain smart contracts. This transmission is facilitated by a decentralized network of validators that securely broadcast the price data onto the blockchain. The fundamental characteristic of the oracle price is its discrete, block-level update frequency, which fundamentally alters market microstructure dynamics. The blockchain's discrete public clock creates an environment in which traders no longer compete primarily on latency but instead play a predictive game around the state of the next oracle update [11].

This description applies primarily to DeFi protocols on general-purpose chains such as traditional Ethereum. On high-performance app-chains such as Hyperliquid and dYdX v4, block times have shrunk to the order of 200 to 400 milliseconds, making latency competition once again a central microstructure issue on these platforms.

This discrete update mechanism provides natural immunity to millisecond-scale price manipulation: even if an attacker succeeds in distorting the spot price for a fleeting instant, the on-chain protocol remains unaffected as long as the manipulation does not persist into the next oracle update cycle. This architecture, however, introduces a new type of risk: during extreme market volatility, the oracle price may lag significantly behind the real-time off-chain price, causing the on-chain protocol's risk assessment to diverge from actual market conditions. This lag can cause delayed liquidations or mispriced collateral, exposing DeFi protocols to systemic fragility [12].

14.1.4 The information-refining chain

The evolution across these three price layers is a process of progressive refinement and institutional empowerment. Table 14-1 compares the transaction price, benchmark price, mark price, and on-chain price across five dimensions: generation method, update frequency, noise level, difficulty of manipulation, and governance decision-maker. The table reveals a progressive pattern: as the price layer evolves from the transaction price to the on-chain price, the update frequency falls from the millisecond scale to the block level, the noise level declines, and the difficulty of manipulation rises—but governance authority also shifts from dispersed market participants to concentrated compiling institutions or protocols.

Dimension	Transaction price	Benchmark price (index)	Mark price	On-chain price (oracle)
Generation method	Order-book matching	Multi-source weighted aggregation	Index plus basis smoothing	Oracle-network relay
Update frequency	Tick-by-tick (millisecond)	1–10 seconds	1–5 seconds	Block-level (1–12 seconds)
Noise level	High	Medium	Low	Low but discrete
Difficulty of manipulation	Low (single market)	Medium (requires multi-market coordination)	Medium-high	Depends on oracle architecture
Governance decision-maker	Market makers and traders	Exchange or index committee	Exchange or protocol	Oracle network or DAO

Table 14-1. Comparison of microstructure characteristics across the layers of the price lineage (Data source: compiled by the author)

Beyond these four explicit price layers, the funding rate constitutes an implicit derived price signal. The funding rate is not a direct quote but a function of the basis between the mark price and the index price, reflecting how the perpetual futures market prices directional exposure. When the funding rate is positive, longs pay a holding cost to shorts, implicitly expressing the market's valuation of the current contract premium; when the funding rate is negative, the direction reverses. Although the funding rate

does not directly participate in liquidation decisions as the four price layers above do, it regulates the long-short balance through economic incentives and thereby indirectly affects the speed of convergence between the mark price and the index price. It therefore occupies a derived tier in the price lineage, sitting between the mark price and the on-chain price; its governance implications are developed in detail in Section 14.6.2.

The chain begins with the last traded price and the mid price, which represent the market's raw, event-driven facts, characterized by high noise but maximal immediacy. The spatial aggregation of the index price filters out the idiosyncratic noise of individual trading venues, producing a consensus price. The temporal smoothing of the mark price adds a buffer against transient shocks, yielding a robust governance price suitable for risk management. Finally, the oracle price carries these refined benchmarks across the network boundary, delivering them to on-chain protocols as discrete pulses. Each step of processing entails a fundamental trade-off: as robustness and manipulation resistance rise, the sensitivity and immediacy of the information decline. This trade-off is not a technical defect but the central proposition of benchmark governance design.

14.1.5 Computational and institutional specifications

To evaluate any price benchmark comprehensively, a dual analytical framework comprising a computational specification and an institutional specification must be applied. The computational specification addresses the question of how the price is calculated, covering the choice of data sources, the allocation of weights, the handling of outliers, the update frequency, and the parameters of the smoothing algorithm [8] [7]. These technical specifications determine the benchmark's statistical properties, including its noise level, degree of lag, and manipulation resistance.

The institutional specification, in turn, addresses the governance and accountability structures surrounding the benchmark. It defines the benchmark's intended use cases—such as computing margin, triggering liquidation, or determining the funding rate—because different applications demand markedly different levels of precision and error tolerance. It identifies the managing entity, distinguishing a centralized risk-management department from a decentralized oracle network, which directly affects the source of credibility and the potential conflicts of interest. The institutional specification also stipulates the procedures for modifying the benchmark, such as adjusting index constituents or weights, as well as the contingency protocols for handling data-source failures or extreme market events. Understanding the institutional specification is essential to recognizing the true function and systemic risk of a price benchmark, because a small deviation in a benchmark used to trigger billions of dollars in liquidations has fundamentally different consequences from a deviation in a purely informational price feed.

14.2 Price benchmarks as implicit governance

In crypto-asset markets, price is often regarded as a neutral mathematical outcome, an objective reflection of supply and demand interacting on the order book. However, once a particular price is extracted, processed, and established as the benchmark for

derivatives liquidation, collateral valuation, or protocol triggers, it sheds its purely descriptive function and becomes a governance instrument with coercive force. In the terms of institutional economics, this transformation maps onto Williamson's (1985) [13] transaction-cost governance framework: the construction of a price benchmark involves highly asset-specific data (the order-book information unique to each exchange), and this specificity creates a hold-up problem—once a derivatives market depends on a particular index-construction rule, the switching cost of changing the benchmark locks market participants into the existing governance structure. At the same time, North's (1990) [14] institutional-analysis framework reveals the nature of a price benchmark as a formal institutional constraint: the benchmark's construction rules, outlier-handling thresholds, and degradation strategies constitute a set of institutional arrangements that govern economic exchange, and their evolutionary trajectory—from the informal *expert judgment* of the LIBOR era to the rule-based benchmarks promoted by the International Organization of Securities Commissions (IOSCO)—follows the general pattern of a transition from informal to formal institutions. Under the framework of incomplete-contract theory, no finite set of algorithmic rules can anticipate every possible market state; consequently, the allocation of residual control rights—the *modification power* discussed in this section—becomes an unavoidable core problem in benchmark governance design. This section reconceptualizes the price benchmark as an implicit governance mechanism, arguing that *defining how price is computed* is itself a power structure that allocates market benefits and risks. Examining the functional role of these benchmarks, their inherent power dimensions, and the LIBOR manipulation case in traditional finance as a historical reference, it traces the systemic fragility that the multiple conflicts of interest in crypto markets produce, and argues that a multi-price system, understood as an institutionalized stratification, is the inevitable result.

14.2.1 Three engineered prices

In crypto derivatives markets and DeFi protocols, what traders directly confront is often not a single spot-market execution price but highly *engineered prices*—namely the price index, mark price, and oracle price defined in Section 14.1. Rather than restating the computational construction of the three, this section focuses on revealing their shared governance attribute: each form is optimized for a specific market friction, and its design choices directly embody the compiler's governance preference between price discovery and risk control.

In the case of the price index, its compiler exercises the power to define *what counts as the real market*. Exchanges included in the index receive an endorsement of their pricing authority, while liquidity excluded from it is treated as noise or an unreliable signal.

The mark price introduces buffering mechanisms such as a moving average or the funding-rate basis; this design is essentially a protective governance intervention. By delaying its confirmation of the market's latest price, the exchange or protocol sacrifices the timeliness of price discovery in exchange for the overall liquidation stability of the system.

The oracle price, in turn, must cross a trust boundary to anchor its result in a smart

contract. An oracle network must not only solve the problem of data-source accuracy but also defend against economic attacks on its price-feed nodes; its governance therefore resides not only in the data-aggregation algorithm but, more deeply, in the staking-and-slashing mechanism for validators and the economic game of decentralized consensus. These three price forms are not neutral measurement tools; they are concrete expressions of specific governance objectives, determining who faces liquidation and who gains arbitrage room under extreme market conditions.

14.2.2 The four dimensions of power in price computation

To view a price benchmark as implicit governance is to deconstruct the exercise of power hidden behind the mathematical formulas. This power is not expressed explicitly through voting or proposals but is embedded implicitly in code and algorithms, taking concrete form in four dimensions of control: the power of selection, the power of weighting, the power of exclusion, and the power of modification. The power of selection determines which data sources are eligible to participate in generating the price. In the highly fragmented crypto market, spot exchanges differ enormously in liquidity depth, compliance standards, and infrastructure stability. By setting admission thresholds, the benchmark compiler is in effect rating the market. When an index announces, for instance, that it will include a given exchange's spot price in its calculation, it not only enhances that exchange's reputation but also turns its order book into an underlying asset for derivatives pricing. Conversely, stripping a data source of its eligibility means severing its influence over the derivatives market.

The power of weighting determines the distribution of pricing influence among the data sources. An equal-weighting model implies equal trust in all data sources, whereas a model weighted by trading volume or liquidity depth concentrates pricing power in the leading platforms. This allocation of weights directly affects the cost of manipulating the benchmark. If a poorly liquid exchange is assigned a disproportionately high weight, an attacker can influence the vast derivatives market with only a modest amount of capital by ramping the spot price on that platform.

The power of exclusion is the benchmark system's emergency circuit-breaker against extreme anomalies. When a data source's price deviates from the median beyond a set threshold, the algorithm triggers exclusion or truncation logic. The choice of this trigger threshold reflects the governor's demarcation between *normal volatility* and *abnormal noise*. An overly sensitive exclusion mechanism may discard a critical data source during genuine, violent market swings, while an overly sluggish setting may allow erroneous data to contaminate the entire benchmark.

The power of modification is the discretionary space that the benchmark compiler reserves outside the rule framework. When facing an unforeseeable black-swan event or a structural change such as a hard fork of the underlying asset, the compiler has the right to intervene in the calculation method on an emergency basis. Although the existence of this power provides the system with necessary flexibility, it also creates potential room for insider rent-seeking and self-dealing. These four dimensions of power intertwine to constitute the implicit governance structure of a price benchmark, turning seemingly objective numbers into the outcome of a multi-party game of interests.

Figure 14-2 compares the price deviations of different benchmarks during the Bitstamp flash crash of May 2019. The spot price on the single platform Bitstamp fell by approximately 20% within minutes, whereas the Brave New Coin index, which used data from multiple exchanges, showed a decline of only approximately 13% to 16%. The gap between the two illustrates how decisively the number of index constituents shapes manipulation resistance.

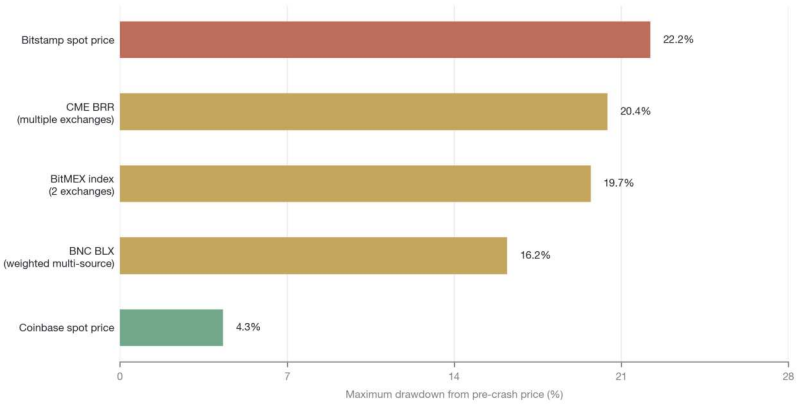


Figure 14-2. Comparison of price deviations across benchmarks during the Bitstamp flash crash

Note: The pre-crash price of $7,944 has been verified against primary records; the deviation magnitudes for each benchmark in the figure are cited from or derived from Brave New Coin's analysis and serve only as an illustrative comparison.

14.2.3 The historical lessons of LIBOR

To fully understand the fragility of a price benchmark as a governance instrument, one must revisit the most severe benchmark-manipulation event in the history of traditional finance: the LIBOR scandal. This historical reference provides a necessary frame for analyzing the structural risks of current crypto-market price benchmarks. LIBOR was once the most widely used benchmark interest rate in the global financial system, underpinning financial derivatives and loan contracts with a notional value as high as $350 trillion [15]. Yet this number, regarded as a core benchmark of financial markets, was ultimately proven to have been systematically manipulated over a long period by a small number of submitting banks.

The collapse of LIBOR did not stem from mere moral failure but from a systemic failure of its inherent governance structure. First, the benchmark's generation relied heavily on the *expert judgment* of a handful of large banks rather than on real underlying market transaction data. This quote-based mechanism, grounded in subjective valuation, offered significant scope for manipulation in the absence of transparency and external audit. Second, the submitting banks held vast positions in derivatives markets, and their trading desks' profits were directly tied to minute movements in the benchmark rate. This institutionalized conflict of interest gave the banks a strong incentive to manipulate the final benchmark rate by submitting false quotes in order to profit their

own trading positions.

When IOSCO and national regulators intervened to investigate and drive reform, they found that the core problem lay in the loss of control over power in the benchmark's construction [16]. The central demand of the reform was to shift from quotes based on subjective judgment to anchoring on real transaction data, and to establish strict benchmark governance and conflict-of-interest firewalls [2]. The lesson of LIBOR is that any price benchmark that lacks transparency, harbors serious conflicts of interest, and affects enormous financial interests will inevitably become a target for attack and manipulation. The fundamental reason LIBOR manipulation persisted for so long was that the benchmark rested on banks' subjective quotes rather than on actual transaction data. Crypto-market index prices are typically anchored to real on-chain or off-chain execution records, and this difference in informational foundation makes the same kind of systematic false quoting harder to achieve in the crypto context. Nevertheless, the core lesson of the LIBOR scandal—that the interweaving of concentrated governance power and conflicts of interest inevitably distorts a benchmark—reappears in crypto markets in an even more concentrated form through the *three-in-one* structure. From the perspective of comparative institutional analysis, the manipulation risk facing crypto-market price benchmarks more closely resembles the manipulation patterns of foreign-exchange fixings or commodity-benchmark window trading—that is, distorting the benchmark result through real trades within a specific time window. This lesson carries important significance for crypto derivatives markets, which depend so heavily on *engineered prices*.

14.2.4 Conflicts of interest in crypto markets

If the LIBOR scandal revealed the destructive power of benchmark manipulation in traditional finance, then the structure of the current crypto derivatives market embodies an even higher degree of systemic risk. In the traditional financial system, the benchmark compiler, the exchange, and the clearinghouse are typically separate entities, whereas in crypto markets a centralized exchange often combines all three roles. This *three-in-one* structure produces highly concentrated conflicts of interest, exposing the implicit governance of price benchmarks to significant moral hazard.

First, a crypto exchange is simultaneously the compiler and the user of the price index. It independently decides the index's constituents, weights, and outlier-handling rules while applying those indices to the perpetual futures and delivery futures on its own platform. More critically, the exchange is also the operator of the liquidation engine and the manager of the risk reserve. When the market swings violently, the exchange holds the power to decide, by adjusting the mark price algorithm, whether to trigger large-scale liquidations. This power affects not only users' profits and losses but also, more directly, the exchange's own liquidation revenue and systemic risk exposure.

The quality of the underlying spot data in crypto markets is itself deeply contentious. Numerous studies show that many unregulated crypto exchanges engage in serious wash trading and volume inflation, severely distorting publicly reported trading volumes [17]. If the compilation of a price index relies excessively on these contaminated data sources, or fails to effectively strip out fake liquidity when allocating weights, then the

derivatives market built upon it stands on an unreliable data foundation. The flash crash on a leading exchange in 2019 occurred precisely because a derivatives platform's price index relied excessively on a single spot exchange, so that a local selling pressure in the spot market was amplified into a chain of liquidations in the derivatives market, causing hundreds of millions of dollars in forced closures [9]. This structural fragility shows that crypto-market price benchmarks face not only the risk of manipulation by external attackers but also the immense pressure of internal governance failure.

This *three-in-one* structural conflict of interest has drawn the attention of the world's major financial regulators. Article 72 of the European Union's Markets in Crypto-Assets (MiCA) regulation sets out explicit requirements for crypto-asset service providers to identify, prevent, manage, and disclose conflicts of interest [18]; the systematic principles specifically addressing the isolation of conflicts of interest in benchmark administration were established in IOSCO's 2013 Principles for Financial Benchmarks [16], and IOSCO's 2023 Policy Recommendations for Crypto and Digital Asset Markets likewise set out general norms for the conflict-of-interest governance of crypto service providers [19]. This means that the benchmark governance power many exchanges currently wield on their own may, in the future, shift from a voluntary design choice to a mandatory legal obligation; Chapter 26 discusses this regulatory evolution in detail.

14.2.5 The inevitability of a multi-price system

Facing the inherent governance dilemmas and structural risks of price benchmarks, crypto markets did not pursue a single, perfect, universal price but instead evolved a multi-price system. This system is not design redundancy but an institutionalized stratification arising from a mathematical incompatibility between the fundamental requirements that different financial functions place on price attributes. In market microstructure theory, price discovery and risk control are two objectives pointing in sharply different directions [5]. Price discovery requires a benchmark to be extremely sensitive to marginal changes in market information, able to reflect the latest supply-demand imbalance rapidly. Yet this sensitivity inevitably introduces microstructure noise and short-term volatility. Conversely, risk control—such as derivatives liquidation and collateral valuation—requires a benchmark to be highly robust, able to filter out transient anomalous swings and prevent the system from falling into a self-fulfilling spiral of chained liquidations.

This inconsistency of objectives dictates that the spot last traded price, the price index, and the mark price must exist in stratified layers. The spot price bears the function of price discovery, allowing arbitrageurs to eliminate spreads across different markets; the price index, through cross-market weighting, provides a relatively smooth fair-value reference; and the mark price, by introducing a time-weighted average or basis limits, further sacrifices timeliness to serve as the terminal safeguard of risk control. This multi-price system is essentially a governance compromise the market reaches between sensitivity and robustness. It acknowledges that no single calculation method can satisfy the needs of all financial scenarios at once, and therefore, through institutionalized stratification, it allocates risks and benefits along different dimensions to the market

participants best suited to bear them. The evolution of this implicit governance mechanism marks the emergence of a distinctive microstructure logic in crypto markets as they confront the challenges of complex financial engineering.

14.3 The five governance links of index construction

The construction of a price index is not a purely technical computation but a systematic engineering of a series of governance decisions. From raw market trade data to the final fair price that serves as the derivatives pricing benchmark, the data must pass through multiple filtering, weighting, and smoothing mechanisms. This section deconstructs the process into *five governance links*: source admission, the weighting scheme, outlier handling, update smoothing, and stress degradation. These five links are tightly coupled, together forming the defensive system and power structure of the price benchmark. The design choices at each link shape the benchmark's manipulation resistance, its market sensitivity, and how gains and losses are distributed under extreme conditions. By dissecting these five links one by one, we reveal the microstructure consequences behind index construction and explore how an index evolves, in reverse, from *reflecting* prices into a financial infrastructure that *steers* prices.

14.3.1 Overview of the five governance links

In crypto derivatives markets, the generation of a price index is a highly engineered process whose core objective is to extract, from spot markets rife with noise, manipulation risk, and fragmented liquidity, a fair value that both derivatives traders and the liquidation engine can trust. This extraction process can be conceptualized as a governance framework composed of five consecutive links—the *five governance links*.

The five governance links unfold in sequence along the direction of data processing. The starting point is source admission, which determines which spot exchanges' order-book and trade data are eligible to enter the index's computation pool; through screening for compliance, liquidity, and technical stability, it excludes low-quality or high-risk markets. Once the eligible sources are determined, the weighting scheme allocates pricing influence among the admitted exchanges; whether it uses volume weighting, equal weighting, or a hybrid tiered model, the weighting scheme directly determines the index's sensitivity to the fluctuations of different markets. Next comes the outlier-handling link, whose function is to monitor and filter extreme deviations in the data sources in real time; by setting a median-deviation threshold or trimming extreme values, it prevents a single exchange's flash crash or malicious manipulation from contaminating the overall index. The outlier-filtered data then enter the update-smoothing stage, which applies primarily to the generation of the mark price; by introducing a time-weighting or moving-average mechanism, it processes the index a second time to buffer sharp short-term fluctuations and prevent the derivatives market from falling into a chain of liquidations triggered by microstructure noise. The terminal link is stress degradation—the system's emergency mechanism when facing an extreme disaster, such as the outage of multiple major exchanges or a widespread failure of data sources. It defines how the system transitions smoothly from multi-source aggregation to single-source reliance, or even, in the most extreme case, activates circuit-breaker

protection. These five links are tightly interconnected, and the parameter settings of each represent a governance trade-off the compiler makes between price discovery and system stability.

Figure 14-3 summarizes the five governance links of index construction and the microstructure transmission mechanism.

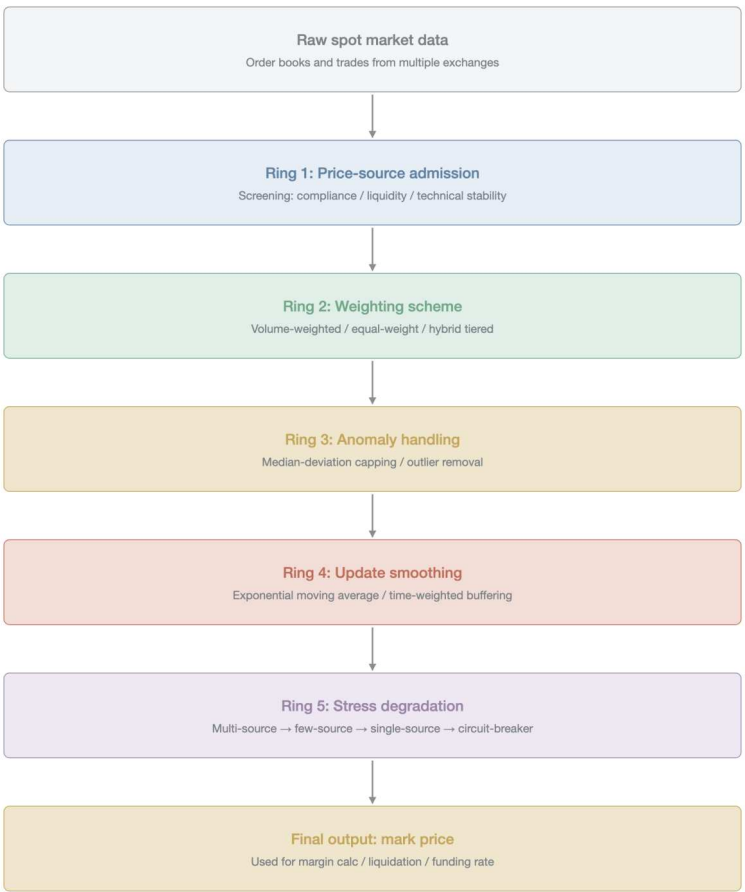


Figure 14-3. The five governance links of index construction and the microstructure transmission mechanism

The figure clearly displays the serial dependency among the five links: the decisions of source admission directly constrain the candidate set for the weighting scheme, the threshold parameters of outlier handling determine the quality of the data entering the update-smoothing stage, and the trigger conditions of stress degradation depend on the degree of redundancy in the first four links. Any parameter adjustment in a single link propagates downstream along this chain structure, ultimately affecting the timing and precision of liquidation triggering.

14.3.2 Source admission

Source admission is the first line of defense in index construction and the link with the most pronounced governance implications. In traditional financial markets, the underlying data for a benchmark typically come from strictly regulated exchanges or large market makers, so their data quality carries a natural compliance endorsement. Crypto markets, however, exhibit high fragmentation and regulatory-arbitrage characteristics, and different exchanges vary significantly in liquidity depth, trading rules, and anti-manipulation mechanisms. Deciding which exchanges to include as index constituents is therefore essentially the process of granting these platforms pricing authority.

A compiler typically establishes a complex set of admission criteria, including but not limited to minimum volume requirements, order-book bid-ask spread limits, the stability of the application programming interface (API), and compliance status. Some leading index compilers, for instance, explicitly exclude from their constituent pool any exchange suspected of serious wash trading or lacking basic anti-money-laundering measures [20]. This admission mechanism not only protects the accuracy of the index but also exerts a significant normative effect on the entire crypto ecosystem. A spot exchange included in a mainstream derivatives index tends to attract more market makers and arbitrage capital, further boosting its liquidity, whereas an excluded exchange may face the marginalizing risk of liquidity depletion.

Source admission, however, also faces severe governance challenges. The first is the *too big to fail* dilemma. If a leading exchange with a dominant market share suffers a compliance crisis or technical failure, the index compiler faces a dilemma: excluding it may cause the index to lose its most representative price signal, while retaining it may inject enormous systemic risk into the derivatives market. Second, excessive concentration of sources can create a single point of failure. When a derivatives platform's index relies excessively on a single or a very small number of spot exchanges, an attacker need only concentrate capital on manipulating the weak link of that spot market to reap outsized profits in the derivatives market. A decentralized oracle manipulation incident in 2022 arose precisely because its price-feed mechanism relied excessively on a single, highly illiquid decentralized exchange (DEX), allowing the attacker to manipulate the underlying asset's price at extremely low cost and thereby drain the protocol's reserves [6]. As global regulatory frameworks gradually take shape, jurisdictional compliance status is becoming an emerging screening dimension for source admission. Under regimes such as Japan's Financial Services Agency (FSA) registration regime, the Monetary Authority of Singapore (MAS) licensing regime, and Hong Kong's Securities and Futures Commission (SFC) virtual-asset-trading-platform licensing regime, constructing an index for regulated participants using price data from unlicensed exchanges may itself face compliance challenges. In late 2022, Avraham Eisenberg was arrested by the U.S. Department of Justice for manipulating the oracle price on Mango Markets, and in April 2024 a jury found him guilty of charges including commodities fraud; that conviction, however, was vacated in May 2025 by the U.S. District Court for the Southern District of New York under Rule 29 of the Federal Rules of Criminal Procedure (on the grounds of improper venue and insufficient evidence on the elements of fraud, with the prosecution having filed an appeal), so its legal characterization remains unsettled [21]. This reversal itself reveals the high degree of uncertainty in the criminal prosecution of

on-chain manipulation, both as to venue and as to establishing the elements of fraud, and it means that the expected legal-cost term in the cost-benefit equation of manipulation analyzed in Section 14.5.8 remains subject to significant uncertainty. Source admission is therefore not merely a technical screen but a profound act of governance over the distribution of market liquidity and the concentration of risk.

14.3.3 Weighting schemes

Once the eligible sources are determined, the next core question is how to allocate pricing weight among these data sources. There is no absolutely correct mathematical answer to the design of a weighting scheme; each method is an expression of specific assumptions about market microstructure characteristics and of governance preferences. The most intuitive approach is the volume-weighted model, which allocates weights according to each exchange's spot trading volume over a recent period. The underlying logic of this model is that liquidity equals representativeness: the larger a market's trading volume, the more effective its price-discovery function, so it should dominate the index. In crypto markets, however, volume data are often prone to contamination by wash trading. If the index compiler fails to identify and strip out fake volume effectively, the volume-weighted model instead transfers pricing power to inferior exchanges that manufacture fake liquidity through wash trading.

To overcome the shortcomings of volume weighting, some indices adopt an equal-weighting model. This model assumes that all exchanges passing the admission test have equal pricing efficacy, thereby avoiding an excessive concentration of pricing power in a single giant. The equal-weighting model has an advantage in guarding against a single point of failure at a leading exchange, but it also has an obvious weakness: when the market swings violently, poorly liquid small exchanges tend to exhibit larger price deviations, and under an equal-weighting model these anomalous swings exert a disproportionate disturbance on the overall index.

A growing number of leading derivatives platforms have adopted a hybrid tiered model [22]. This model jointly considers multiple dimensions such as trading volume, order-book depth, and price deviation. When computing weights, for example, it looks not only at historical volume but also samples the resting depth on the bid and ask in real time; at the same time, if one exchange's latest price deviates from the median of all constituent exchanges, its weight may be dynamically reduced or even set to zero. By introducing a dynamic feedback mechanism, this hybrid model allows the index to reflect the consensus of mainstream liquidity under normal market conditions while rapidly isolating the risk of anomalous nodes under extreme conditions. This sophisticated weight design is essentially a governance practice in which the compiler, through algorithmic rules, finely regulates the strategic behavior of market participants.

Figure 14-4 contrasts the differences among major derivatives platforms in the number of price sources and the methods of outlier handling.

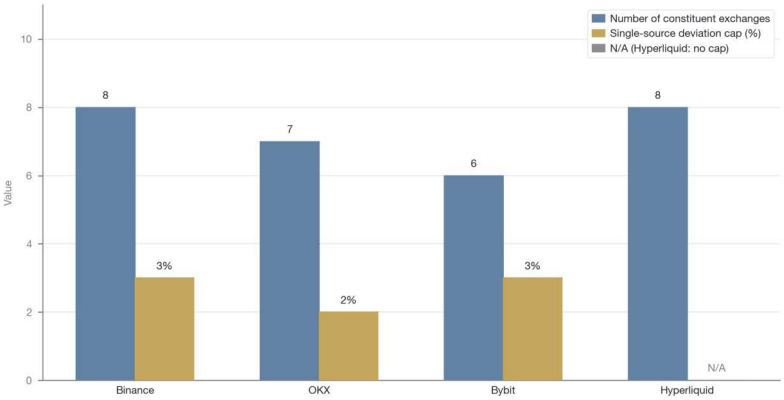


Figure 14-4. Comparison of index-construction methods across major derivatives platforms

Note: The number of constituent exchanges and the single-source deviation clamp threshold in the figure were compiled by the author from each platform's official documentation (as of March 2026); construction rules change over time, and the definitive configuration is whatever each platform's latest documentation states. In addition, the *number of constituent exchanges* in this figure and the *number of price sources* in Table 14-2 are metrics on different bases (the former is the count of constituent exchanges included in index construction, the latter is the number of price sources referenced beneath the mark price), so any discrepancy between the two values is not a contradiction.

The figure shows that platforms differ by several times in the number of price sources, which directly determines the level of robustness each index has when facing a single-exchange anomaly. Platforms that use more price sources typically pair them with more complex outlier-exclusion logic, whereas platforms with fewer sources tend to compensate by relying on stronger median-truncation mechanisms. This differentiated construction strategy reflects each platform's governance trade-off between operational complexity and manipulation resistance.

14.3.4 Outlier handling

In the process of aggregating multi-source data, outlier handling is an indispensable governance link. Because exchanges differ in liquidity depth, network latency, and market-maker strategies, it is common for one constituent exchange's price to deviate momentarily at a particular instant. The core task of the outlier-handling mechanism is to strike a delicate balance between *filtering microstructure noise* and *preserving genuine market signals*.

The most common outlier-handling method is a median-based deviation limit. The system computes the median of all constituent exchanges' prices in real time and sets an allowable deviation threshold (for example, 2% or 3%). If one constituent exchange's

price exceeds this threshold, the system does not adopt that extreme price directly but instead clamps it to the threshold boundary [8]. If the median is \$10,000 and the deviation cap is 3%, for instance, then even if one exchange's price surges to \$11,000, the system will treat it as only \$10,300 when computing the index. This clamping mechanism effectively prevents a single exchange's *wick* from damaging the overall index.

The design of the outlier-handling mechanism, however, is fraught with governance challenges. If the deviation threshold is set too narrow, the system may misjudge a genuine, drastic market change as anomalous noise. When the market faces a systemic shock (such as a major macro event or the release of a regulatory policy), different exchanges will inevitably react at different speeds because of their liquidity differences. Overly strict outlier handling may cause the index to lag behind the real market price at a critical moment, thereby inviting arbitrage attacks on the derivatives market. Conversely, if the threshold is set too wide, it cannot effectively guard against extreme deviations caused by malicious manipulation or technical failure. Outlier handling is therefore not merely a statistical trimming operation but an institutionalized expression of the governor's tolerance for market volatility.

14.3.5 Update smoothing

After price aggregation and outlier filtering, the index often needs to undergo update smoothing to further adapt to the risk-control needs of derivatives markets, thereby generating the final mark price. The core tools of smoothing include algorithms such as the EMA, the median filter, and the time-weighted average price (TWAP), each with different characteristics in noise filtering and the amount of latency it introduces. The fundamental reason for introducing a smoothing mechanism is that the liquidation logic of derivatives markets has an inherent *self-exciting* feature: when the market price briefly touches a certain level, it may trigger the forced closure of a large number of leveraged positions, and these closing orders flood the market as market orders, pushing the price further in the adverse direction and triggering still more liquidations, forming a positive-feedback loop. The smoothing mechanism breaks this loop by artificially slowing the update speed of the mark price: a transient price spike, buffered by the moving average, causes only a mild displacement of the mark price [7], giving leveraged traders an adjustment window in which to post additional margin or wait for liquidity to recover. Yet the smoothing mechanism trades timeliness for stability: during the smoothing period, a basis necessarily exists between the mark price and the spot price, providing room for arbitrageurs, and it also means that in a one-sided move the derivatives market's price signal always lags the spot market. The choice of smoothing parameters therefore directly reflects the governor's value orientation between protecting leveraged traders and maintaining pricing efficiency. The quantitative analysis of EMA latency and its liquidation consequences is developed in detail in Section 14.4.3.

14.3.6 Stress degradation

The final governance link in index construction is the stress-degradation mechanism. Ideally, multi-source aggregation can provide a stable and representative price bench-

mark. Yet crypto-market infrastructure remains relatively fragile, and extreme events such as exchange outages, network blockages, or API failures occur from time to time. When multiple constituent exchanges lose connectivity simultaneously, or when a momentary evaporation of market liquidity causes quotes to fail on a large scale, the index system must have a contingency plan to ensure that the derivatives market is not paralyzed by the loss of its pricing benchmark.

The stress-degradation mechanism is typically designed as a progressive degradation strategy. When some data sources fail, the system first attempts to reallocate weights, transferring the weight of the failed nodes to the remaining normal nodes. If too many nodes fail and the remaining constituent exchanges are insufficiently representative, the system may trigger a second level of degradation—for example, interpolating with historical prices or estimating by reference to the trajectory of other comparable indices.

The degradation strategy above harbors a key assumption: that source failure is detectable—that is, the system can clearly distinguish between the states of *a source being unavailable* and *a source providing data*. In the security literature on distributed systems, however, the more dangerous case is Byzantine failure, in which a data source is still operating normally and continuously outputting prices, but its output has been maliciously or erroneously shifted. A sophisticated attacker would not take an exchange offline (which would trigger an obvious degradation process); instead, the attacker might maintain a small but systematic price shift in a heavily weighted constituent exchange, keeping the anomalous price within the median-deviation threshold so that the outlier-handling mechanism does not identify it. When the deviation threshold described in Section 14.3.4 is set at 2% to 3%, a persistent shift below that threshold will be treated as normal volatility and fully incorporated into the index calculation. Defending against this attack requires going beyond simple threshold detection to introduce statistical anomaly tests across time windows and consistency checks across data sources.

In the most extreme case, when all external data sources are unavailable, the system must activate its last line of defense. Some exchanges adopt a *last-price protection* mechanism: in the absence of an external reference, they temporarily use the derivatives contract's own last traded price as the interim benchmark for computing unrealized P&L and triggering liquidation [11]. Although this single-source degradation is extremely dangerous and prone to internal manipulation, it is the only choice for keeping the market running when the system faces total failure. The design of the stress-degradation mechanism depends on the governor's crisis-management capability under extreme uncertainty. It determines whether, when disaster strikes, the system chooses a rigid circuit-breaker halt to protect the accuracy of the price, or a compromising degradation to maintain the continuity of trading.

14.3.7 Microstructure consequences

Taking the five governance links of index construction together, one microstructure phenomenon deserves attention: in crypto markets, a price index is no longer merely a measurement tool that objectively *reflects* prices but has evolved into a financial infrastructure that can, in reverse, *steer* prices. This reflexivity effect is transmitted

mainly through market-maker behavior and arbitrage mechanisms.

First, the design of the index's constituents and weights directly shapes market makers' hedging strategies. Because the pricing and liquidation of derivatives depend heavily on the index, market makers must hedge their exposure on the spot market when providing derivatives liquidity. To minimize basis risk, market makers tend to conduct their hedging trades on the heavily weighted constituent exchanges. This behavioral logic causes liquidity to concentrate further in the leading constituent exchanges, continually reinforcing their pricing power in the spot market. The allocation of index weights thus becomes, in effect, the core variable that steers the market's decisions about where liquidity flows.

Second, the index's outlier-handling and smoothing mechanisms create specific arbitrage opportunities. When the spot market swings violently while the mark price has not yet fully caught up because of the moving-average buffer, arbitrageurs are quick to exploit this deterministic lag. By taking offsetting positions between the spot and derivatives markets, they seek to profit from the process by which the mark price ultimately reverts to the spot price. Although this arbitrage activity objectively promotes the convergence of prices across markets, at the microstructure level the influx of arbitrage capital also exerts a significant impact on the spot market's order book and, under specific conditions, can even amplify the spot market's volatility.

Empirical research supports the reflexivity effect described above. Makarov and Schoar's (2020) [23] study of cross-exchange arbitrage in cryptocurrency markets shows that the speed at which price spreads between exchanges are eliminated is positively correlated with an exchange's weight in mainstream indices: the higher an exchange's weight, the faster arbitrage capital flows in and the more rapidly spreads converge, which corroborates precisely the mechanism by which index weights, in reverse, shape the distribution of spot-market liquidity by steering the flow of hedging and arbitrage capital. The causal interpretation of this correlation, however, warrants caution. A high-liquidity exchange is both more likely to be included in an index and given a higher weight and naturally has a faster spread-elimination speed, and liquidity, as a common driver, may simultaneously determine both the weight and the arbitrage speed. To identify rigorously the causal effect of index weight on arbitrage activity would require a natural-experiment design that uses index-constituent adjustment events as exogenous shocks. More fundamentally, when a price index is widely adopted and becomes the settlement benchmark for tens of billions of dollars in derivatives contracts, it becomes a systemically important financial infrastructure in its own right [12]. Any small rule adjustment to that index—whether adding or removing a constituent exchange or modifying the 1% deviation threshold—triggers an enormous redistribution of wealth in the derivatives market. Index construction is therefore no longer a purely quantitative engineering task but a governance practice involving a multi-party game of interests. In this process, the power to define how price is computed has become the most consequential and least transparent governance power in the crypto financial system.

14.4 Mark price and liquidation triggering

The mark price is not a market execution price in the traditional sense but the core trigger within a derivatives exchange's risk-control system. It determines the conditions under which a trader's position will be forcibly liquidated, and its design faces a trilemma among accuracy, robustness, and timeliness. The EMA smoothing mechanism offers protection against short-term manipulation but can create a retroactive-liquidation trap in violent reversals. Each method of computing the mark price is essentially an expression of the exchange's governance preference across different market-risk scenarios, directly shaping the microstructure consequences of liquidation events.

14.4.1 The institutional status of the mark price

In crypto derivatives markets, the mark price is granted the highest institutional status, reflected in its designation as the sole benchmark for three core risk-control functions. The mark price is the absolute benchmark for computing unrealized P&L: the floating profit and loss displayed in a trader's account is not based on the last traded price but is determined by the mark price. This design defines a trader's book wealth at any given moment, thereby protecting the trader from the psychological interference of ultra-short-term order-book noise and preventing the sharp book swings caused by a single anomalous trade from triggering irrational behavior.

The mark price is also the basis for computing the margin ratio. Major exchanges all use the mark price to compute position value and the margin ratio, with the core aim of preventing users from being liquidated within a short window because the last traded price has been maliciously manipulated. When market liquidity dries up momentarily or suffers a large directional shock, the last traded price on the order book may deviate sharply, whereas the mark price—built from global data—can remain relatively stable, ensuring the fairness of margin computation. In a high-leverage environment, even a 0.5% price deviation is amplified into a 25% change in the margin ratio at 50× leverage, so the stability of the mark price becomes a key safeguard against systemic misjudgment.

Above all, the mark price is the sole arbiter that triggers liquidation. Whether a position is liquidated is judged by one criterion alone: whether the mark price reaches the liquidation price. However violently the last traded price swings, the liquidation procedure will not start as long as the mark price stays above the safety line. This fair price marking principle, pioneered by BitMEX [10], decouples the liquidation trigger from the instantaneous execution price of any single exchange, effectively isolating derivatives positions from the destructive impact of local liquidity crises; it has now become the industry-standard risk-control paradigm. The fair price marking paradigm pioneered by BitMEX has been extended further within the unified-margin systems of contemporary mainstream exchanges: the mark price not only determines liquidation triggering and unrealized P&L but also affects the cross-asset offsetting calculation of portfolio margin, the ordering of auto-deleveraging, and the adjustment of dynamic risk limits. BitMEX's documentation explicitly delimits the scope of fair price marking: it affects only the computation of the liquidation price and unrealized P&L, whereas realized profit and loss (realized P&L) is still based on the actual execution price [10].

The institutional authority of the mark price within the liquidation mechanism is the institutional empowerment at the heart of its role as a governance price. It is not merely a technical parameter but a key node of power allocation within the entire risk-control system, ultimately determining whose positions will be liquidated and whose assets will be protected.

14.4.2 The mark price trilemma

The design of the mark price faces a fundamental trilemma: accuracy, robustness, and timeliness cannot all be optimized at once. Let P_t^* be the true fair price at time t (measured as the volume-weighted median across multiple exchanges), M_t the mark price function, ΔM_{attack} the mark-price shift induced by a single-exchange price manipulation ϵ, and $\text{Lag}(M)$ the latency (in seconds) with which the mark price reflects new information. Accuracy requires minimizing $\mathbb{E}[|M_t - P_t^*|]$, robustness requires minimizing $\sup_\epsilon |\Delta M_{\text{attack}}(\epsilon)|$, and timeliness requires minimizing $\text{Lag}(M)$. In an information-theoretic sense, these three objectives form a trade-off on the Pareto frontier: any mark price function can optimize only two of them at once, and the third is inevitably sacrificed.

Figure 14-5 shows the different design orientations of major trading platforms within the mark price trilemma.

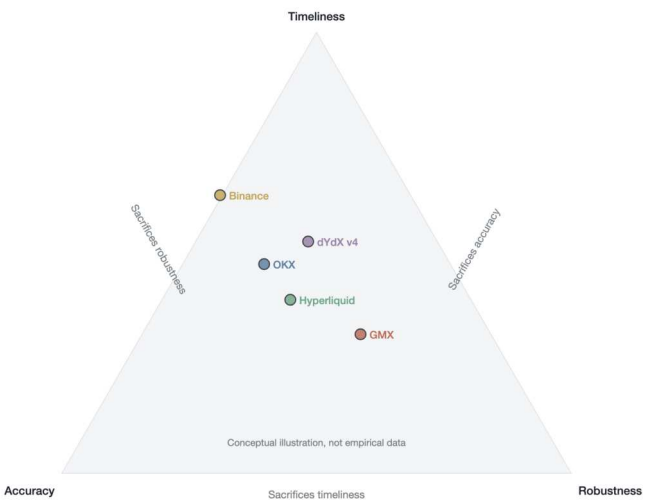


Figure 14-5. The mark price trilemma and each platform's design choice

There are deep, inherent conflicts among these three objectives. If a system pursues the extremes of accuracy and timeliness, it can only use the last traded price directly, but this leaves it almost defenseless against price manipulation, severely sacrificing robustness. If a system chooses the combination of accuracy and robustness, it typically must introduce multi-source weighting and a long-window EMA algorithm, which

inevitably causes the mark price to lag badly in a rapid one-sided move, thereby sacrificing timeliness. If a system tries to balance robustness and timeliness, it often adopts a short-window median algorithm, but when liquidity is unevenly distributed across platforms, the median may systematically deviate from the true fair value, thereby impairing accuracy.

Each major platform's mark price formula is essentially its specific choice within the trilemma, reflecting a different governance preference. Binance takes the median of three values—the price index, an adjusted price containing a one-minute basis moving average, and the last traded price [24]—leaning toward a balance of accuracy and timeliness. In May 2025, Binance shortened the basis moving-average window from 2.5 minutes to 1 minute, and in September 2025 shortened it further to 30 seconds [24]; this series of adjustments clearly embodies a governance intent tilted toward timeliness. This median-of-three design is naturally resistant to a single-point anomaly, but under extreme conditions, when all three inputs are shocked simultaneously, it still risks being skewed by a local price. OKX's design instead anchors directly to the index price and adds a moving average of the basis [25], leaning toward accuracy and robustness; its smoothing term reflects only the contract premium and is structurally clear, but the moving average introduces substantial latency in a rapid move. Hyperliquid adopts a more complex three-component median mechanism [7]. Its mark price takes the median of the following three independent components: the first is the oracle price plus a 150-second EMA of the difference between Hyperliquid's weighted mid price and the oracle price; the second is the weighted mid price of Hyperliquid's own order book; and the third is the median of the perpetual futures prices across multiple centralized exchanges. In this architecture, the object smoothed by the 150-second EMA is the basis (the difference between Hyperliquid's mid price and the oracle price) rather than the price itself, which means that when the oracle price moves in tandem, the mark price follows the oracle price instantaneously, with the lag manifesting only in the adjustment of the basis. Its EMA decay formula uses a time constant of about 2.5 minutes (150 seconds) (decay follows $\exp(-\mathrm{t}/150\ \mathrm{s})$) and updates roughly every 3 seconds as validators publish the oracle price [7]. Hyperliquid's oracle is operated by its own validator set rather than an independent external oracle network; the security implications of validators simultaneously bearing the dual roles of consensus and price feeding are discussed further in Section 14.5.4. This design leans heavily toward robustness, but because it relies on multiple external sources and a long smoothing window, it exhibits significant reaction latency under extreme conditions. dYdX's fourth version, by contrast, embeds price aggregation and a consensus mechanism within its validator nodes [26], leaning toward robustness and timeliness; it eliminates dependence on an external oracle but introduces the potential governance risk of validator centralization.

14.4.3 The latency effect of EMA smoothing

To seek a balance within the trilemma, most exchanges introduce an EMA algorithm to smooth the mark price. Its recursive formula is $M_t = \alpha \cdot P_t + (1-\alpha) \cdot M_{t-1}$, where $\alpha = 1 - e^{-\Delta t/\tau}$, τ is the time constant (the e-folding time, with half-life $t_{1/2} = \tau \ln 2 \approx 0.69\tau$), and Δt is the update interval. In a step-down scenario, if the last traded price undergoes a sudden change of magnitude ΔP at time t_0, the mark price's response at

time t is $M_t = P_0 + \Delta P \cdot (1 - e^{-(t-t_0)/\tau})$. This formula shows that at $t = \tau$ the mark price reflects only about 63% of the price change, and at $t = 0.2\tau$ only about 18%. Taking Hyperliquid's $\tau = 150$ seconds as an example, if the last traded price falls 10% within 30 seconds, the mark price falls only about 1.8% over the same period (30 seconds is 0.2τ, so 1.8% here equals 18% × 10%, the same point expressed two ways as the earlier "about 18% at $t = 0.2\tau$"). The 1.8% here rests on a simplified model that applies smoothing to the price itself; in Hyperliquid's actual architecture the smoothing acts on the basis rather than the price, and the mark price's actual response also depends on the oracle update frequency (about once every 3 seconds)—the lag is smaller when the oracle closely tracks the last traded price, and the two layers of lag compound when the oracle itself is delayed. The choice of the decay coefficient α and the time constant τ constitutes the most concrete parametric governance decision: the larger τ is, the more slowly the mark price changes and the stronger its immunity to flash crashes and short-term manipulation, but the more severe the latency effect it produces in a rapidly one-directional market.

Figure 14-6 simulates the direct impact of different smoothing-window lengths on the distribution of liquidation events in a rapidly falling market.

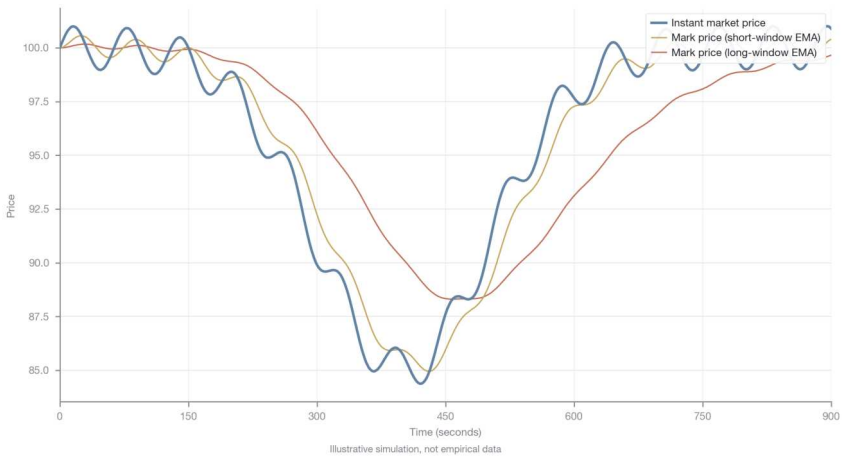


Figure 14-6. The effect of smoothing-window length on the temporal distribution of liquidation events

When the market price falls rapidly, the mark price is often significantly higher than the last traded price, making a trader's margin ratio appear safer on paper than it actually is. Liquidation is therefore triggered with a delay, and when the mark price finally catches up to the last traded price, the liquidations that should have occurred dispersed across different price levels are artificially compressed into a concentrated burst at a single point in time, sharply increasing the intensity of the liquidation cascade. As the worked example above shows, this lag of the mark price means that a large number of liquidations that should have been triggered at different price levels are temporarily backlogged, awaiting the mark price's slow catch-up.

In a reversal, this smoothing mechanism also produces a cumulative-release effect: when a rapid fall is quickly followed by a rebound, the lagging mark price may trigger delayed

liquidations that should not have occurred—this *mark price trap* is developed in detail in Section 14.4.4. Conversely, if a system adopts a short-window EMA, the mark price can closely track the last traded price and timeliness improves substantially, but it becomes extremely sensitive to brief price manipulation and flash crashes. A manipulator need only push the last traded price for a very short time, and the mark price will quickly follow and trigger liquidations, which significantly lowers the cost of attack and widens the system's attack surface. Binance's decision in 2025 to shorten the basis moving-average window successively from 2.5 minutes to 1 minute and then to 30 seconds [24] is exactly such a recalibration between the latency effect and manipulation sensitivity.

14.4.4 The mark price trap

The EMA smoothing mechanism creates a phenomenon in reversals known as the mark price trap. In a market environment of a violent fall accompanied by a rapid rebound, the lag of the smoothing algorithm causes the mark price to keep running downward even after the true market price has recovered significantly. This algorithmic inertia produces a retroactive effect in which the price has recovered but the position is still liquidated.

Figure 14-7 dissects the mechanism of this trap by comparing the trajectory of the mark price under long-window and short-window EMA parameters. The red shaded region marks the time window in which the mark price is below the liquidation line while the last traded price has already recovered above it—that is, the interval in which retroactive liquidation occurs.

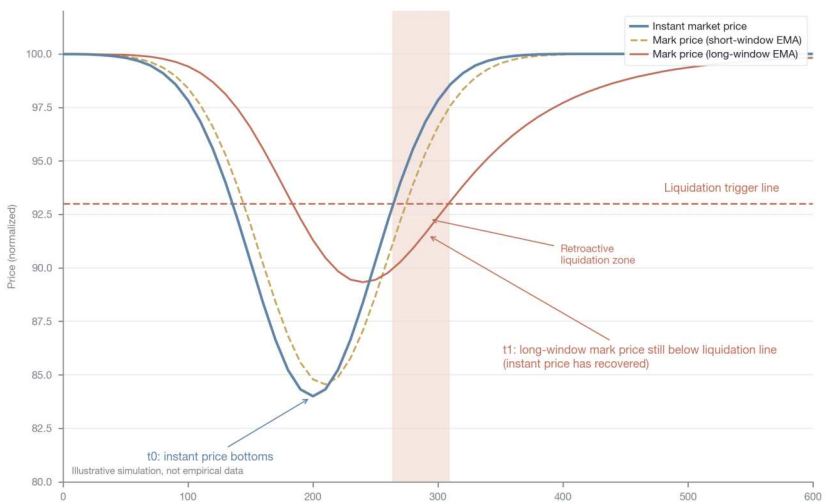


Figure 14-7. The mark price trap and retroactive liquidation in a V-shaped reversal

During the phase of a sharp fall, the true price bottoms out rapidly and the mark price begins to follow it down, but exhibits a clear lag because of the smoothing algorithm. Subsequently, the true price rebounds, but by this time the mark price is still digesting the earlier steep drop and continues to probe downward. When the mark price finally touches the trader's liquidation line, the user's position is forcibly liquidated. Only

afterward does the mark price begin to recover slowly, but the liquidation is already irreversible and the trader's substantive loss has been incurred. The red shaded region in Figure 14-7 marks precisely this trap interval: the last traded market price is already significantly above the liquidation line, yet the mark price computed by the long-window EMA still hovers below it, causing retroactive liquidation to be executed at a stage when the market appears to have already recovered.

The inertia of the EMA is meant to serve as a buffer that protects traders in a one-directional plunge, but in a reversal it becomes a mechanism of harm. Observing that the last traded price has already recovered, traders often mistakenly believe their positions are already safe, overlooking the hidden risk that the mark price is still approaching the liquidation line. In the end, traders are retroactively liquidated at the very moment the crisis appears to have passed. This phenomenon reveals a profound governance implication: choosing the length of the smoothing window is essentially choosing which class of market conditions' users to protect—an implicit risk-allocation decision. Even a trader who judges the market direction perfectly correctly and successfully buys the bottom may still be liquidated for lack of a full understanding of the mark price's algorithmic behavior. Beyond retroactive liquidation in a V-shaped reversal, the mark price can also produce a symmetric *mark price illusion* in a slow trending market: EMA smoothing keeps the mark price persistently below the last traded price during a slow rise, giving traders the illusion that their margin is more ample than it really is. When the trend suddenly reverses, the mark price's *catch-up* effect leads to a larger-than-expected concentration of liquidations. In this specific scenario, the trilemma resolves into a dilemma between protecting the majority from manipulation and protecting the minority from delayed liquidation. Although a short-window EMA can effectively avoid the mark price trap, its high sensitivity to the last traded price—as shown by the orange curve in Figure 14-7—means that it produces more false liquidation triggers amid everyday volatility.

The mark price trap not only causes immediate liquidation losses but also produces lasting microstructure effects through behavioral feedback. Rational traders who repeatedly experience retroactive liquidation will systematically reduce their leverage use or set stop-losses at distances farther than necessary, even when such conservative parameter settings are not optimal from a fundamental standpoint. This experience effect reduces the overall market's leverage supply and liquidity depth, and its long-term welfare loss exceeds that of any single liquidation event.

14.4.5 Liquidation hunting and institutional isolation

In the early derivatives markets that lacked mark price protection, liquidation hunting was a high-frequency and highly impactful attack method. When a liquidation system relied directly on the last traded price, an attacker would first scout for a target market with shallow order-book depth, a concentration of high-leverage positions, and thin liquidity. Having selected the target, the attacker would suddenly deploy a massive market sell order during the period of lowest liquidity, directly consuming all of the bid-side resting orders and causing the last traded price to plunge in a cliff-like drop, forming the classic *wick* pattern on the candlestick chart. This instantaneous price

collapse would immediately trigger the risk-control system tied to the last traded price, forcibly liquidating a large number of high-leverage long positions. The market sell orders generated by the liquidation engine would push the price down further, forming a destructive liquidation cascade. In the end, the attacker would profit through a short position established in advance, or buy in at the extremely low prices of the panic sell-off.

The introduction of the mark price mechanism fundamentally changed this attack-defense dynamic through institutional isolation. Along the spatial dimension, the multi-source aggregation of the index price makes the price shock an attacker creates in a single derivatives market negligible for the global index; to affect the index price significantly, an attacker would have to manipulate all constituent spot markets simultaneously, so the cost of attack grows exponentially. Along the temporal dimension, the mark price's moving-average algorithm provides a further buffer: even if the derivatives market's last traded price momentarily deviates sharply from the index price and the basis widens abruptly, the smoothing dilutes the impact of this transient anomalous basis on the final mark price enormously, so the mark price may move only minutely—far too little to touch most traders' liquidation lines.

The Bitstamp flash crash of May 2019 is a classic cautionary tale of insufficient institutional isolation. At the time, a massive sell order of approximately 5,000 BTC instantly drove the platform's spot price from about $7,900 down to about $6,180, a drop of approximately 20% [27]. Because BitMEX's price index was then composed of only two exchanges, Bitstamp and Coinbase, equally weighted, this local price manipulation directly dragged down the global index and triggered more than $200 million in long liquidations. Brave New Coin's multi-exchange index only briefly dipped below $6,660 in the same period, whereas BitMEX's index momentarily touched $6,382 [27]; the gap of nearly $300 between the two fully illustrates the decisive influence of the number of index constituents on robustness. This event directly spurred industry-wide index reform, with mainstream platforms expanding the number of index-constituent exchanges and introducing outlier-filtering mechanisms.

As platforms today introduce broader index constituents and more complete smoothing mechanisms, transient price manipulation targeting a single derivatives order book usually ends in failure. The mark price stays stable, avoiding the unfair liquidation of innocent positions, while the attacker may instead suffer enormous slippage losses from dumping at low prices in an illiquid market. The mark price mechanism is not a simple technical optimization but a fundamental institutional design: it shifts the authority to adjudicate risk control from noisy local order-book signals to a highly robust global price signal, significantly raising the barrier to market manipulation.

14.4.6 Platform comparison

The methods by which different derivatives exchanges compute the mark price embody markedly different design philosophies, yet their core idea remains highly consistent: to take an externally, spatially aggregated spot price as the absolute anchor and to absorb short-term irrational swings through a temporal smoothing mechanism. Table 14-2 compares the mark price mechanisms of mainstream trading platforms.

Platform	Core algorithmic architecture	Smoothing parameter (window / time constant)	Number of price sources	Update frequency	Governance preference
Binance	Median of the index price, adjusted price, and last traded price	30-second window (since Sept 2025)	10+	~1 second	Balance of accuracy and timeliness
OKX	Index price plus a basis moving average	Variable	7+	~1 second	Accuracy and robustness
Hyperliquid	Median blend of oracle, local order book, and multi-CEX prices	150 seconds	3 input components (10+ underlying sources)	~3 seconds†	Robustness first
dYdX v4	Validator-node consensus aggregation	No explicit EMA (median aggregation + ~1–2 s consensus latency)	Validator aggregation	~1 second†	Decentralization and robustness
GMX	Chainlink and Chaos Labs decentralized oracles	Heartbeat-triggered	Multiple nodes	0.5% deviation or 27-second heartbeat	Robustness and accuracy

Table 14-2. Comparison of mark price mechanisms across mainstream trading platforms (Data source: compiled by the author from platforms' public documentation)

Note: † The figures shown are nominal price-publication frequencies, which may not be synchronized with the liquidation engine's check frequency; the actual frequency may decrease under extreme conditions. The Binance formula described applies to the normal-market state of USDT-margined perpetual futures; under extreme market conditions it may switch to a degraded logic.

Table 14-2 has already listed each platform's choice within the mark price trilemma; here we add only two differences not developed in Section 14.4.2. First, Binance's median-of-three formula contains an implicit form of outlier filtering: when any one input deviates abnormally, the median mechanism automatically excludes it from the final result, giving natural resistance to a single-source anomaly [24].

Second, Hyperliquid's most complex three-component median mechanism [7] exposed, in the JELLY attack of March 2025 (see Section 14.7.5), the fragility of such complex mechanisms when liquidity is extremely depleted: the attacker neither tampered with the smart contract nor exploited a code vulnerability but—relying solely on a thorough understanding of the mark price computation logic—contaminated the oracle through compliant spot manipulation, skewed the mark price via a loophole in the median aggregation, and forced the platform into an emergency delisting of the token [28].

The liquidation impact of the mark price also depends on the margin architecture. Under isolated-margin mode, a change in the mark price affects only the margin ratio of a single position; under the cross-margin or unified-margin mode now widely adopted by mainstream platforms, a small swing in the mark price affects the total margin adequacy of the entire account and the cross-asset offsetting calculation, so liquidation sensitivity rises significantly.

14.4.7 The intensity of liquidation cascades

The choice of mark price parameters determines not only the fate of individual traders but also, directly, the intensity of systemic liquidation cascades. As described in Section 14.4.3, a longer-window smoothing parameter takes the liquidations that should have fired at dispersed price levels as the market fell and compresses them into a single concentrated burst: when the mark price finally breaks through a key threshold, a large number of positions are liquidated at once within an extremely short time, the single-round liquidation volume increases sharply, and the market's self-reinforcing divergence condition is readily met.

In the crypto-market liquidation cascade that erupted on October 10, 2025, more than $19 billion in leveraged positions were forcibly liquidated within about 24 hours, of which long liquidations accounted for approximately 87%, affecting about 1.6 million traders [1]. Bitcoin's price plunged from about $122,000 to about $104,000, and Ethereum fell about 12% over the same period [1]. This event reveals how a smoothing mechanism originally intended as a buffer can, through delay and concentrated release, instead become a catalyst that intensifies a systemic collapse in an extreme one-sided move. Ali's (2025) [1] study notes that the trigger of this liquidation cascade was an escalation of tariffs between the United States and China, but that its intensity was largely amplified by the microstructure of the derivatives market: the latency effect of the mark price caused liquidation events to become highly concentrated along the time dimension, forming a self-reinforcing downward spiral. Figure 14-8, based on the flash-crash data of October 2025, shows the dynamic divergence between the last traded price and the mark price within a 5-minute window: the last traded price fell rapidly while the mark price lagged significantly because of the EMA buffer, and the positions within the price gap formed between them were temporarily protected by the smoothing mechanism—but they also constituted the backlog that would later be concentrated into liquidations as the mark price caught up.

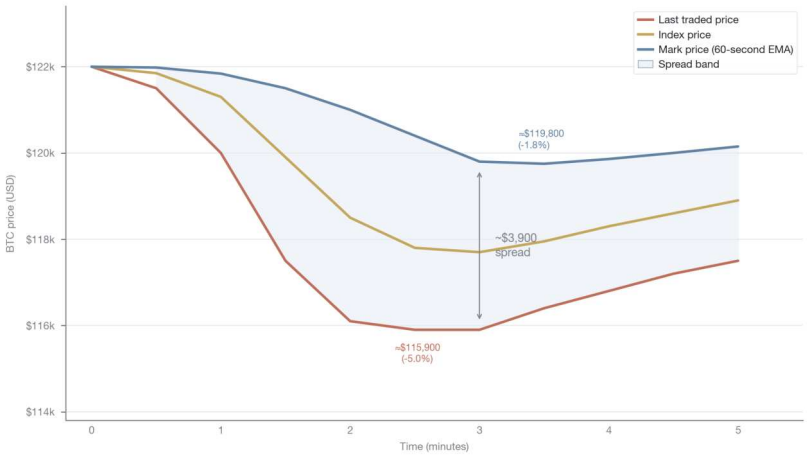


Figure 14-8. The dynamic divergence between mark price and last traded price in a flash-crash scenario

Conversely, if an exchange leans toward timeliness and chooses a shorter smoothing window, the mark price will track market swings extremely sensitively. Under such a mechanism, liquidation events are triggered in a more dispersed manner, each round of forced closure releases relatively little selling pressure to the market, and the divergence condition for systemic risk is harder to meet. Yet such a sensitive mark price, amid everyday choppy conditions, is highly prone to triggering a large number of unnecessary liquidations because of a brief price wick, harming the interests of individual traders. This is a profound and counterintuitive governance trade-off: a more robust design lowers the probability of an individual being improperly liquidated but raises the risk of systemic collapse, whereas a more sensitive design protects the stability of the system but at the cost of sacrificing individuals' safety amid everyday volatility. Protecting individuals and protecting the system here point to entirely opposite design choices, which constitutes the core normative challenge in mark price mechanism design. This trade-off cannot be dissolved by purely technical means, because it is essentially a normative question about risk allocation: in an inevitable extreme event, whose interests should the system prioritize?

This analytical framework must also incorporate the divergence between the mark price and the liquidation execution price. The mark price determines the timing of the liquidation trigger, but the actual closing is executed as market orders against an order book that may already have severely contracted. In the October 2025 event, in an environment where order-book depth contracted by more than 90%, the execution slippage of large liquidated positions may have far exceeded the construction error of the mark price itself. Therefore, measuring the true impact of mark price mechanism design on trader welfare should examine not only the deviation of the mark price from fair value but also the execution gap between the moment the mark price triggers and the actual liquidation execution price.

14.5 Oracle architecture and the information bottleneck

Smart contracts are deterministic in their execution logic, but their runtime environment is essentially a closed system, unable to perceive the state of the external world on its own. For an on-chain financial protocol to provide financial services priced on real-world assets, it must rely on a mechanism that transmits off-chain data to the chain safely and reliably. This information-transmission mechanism is the oracle. In the context of DeFi, an oracle is not merely a simple data-transmission channel but the core node that determines the security of the entire system. If the data an oracle provides are biased or maliciously manipulated, then however well the on-chain protocol's liquidation engine and risk-control models are designed, the result will inevitably be erroneous liquidations, asset loss, or systemic collapse. The oracle is therefore regarded as the security bottleneck of the on-chain financial ecosystem.

As the total value locked (TVL) in DeFi continues to climb, the importance of oracles within the entire ecosystem grows ever more prominent. As of the first quarter of 2026, leading oracle networks secured more than $100 billion in on-chain assets, and the price data they provided triggered trillions of dollars in transactions and liquidation instructions [29]. Yet this large asset scale also makes oracles the primary target for attackers. Although oracle manipulation is not the largest source of DeFi losses in 2025 (by Immunefi's accounting, pure manipulation attacks fell from nearly 19% of DeFi losses in 2022 to about 1% in 2025, and the dominant loss types that year were private-key compromise and phishing), the destructive power of a single incident and its high degree of targeting at low-liquidity long-tail assets make it one of the core threats that protocol designers must confront [30]. This asymmetric attack-defense posture forces protocol designers into difficult trade-offs in their choice of oracle architecture.

14.5.1 The oracle trilemma

Oracle design is not merely a matter of engineering implementation but is constrained by a fundamental trilemma. Theoretical research shows that in constructing a decentralized oracle network, designers must trade off among three dimensions—accuracy, manipulation resistance, and timeliness—and cannot reach a theoretically optimal state on all three at once [31]. This trilemma fundamentally determines the evolutionary direction of different oracle architectures and their suitability for specific financial scenarios.

Accuracy requires that the on-chain price an oracle provides reflect, to the greatest extent possible, the true trading state of the off-chain spot market. To improve accuracy, an oracle typically needs to obtain prices from multiple independent data sources and use complex aggregation algorithms to remove outliers. Manipulation resistance concerns the oracle's robustness in the face of capital attacks—that is, how large an economic cost an attacker must pay to make the oracle output a price that deviates from the true market level. Improving manipulation resistance often requires introducing more data providers, longer observation windows, or stricter staking-and-slashing mechanisms. Timeliness measures the time delay for an off-chain price change to be reflected on-chain. In a rapidly volatile market, timeliness directly bears on whether the liquidation mechanism can trigger in time and whether arbitrageurs can effectively

close cross-market spreads.

There are inherent conflicts among these three dimensions. The multi-source aggregation and time-weighted averaging mechanisms introduced to improve manipulation resistance inevitably increase the latency of data processing, thereby sacrificing timeliness. Conversely, using a single data source or an instantaneous price in pursuit of the utmost timeliness sharply lowers the cost for an attacker to manipulate the price, impairing the system's manipulation resistance. In existing DeFi practice, no oracle scheme can perfectly resolve this dilemma; all architectural innovations are essentially a rebalancing of these three dimensions under specific business requirements.

14.5.2 Push-based architecture

The push-based architecture is the oracle model most widely adopted in the early days of DeFi; its core logic is that a network of oracle nodes actively pushes price-data updates onto on-chain smart contracts. Under this architecture, nodes continuously monitor the price movements of a target asset off-chain, and when the price fluctuation exceeds a preset deviation threshold, or when the time since the last update exceeds the limit of the heartbeat interval, the node network initiates a consensus process. After reaching agreement through an off-chain reporting protocol, the nodes package the aggregated price data into a transaction and send it to the blockchain network for all protocols relying on that oracle to read [32].

The greatest advantage of this design lies in its extremely high manipulation resistance and broad compatibility. Because the data have been signed and aggregated by multiple nodes off-chain, the failure of a single data source or the malfeasance of a single node can hardly have a substantive effect on the final on-chain price. At the same time, because the price data are stored directly in on-chain state, any smart contract can read these data with zero latency, greatly simplifying the development process of decentralized applications.

The limitations of the push-based architecture, however, are equally significant. To maintain high-frequency price updates, the oracle network must continuously consume large amounts of network fees. When the blockchain network is congested, oracle update transactions may sit waiting for a long time because of insufficient fees, causing the on-chain price to lag badly behind the real market. Moreover, constrained by on-chain computation and storage costs, push-based oracles can typically support only a limited number of asset instruments and cannot provide millisecond-level price updates. This compromise in timeliness and scalability makes the push-based architecture ill-suited to the needs of price-sensitive applications such as derivatives trading.

14.5.3 Pull-based architecture

To break through the bottlenecks of the push-based architecture in update frequency and cost, the pull-based oracle architecture emerged. In this model, oracle nodes no longer actively push prices on-chain but continuously generate cryptographically signed price reports off-chain. When a decentralized application's user needs to execute a price-dependent transaction, the user or the front-end interface fetches the latest price report from the oracle's off-chain data-availability layer and packages it together with

the business transaction for submission to the blockchain. Before executing the business logic, the smart contract first verifies the report's signature and timestamp, and only after confirming their validity does it use that price in its computation [33].

Figure 14-9 shows the core differences in data-flow paths among the three oracle architectures: push-based, pull-based, and app-chain native.



Figure 14-9. Comparison of the data-flow paths of the three mainstream oracle architectures

The pull-based architecture fundamentally changes the economic model of the oracle. The cost of a price update is no longer borne by the oracle network but transferred to the user who actually uses that price, achieving a precise match between cost and benefit. This design entirely removes the constraint that network congestion places on price-update frequency, enabling the oracle to provide high-frequency price data for hundreds or thousands of assets at extremely low latency. For a perpetual futures DEX that needs millisecond-level response, the pull-based architecture provides the necessary infrastructural support.

Despite its major breakthroughs in timeliness and scalability, the pull-based architecture also introduces a new risk dimension. Because the trigger for a price update rests in the user's hands, a malicious trader may exploit this mechanism for selective updating. In a liquidation scenario, for example, if the latest price is unfavorable to the borrower, the borrower may deliberately avoid triggering a price update in an attempt to delay the liquidation process. Although a protocol can mitigate this problem by

introducing independent liquidation bots, this increases the overall complexity of the system. In addition, the pull-based architecture requires the smart contract to verify multiple signatures in real time within every transaction, which may become a new performance bottleneck on some blockchains with limited computational resources.

14.5.4 App-chain native oracles

As application-specific blockchains have proliferated, a new architecture that embeds the oracle mechanism directly into the blockchain consensus layer has gained adoption. In this deeply coupled model, the oracle is no longer an external service independent of the blockchain but becomes a native responsibility of validator nodes. While participating in block packaging, validators must submit the external asset prices they observe. The consensus mechanism votes not only on transaction ordering but also on the median aggregation of these price data, and the resulting price is written directly into the block header for the on-chain state machine to use [34].

The native-oracle architecture pushes timeliness to the extreme. Because price updates proceed in step with block generation, the latency of the on-chain price is limited only by the block time and can typically reach the sub-second level. More importantly, this architecture entirely eliminates the transaction fees required for oracle updates, because the price data are no longer transmitted through ordinary smart-contract transactions but take effect directly as a byproduct of the consensus process. This zero-cost, high-frequency price-acquisition capability enables an app-chain to support extremely high-leverage derivatives trading while sharply reducing the systemic risk posed by liquidation latency.

Yet this design, which deeply binds financial logic to the underlying consensus, also raises profound questions about the system's security boundary. A native oracle requires validators to take on extra data-acquisition and processing work, increasing the complexity of node operation. If some validators cannot provide accurate prices because of data-source failures or network latency, it may affect the consensus process of the entire blockchain, causing block production to stall. Moreover, this architecture forgoes the generality of the oracle; the price data it provides can be used only within the specific app-chain and cannot serve the broader multi-chain ecosystem. This concession in generality is the price that must be paid in pursuit of ultimate performance.

The app-chain native-oracle architecture raises a deeper question about the incentive compatibility of validators. When validators bear the dual responsibilities of block production and price submission, a validator holding derivatives positions faces an economic inducement to trigger favorable liquidations by submitting a shifted quote. Furthermore, in systems such as dYdX v4 and Hyperliquid, validators can also extract returns similar to maximal extractable value (MEV) by selectively delaying or advancing the submission of price updates. The severity of this incentive conflict depends on the size of the validator set and the strength of the penalty mechanism. dYdX v4's validator cap was lowered from 60 to 50 in mid-2025, with roughly 30-odd actually active, and Hyperliquid's validator set is even more limited in size; both differ by orders of magnitude in decentralization from the hundreds to thousands of validators on general-purpose chains, and a smaller validator set significantly lowers the threshold

for multiple nodes to collude in manipulating price quotes.

Beyond the three architectures above, all of which depend on external data sources, the TWAP oracle based on an on-chain automated market maker (AMM) constitutes an entirely different paradigm. This mechanism relies on no off-chain data feed; instead, it computes the price directly from the liquidity pool of an on-chain DEX. In Uniswap V3, for example, a built-in price accumulator records the logarithm of the instantaneous price at every trade, and an external protocol can compute the geometric mean price over an arbitrary window length by reading the difference in the accumulator between two points in time. Architecturally, this design eliminates the oracle's trust-boundary problem, because the price data are generated entirely by on-chain trading activity, with no need for a validator network or signature verification. However, the manipulation cost of an on-chain TWAP oracle is proportional to both the depth of the liquidity pool and the length of the time window: an attacker must sustain a disequilibrium trading price throughout the entire TWAP window, which means that the manipulation cost grows linearly with window length—but at the price of a latency that scales proportionally with that same window length. In the approximately $34 million attack on Harvest Finance in October 2020, the attacker used a flash loan to sharply distort the price in Curve's USDC/USDT liquidity pool and exploited the vulnerability of the excessively short window of the on-chain TWAP oracle that Harvest relied on, completing the price manipulation and the malicious deposits and withdrawals within a few transactions [35]. This incident demonstrates that the security of an on-chain TWAP oracle depends heavily on the depth of the underlying liquidity pool, not merely on the length of the time window. Building on the order-book and AMM microstructure discussed in Chapter 7, the on-chain TWAP oracle can be viewed as an institutional arrangement that translates market makers' quoting behavior directly into a price benchmark; its governance implication is that it shifts the benchmark's credibility from the honesty of validators to the capital commitment of liquidity providers.

14.5.5 The cross-layer information bottleneck

Whatever architecture is adopted, the process by which an oracle transmits off-chain prices on-chain inevitably entails a lossy compression of information. This phenomenon is called the oracle cross-layer information bottleneck, and it is the core theoretical foundation for understanding the fragility of on-chain finance. In traditional centralized financial markets, price is a continuous, high-precision, multi-dimensional information flow rich in order-book depth. Yet when this information enters the blockchain through an oracle, it is forced to undergo an irreversible compression along three dimensions: time, precision, and source.

Along the time dimension, because of the blockchain's discrete block-production mechanism and the trigger conditions limiting oracle updates, the continuous price flow is truncated into discrete data points. Within the time window between two oracle updates, the on-chain protocol is in a blind spot with respect to external market changes. Along the precision dimension, to lower on-chain computation costs and reach consensus among nodes, the complex order-book structure and minute price fluctuations are typically simplified into a single aggregate scalar value, masking the true depth of mar-

ket liquidity. Along the source dimension, the price data of hundreds or thousands of centralized exchanges and decentralized liquidity pools are ultimately condensed into a single unified benchmark price, erasing the microstructure differences between markets.

The consequences of this information bottleneck are systemic. Because all protocols relying on the same oracle share this compressed informational reality, the oracle effectively becomes the single point of failure for the entire on-chain financial ecosystem. When the off-chain market swings violently and the oracle, constrained by the bottleneck, cannot transmit this swing in a timely and accurate manner, the on-chain protocol's risk-control mechanisms make decisions based on stale or distorted price data. This systemic misjudgment, arising from information asymmetry, is precisely the root cause of large-scale liquidation cascades and pool depletion. However rigorous the logic of an on-chain protocol's design, its security ceiling is always constrained by the quality of the oracle's cross-layer information transmission.

14.5.6 Latency and information asymmetry

The most direct manifestation of the oracle cross-layer information bottleneck is the latency of price updates, which creates a brief but significantly impactful window of information asymmetry between on-chain and off-chain. When the true price in the external market suddenly changes while the oracle's on-chain price has not yet been updated, an arbitrageur with faster information channels can clearly see the future direction of the on-chain price. This time difference confers on the arbitrageur a structural informational advantage. In actual execution, frictions such as the uncertainty of gas costs, the front-running competition of MEV bots, the execution risk of smart contracts, and the protocol's position limits prevent this theoretical informational advantage from being fully captured—which is precisely why the oracle extractable value discussed in Section 14.5.7 is partly dissipated through MEV competition.

Figure 14-10 depicts the window of information asymmetry created by oracle update latency when the true off-chain price suddenly changes.

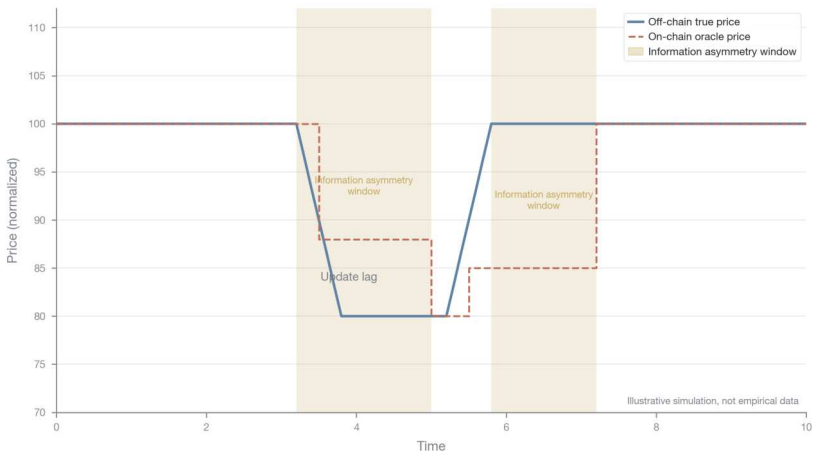


Figure 14-10. Oracle update latency and the formation of the information-asymmetry window

In a lending protocol, if the true price of the collateral asset plunges but the oracle price still sits at a high level, a malicious user can, in the few seconds before the oracle updates, use the overvalued collateral to borrow a large amount of assets and then let the position be liquidated after the oracle updates. On a derivatives exchange, if the true price surges, a trader can establish a long position while the on-chain price has not yet reacted and close it for a profit immediately after the oracle completes its update. This behavior essentially shifts the oracle's latency cost onto the protocol's liquidity providers and other honest users.

To counter this latency arbitrage, protocol designers typically introduce more conservative parameter settings, such as lowering the maximum leverage, raising the initial collateralization ratio, or extending the withdrawal lock-up time. Yet these defensive measures inevitably reduce capital efficiency and harm the user experience. The deeper problem is that as long as oracle updates have a physical time interval, this window of information asymmetry cannot be entirely eliminated. The speed race between arbitrageurs and oracle nodes becomes a zero-sum game that consumes vast resources, with the result of continually driving up the operating cost of the entire ecosystem.

14.5.7 Oracle extractable value

An oracle update transmits not only price information but also determines the timing of on-chain state transitions. In a lending protocol, a change in the oracle price is the sole condition that triggers the liquidation procedure. When the oracle submits a price that renders a borrowing position insolvent, whoever can be the first to execute the liquidation transaction reaps a rich liquidation penalty. This risk-free profit, directly spawned by the oracle update, is called oracle extractable value (OEV) [36].

Figure 14-11 shows the extraction and redistribution flow of OEV under the traditional path and under the order-flow-auction path.

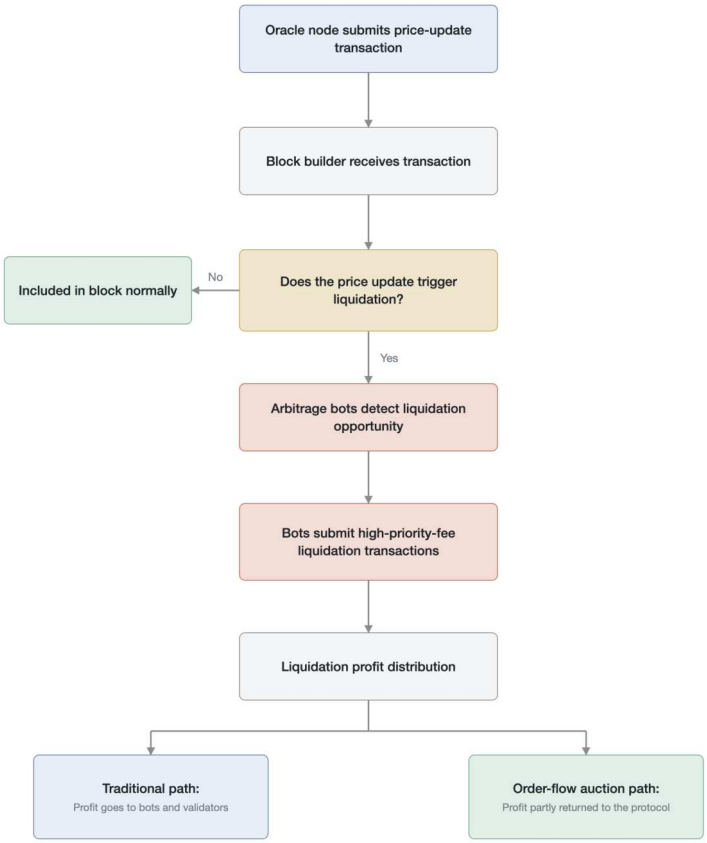


Figure 14-11. The extraction and redistribution of oracle extractable value

OEV reshapes the underlying incentive structure of the blockchain network. To compete for the right to liquidate, arbitrage bots—within the same block as the oracle-update transaction—try to have their own liquidation transaction executed immediately after the oracle update, either by paying an extremely high network priority fee or by directly bribing the block builder. This fierce bidding not only causes network congestion and a surge in ordinary users' transaction costs but also transfers value that would otherwise belong to the DeFi protocol to the underlying miners or validators.

To reclaim this leaked value, oracle networks have begun to explore new economic models. One representative scheme introduces an order-flow auction mechanism that allows liquidators to bid off-chain for the right to execute an oracle update. The highest-bidding liquidator obtains the exclusive right to package the oracle update together with its liquidation transaction into a single atomic operation for submission on-chain. The proceeds of the auction are then returned to the lending protocol or the oracle network according to a predetermined ratio [37]. This mechanism transforms the timing of the oracle update itself into a quantifiable, distributable economic asset,

marking the evolution of oracle design from a pure data-transmission layer toward a complex economic-game layer.

14.5.8 The cost-benefit of oracle manipulation

An oracle may fail not only because of technical faults or network latency but also faces the severe threat of deliberate manipulation. An oracle-manipulation attack is essentially a precisely calculated economic game: the attacker conducts large trades in the off-chain spot market or on-chain liquidity pools to artificially distort the target asset's price, forcing the oracle to transmit this false price on-chain. The attacker then uses this manipulated on-chain price to borrow maliciously in a lending protocol or to trigger erroneous liquidations in a derivatives market, thereby seizing enormous profits.

The core of whether an attack is feasible lies in the comparison of cost and benefit. The cost of attack mainly comprises the capital slippage needed to move the price in the spot market, transaction fees, and the friction cost of cross-market hedging. Because oracles typically use multi-source aggregation and time-weighted averaging algorithms, an attacker must maintain the anomalous price for a sufficiently long time and across multiple major liquidity centers at once, which greatly raises the manipulation cost [38]. However, with the spread of uncollateralized borrowing tools such as flash loans, an attacker can mobilize hundreds of millions of dollars within a single block and instantly break through a weak liquidity pool, so the originally very high capital barrier loses its binding force.

Figure 14-12 compares the cost-benefit curves of launching an oracle-manipulation attack against a high-liquidity asset versus a low-liquidity asset.

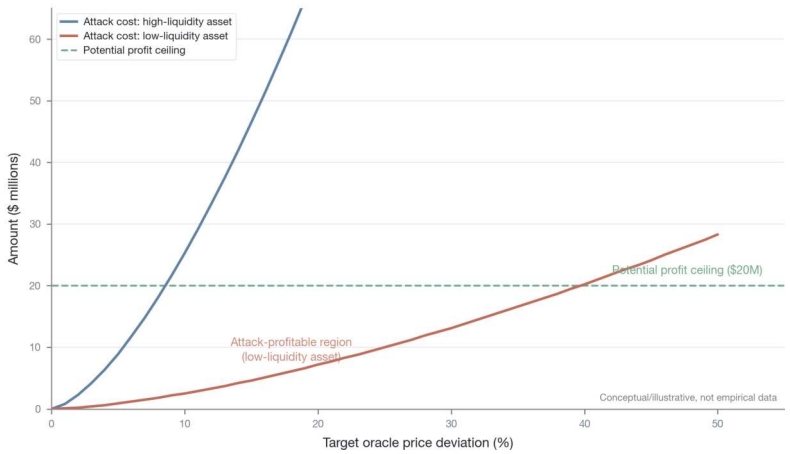


Figure 14-12. Cost-benefit analysis of oracle manipulation attacks

The potential benefit of an attack depends on the on-chain protocol's risk exposure. If a lending protocol allows users to use an extremely illiquid long-tail token as collateral to borrow highly liquid stablecoins, then an attacker need only spend a minimal cost to ramp the long-tail token's price and can drain the protocol's entire stablecoin reserve.

In the 2022 attack on a decentralized lending-and-derivatives protocol, the attacker extracted approximately $116 million in liquidity within just a few dozen minutes by manipulating the oracle price of a governance token [3]. This extremely high return-on-investment ratio shows that defending against oracle manipulation cannot rely solely on strengthening the robustness of the oracle itself but requires the on-chain protocol to build a multi-layered defense-in-depth system across asset admission, debt ceilings, and liquidity-depth monitoring.

14.6 Feedback loops and systemic transmission

In the on-chain financial system, a price benchmark is not merely a passively read data point but a dynamic variable that drives the operation of the entire system. When we examine a price benchmark within the dynamic evolution of market microstructure, we find that it not only reflects market state but also shapes market behavior. Under the joint action of leverage and liquidity, a price benchmark becomes the hub of both positive-feedback and negative-feedback loops. Liquidation spirals, funding-rate arbitrage, and oracle synchronization congestion are all, in essence, the systemic results of a price benchmark transmitting through different feedback mechanisms. Understanding the trigger conditions and transmission paths of these loops is a prerequisite for guarding against on-chain systemic risk.

14.6.1 Positive feedback and the liquidation spiral

In the finance literature, Brunnermeier and Pedersen (2009) decompose the margin spiral into two mutually reinforcing sub-mechanisms: the loss spiral and the margin spiral [39]. The transmission logic of the loss spiral is that a price fall causes leveraged traders' existing positions to incur unrealized losses; the decline in their capital adequacy forces them to cut positions, and the resulting selling pressure depresses the price further. The transmission logic of the margin spiral is different: a price fall raises market volatility, and the risk-control system accordingly raises its margin-ratio requirement, so that even if a trader's existing position has not yet touched the liquidation line, the higher margin requirement forces them to reduce positions proactively to satisfy the new capital constraint. In traditional finance, these two sub-mechanisms are coupled through market makers' funding constraints. Gromb and Vayanos's (2002) [40] study shows that when arbitrageurs face funding constraints, their ability to dampen price deviations declines, market liquidity deteriorates accordingly, and the price shock is further amplified. The model also implies a contagion channel that is often overlooked: when market makers and proprietary trading firms operate simultaneously in both crypto and traditional financial markets, a funding-constraint shock in the crypto market may transmit to other asset classes through the balance sheet of the common intermediary, and vice versa. Brunnermeier and Pedersen's original model assumes that margin requirements adjust dynamically with value at risk (VaR)—that is, rising volatility drives up the margin ratio and forces traders to reduce positions before the price reaches the liquidation line. Crypto perpetual futures markets, however, generally adopt a fixed-leverage margin regime, in which the initial margin is the reciprocal of the leverage multiple and does not adjust in real time with volatility. This means that the

transmission path of the *margin spiral* in the original model—rising volatility leading to higher margin requirements leading to forced position reduction—is replaced in crypto markets by the *loss spiral* path: a price fall directly erodes the margin balance until it touches the fixed liquidation threshold. The dynamics of the two mechanisms have different temporal characteristics: the VaR margin spiral can start at the early stage of a volatility rise, whereas the loss spiral under fixed leverage is triggered only after the price change accumulates to a sufficient magnitude. When this corrected theory is mapped onto on-chain finance, the liquidation spiral exhibits even more deterministic and automated characteristics, because the smart-contract liquidation engine has no buffer of human judgment: once the mark price touches the liquidation line, the closing instruction executes within milliseconds (Chapter 11 provides the mathematical derivation of this mechanism). The liquidation spiral is not merely the result of excessive leverage or the depletion of market liquidity; its fundamental driving force is embedded deep in the design of the price benchmark. The smoothing parameters of the mark price, the update frequency of the oracle, and the degradation strategy of the index together constitute the control variables of the liquidation spiral, determining the spiral's activation threshold, amplification speed, and duration.

The smoothing mechanism of the mark price serves as a buffer in normal markets, but under extreme conditions it can become a catalyst for the liquidation spiral. The *mark price trap* described in Section 14.4.4 manifests in the liquidation spiral as a reflexive closed loop: retroactive liquidation under long-window smoothing forcibly closes positions that had already avoided liquidation, the resulting selling pressure again depresses the true price, and this in turn drags the mark price down further, turning a one-time price shock into sustained systemic selling pressure.

The index's degradation strategy likewise plays a key role in the spiral. When some exchanges' price sources fail because of liquidity depletion or system outages, if the index-construction rules lack a robust exclusion mechanism, anomalous prices are incorporated into the calculation. This distorts the index price and in turn triggers erroneous liquidations by protocols relying on that index. In the market crash of October 2025, more than $19 billion in leveraged positions were liquidated within 24 hours [41]. During this process, the near-touch available liquidity in Bitcoin's order book contracted sharply by more than 90% (on a major-venue, major-pair basis), and the bid-ask spread widened from single-digit basis points to double-digit percentages. This liquidity vacuum and the mechanical execution of the mark price reinforced each other, forming a classic positive-feedback loop.

14.6.2 Negative feedback and the funding rate

In contrast to the positive feedback of the liquidation spiral, the funding rate is the core negative-feedback mechanism in the perpetual futures market. Its core function is to use economic incentives to push the contract price back toward the spot index price. When the contract price is higher than the index price, the funding rate is positive and longs pay a fee to shorts, which incentivizes arbitrageurs to short the contract and buy the spot, thereby depressing the contract price. This algorithmic feedback rule effectively maintains the anchoring between the derivatives market and the spot

market most of the time.

Yet the effectiveness of the funding-rate mechanism rests entirely on the accuracy of the index price. When the index price that serves as the benchmark is itself distorted, this negative-feedback mechanism not only fails but may even turn into destructive positive feedback. If the spot index deviates from true fair value because of an extreme anomaly at an individual constituent exchange, the funding rate is computed based on this erroneous benchmark. Acting on erroneous incentives, arbitrageurs instead push the derivatives price in a direction that deviates from true value.

In extreme market volatility, the failure of the funding rate is often accompanied by the depletion of liquidity. When market participants expect the index price to be unreliable, they demand a higher risk premium, manifested as wider spreads and smaller resting-order sizes. At this point, even if the funding rate offers an enormous arbitrage opportunity, arbitrage capital chooses to exit the market because of excessive slippage costs and a sharp rise in liquidation risk. On the eve of the October 2025 crash, the annualized funding rate in the derivatives market briefly surged to 30%. A 30% annualized funding rate converts to approximately 0.0274% every 8 hours; although this level is far above the 0.01% baseline, its negative-feedback effect is subject to three constraints in a one-sided extreme move. The first is the constraint of settlement frequency: within the 8-hour window between two consecutive settlements, the funding rate exerts no direct economic constraint on the market. The second is the saturation of the rate cap: most exchanges impose a cap on the funding rate. On Binance, for example, the cap is not a fixed value but is set according to the contract's maintenance margin rate (approximately $0.75\times$ the maintenance margin rate for high-leverage contracts, and approximately $\pm 2\%$ for low-leverage contracts); once the degree of market imbalance exceeds the basis level corresponding to that cap, the funding rate can no longer increase to reflect the true extent of the long-short imbalance. Third, when the expected short-term return of extremely leveraged longs far exceeds the cost of the funding rate, the rate mechanism degenerates from a substantive deleveraging incentive into a negligible holding cost. When the market collapses, this steep holding cost accelerates the closing of long positions, the negative-feedback mechanism breaks down entirely, and it instead becomes a force that intensifies the market sell-off.

From a macro-finance perspective, the funding rate can be decomposed into a risk-free interest rate component and a leverage premium specific to crypto markets. During a Federal Reserve rate-hiking cycle, the opportunity cost of holding a long position in crypto perpetual futures rises and the equilibrium funding rate increases accordingly, which means that part of the funding rate's *anomaly* may merely reflect changes in the macro interest-rate environment rather than a failure of market microstructure. However, the annualized funding rate of as high as 30% on the eve of the October 2025 crash far exceeded contemporaneous U.S. dollar interest rates, revealing the enormous leverage premium embedded within it. This extreme premium is itself a signal of behavioral crowding: the high concentration of directional positions (post-hoc statistics show that longs accounted for approximately 87% of liquidations) indicates that narrative-driven chasing and social-media-amplified optimism had pushed the market into an extremely fragile state. Far from suppressing this crowding, the high funding rate was instead read by some traders as confirmation that the *market is extremely bullish*, forming

a positive-feedback cognitive loop that runs counter to the negative-feedback design intent.

Figure 14-13 places the operating mechanisms of the positive-feedback loop and the negative-feedback loop side by side, revealing the symmetry between them in trigger conditions and transmission direction.

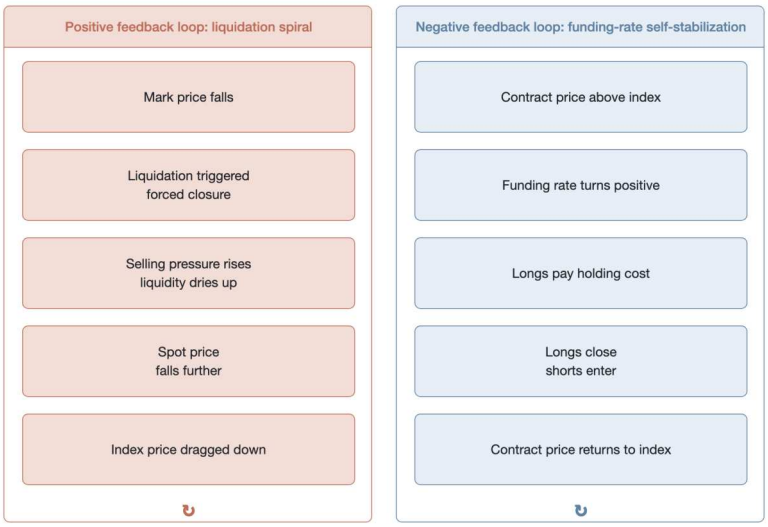


Figure 14-13. Comparison of the positive-feedback and negative-feedback loops

The core finding of this figure is that under normal market conditions, the negative-feedback channel of the funding rate dominates and can effectively pull the derivatives price back toward the spot index; however, when the benchmark price itself is distorted, the anchoring foundation of the negative-feedback loop is destroyed and the system rapidly switches to a positive-feedback mode dominated by the liquidation spiral. The switch between the two modes is often nonlinear, and the location of the critical point depends on the accuracy of the index and the depth of market liquidity.

14.6.3 Oracle synchronization risk

As the sole channel through which off-chain data enter the chain, the oracle's information-compression characteristic manifests not only within a single protocol but also, more significantly, at the cross-protocol systemic level. In the current on-chain financial ecosystem, oracle networks exhibit a highly concentrated character. The vast majority of mainstream protocols rely on a very small number of leading oracle service providers to obtain their price benchmarks. While this architecture lowers the integration cost of an individual protocol, it introduces a serious synchronization risk.

When multiple lending protocols, derivatives exchanges, and synthetic-asset platforms share the same oracle price feed, they effectively share the same liquidation trigger. Once the oracle updates a sharply lower price, all protocols subscribing to that feed

trigger their liquidation logic simultaneously. On a throughput-constrained blockchain such as Ethereum, this synchronized liquidation instantly generates a massive volume of liquidation transactions, causing severe network congestion and a surge in gas fees.

This synchronized congestion in turn obstructs users' margin top-ups and arbitrageurs' dampening trades. Borrowers cannot add margin before liquidation occurs, arbitrageurs cannot execute cross-market arbitrage in time, and liquidity providers cannot adjust their market-making ranges. The delay caused by network congestion is equivalent to injecting enormous friction into the entire system, so that a local imbalance that could have been regulated spontaneously by the market evolves into a systemic liquidity crisis. In the crash of May 19, 2021, network-wide liquidations exceeded $8 billion, and the synchronization of oracle updates caused severe on-chain congestion, so that many users could not stop their positions from being liquidated because they were unable to send transactions [4]. More insidious is the second-order effect of synchronized congestion: when network congestion prevents oracle price updates, the on-chain price stalls at a stale level, and the liquidations that should have been triggered are therefore delayed. Yet once the congestion eases and the oracle price completes a discrete jump update, all the positions awaiting liquidation that accumulated during the congestion are released in a concentrated burst within an extremely short time, and the resulting liquidation shock is far greater than the gradual liquidation that would have occurred had the oracle updated continuously. In other words, the recovery phase of synchronized congestion may be more destructive than the congestion itself.

14.6.4 The four transmission paths

The failure of a price benchmark is not an isolated event but spreads through the entire financial system along specific paths. Synthesizing the preceding analysis, we can outline four main transmission paths by which a price benchmark triggers a systemic crisis. These four paths describe the dynamic process by which a crisis spreads through the system, answering the question of *how a crisis evolves from a local anomaly into a systemic collapse*. Complementary to them is the taxonomy of five failure modes to be established in Section 14.7.2, which classifies the system's structural vulnerabilities from the static angle of architectural flaws, answering the question of *why the system is prone to failure at specific links*. A correspondence exists between the two frameworks: source contamination is both the starting point of a transmission path (this section) and an architecture-level failure mode (Section 14.7.2); the distortion of the arbitrage window is the joint runtime manifestation of the two failure modes of aggregation distortion and transmission latency. These four paths often do not occur in sequence but activate simultaneously and interweave under extreme conditions.

Source contamination is the starting point of crisis transmission, usually originating from a local liquidity depletion in the spot market or a system failure at an individual exchange; the anomalous price enters the calculation system of the benchmark price through the weight allocation of index construction and, in the absence of an effective exclusion mechanism, is amplified and passed to downstream. When the distorted benchmark price enters an anchoring mechanism such as the funding rate, the distortion of the arbitrage window follows: arbitrage capital, guided by erroneous economic signals,

withdraws from the market or takes offsetting positions, so that a mechanism that normally stabilizes turns into one that amplifies. At the same time, the liquidation spiral, through the smoothing algorithm of the mark price, converts a price shock into a stream of sustained liquidation instructions, and the intervention of the auto-deleveraging mechanism further contagiously spreads the risk to originally profitable hedging positions, causing systemic deleveraging. At the underlying-infrastructure level, the synchronized liquidation caused by multiple protocols sharing an oracle instantly exhausts the blockchain's computational resources, and the network congestion not only blocks users' self-rescue avenues but also paralyzes arbitrageurs' ability to intervene, so that the system loses all its regulatory functions at the very moment it most needs liquidity. These four paths do not occur in sequence but activate simultaneously and couple with one another under extreme conditions, ultimately amplifying a local price anomaly into an on-chain systemic crisis [42].

The interaction effects among the four paths are superadditive. Among them, the coupling of source contamination and feedback amplification is the most destructive path pair: a distorted source transmits through the index to the mark price, triggers erroneous liquidations, and the liquidation selling pressure in turn further distorts the spot price, constituting a new round of source contamination. When oracle synchronization risk is superimposed, multiple protocols sharing the same oracle feed trigger the above cycle at the same instant, amplifying a price distortion that should have been confined to a single protocol into a cross-protocol systemic contagion. The key path to breaking this cycle lies in the accuracy of the source layer: once the credibility of the index price is guaranteed, the downstream feedback amplification and arbitrage-window distortion lose their driving force.

Figure 14-14 visualizes the four paths above and their cross-coupling relationships in the form of a flowchart.

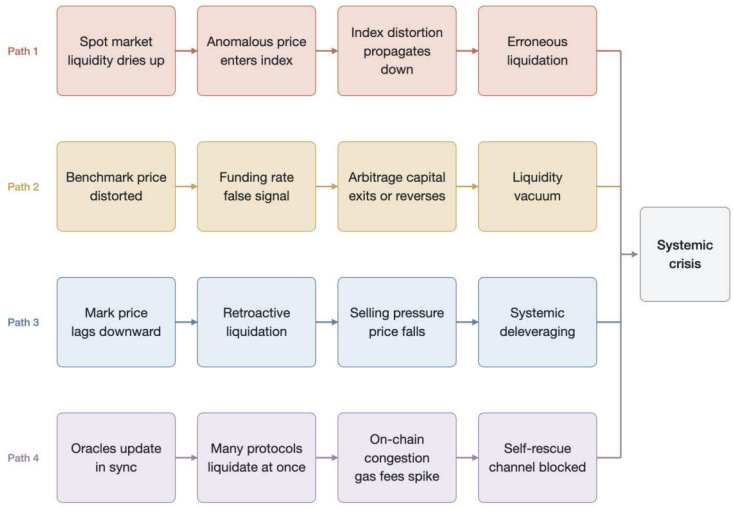


Figure 14-14. The four transmission paths of price-benchmark failure

Although the figure presents the four paths in parallel, their true danger lies in the fact that they are not isolated linear chains but couple with one another to form a mesh structure: the output of source contamination flows directly into the input of arbitrage-window distortion, the selling pressure produced by the liquidation spiral in turn aggravates the liquidity depletion of the source layer, and synchronized congestion acts as a lateral constraint that simultaneously limits the self-repair capability of all paths. This multi-path coupling explains why the scale of losses in extreme conditions often far exceeds what any single path would predict.

14.7 Attack surfaces, failure modes, and lines of defense

Having understood how a price benchmark transmits risk through the system, we must confront an objective fact: a price benchmark can not only fail naturally in an extreme market environment but can also become the target of deliberate manipulation by malicious attackers. The permissionless nature of on-chain finance and the openness and transparency of its code mean that any small design flaw can be turned into enormous economic gain. Attack-surface analysis, the classification of failure modes, and the construction of a defensive system constitute the practical core of price-benchmark governance.

14.7.1 A unified attack-surface analysis

The attack surface of a price benchmark is not a single-dimensional vulnerability but is distributed across the complete lifecycle from data collection to on-chain execution. We can deconstruct this lifecycle into three core layers: the index layer, the mark-price

layer, and the oracle layer. Any attack on a price benchmark is essentially a search for the weak link among these three layers, breaking their normal operating logic through capital advantage or technical means.

The attack surface of the index layer is concentrated mainly in liquidity manipulation in the spot market. An attacker artificially distorts the index's computed result through wash trading or malicious price ramping on a heavily weighted spot exchange. The success rate of this attack depends on the robustness of the index-construction rules, especially the ability to exclude outliers and the consideration of liquidity depth. If the index is merely a simple volume weighting without truncation of extreme prices, an attacker can influence the entire index with relatively little capital.

The attack surface of the mark-price layer focuses on the vulnerabilities of the smoothing algorithm. The mark price is designed to filter short-term volatility, but this smoothing mechanism can itself be exploited. In some designs, for example, if an attacker can maintain the spot price at an anomalous level for a sufficiently long time, the mark price will inevitably converge toward that anomalous price, thereby triggering the intended liquidations. This game along the time dimension depends on the contest between the attacker's capital cost and the length of the smoothing window.

The attack surface of the oracle layer concerns the security and timeliness of data transmission. Systematic research classifying oracle manipulation shows that attack vectors can unfold layer by layer, from the data source through off-chain aggregation to on-chain consumption [43]. Because the oracle is the main channel through which off-chain data enter the chain, it becomes the primary target of attack. Attack methods include manipulating a single oracle's data source through a flash loan, or exploiting network congestion to delay the oracle's update and thereby create arbitrage room. When the attack surfaces of these three layers are incorporated into a unified analytical framework, we can clearly see that the attacker's decision follows a simple condition of economic rationality: as long as the cost of manipulating the price benchmark is lower than the resulting liquidation gain or arbitrage profit, the attack will inevitably occur.

14.7.2 Five failure modes

To construct a defensive system systematically, we need to translate the attack surface into concrete failure modes. The failure of a price benchmark can be summarized into five typical modes covering every link from data input to governance intervention. Source contamination is the most basic failure mode, usually occurring in the upstream spot market: when an exchange serving as an index constituent suffers liquidity depletion or malicious manipulation, its anomalous output price, if not identified and excluded by the index-construction rules in time, enters the downstream system as a contamination source. Even if none of the data sources is itself contaminated, aggregation distortion can still arise at the index-calculation or oracle-node-aggregation stage, where unreasonable weight allocation or median-calculation logic can cause the final result to deviate severely from fair market value. The event in March 2026 in which a mainstream lending protocol, because of an oracle-configuration error, undervalued staked Ethereum and triggered $27 million in erroneous liquidations is a classic case of aggregation distortion [44]. The specific technical root cause of that event involved

the exchange-rate oracle between the liquid staking token (LST) and the underlying Ethereum, rather than a simple price-oracle error; in a scenario where the staking token traded at an exchange-rate discount, an erroneous exchange-rate parameter triggered liquidations based on an incorrect valuation. At the data-transmission stage, transmission latency constitutes another kind of failure: under extreme conditions, the oracle's update frequency may fail to keep pace with the violent price swings, and the resulting on-chain price lag creates a window of information asymmetry for high-frequency traders, allowing arbitrageurs to extract value from the protocol using stale on-chain prices. When a distorted price benchmark triggers unreasonable liquidations or erroneous funding-rate payments, the feedback-amplification effect further worsens the situation: the selling pressure produced by liquidation continues to distort the price, forming a vicious cycle that turns a local failure into a global systemic crisis. Beyond these technical failures, governance rigidity itself constitutes a failure mode. If the governance process is too cumbersome to make decisions within the critical window of a crisis, the system loses its last means of correction; and if governance power is too concentrated, so that a few people can arbitrarily modify the parameters of the price benchmark, it will provoke a crisis of market trust, causing participants to withdraw liquidity for fear that the rules will be tampered with.

14.7.3 Three lines of defense

Facing multi-dimensional attack surfaces and complex failure modes, a single defensive mechanism is clearly insufficient. An on-chain financial protocol must build a multi-layered system that ranges from passive response to active defense. This defense in depth can be divided into three lines: the prevention layer, the buffer layer, and the recovery layer.

The outermost layer of the defense-in-depth system is the prevention mechanism, whose core objective is to block an anomaly before it occurs—for example, by adopting dynamic weight allocation based on liquidity depth in place of static volume weighting, introducing a strict outlier-truncation mechanism to exclude quotes deviating from the median beyond a specific threshold, and using a multi-oracle redundant design to cross-validate data from different networks and reduce the risk of a single oracle being manipulated. When an anomalous price breaks through the prevention mechanism, the buffer mechanism takes on the function of limiting its destructive power: a circuit-breaker mechanism can automatically pause trading or liquidation during violent price swings, providing the market a cooling-off period; the mark price's EMA algorithm absorbs short-term price shocks through smoothing along the time dimension; and dynamically adjusting the leverage cap and raising the initial-margin requirement constrain the systemic risk exposure under extreme conditions. Where the system has already suffered actual losses, the recovery mechanism intervenes as a last resort: the insurance fund absorbs the bankruptcy losses that arise when liquidation is not timely, protecting the protocol from insolvency, and when the insurance fund is exhausted, the auto-deleveraging mechanism makes up the shortfall by forcibly closing profitable positions. Beyond these three layers of mechanisms, transparent governance intervention serves as a supplementary means, allowing emergency adjustment of system parameters or a state rollback to be carried out under extreme circumstances.

The effectiveness of the three lines of defense rests on the assumption that each layer independently bears part of the risk, but in a systemic-collapse scenario the layers may come under pressure simultaneously rather than being activated in sequence. When all sources deviate simultaneously under a systemic shock, the prevention layer's outlier filtering loses its reference benchmark for judging *what is normal*; the protection the buffer layer's EMA smoothing provides in a sustained one-directional move decays exponentially over time; and the emergency governance vote on which the recovery layer relies may be far too slow to keep pace with the market's collapse. As Reason's Swiss cheese model reveals, the safety of multi-layered defense depends on the probability that the flaws of each layer are not aligned at the same moment—and an extreme tail event is precisely the scenario in which all layers fail at once. It is therefore insufficient to evaluate the effectiveness of each defensive line in isolation; one must also identify the joint-failure threshold of the multi-layer defenses through reverse stress testing—that is, by fixing the worst-case liquidation outcome and working backward to the minimal market-shock scenario that would produce it.

14.7.4 The attack-defense matrix

Combining the five failure modes with the three lines of defense, we can construct a comprehensive attack-defense matrix. This matrix reveals not only the response strategy for each failure mode but also the weak links in the current defensive system. Against source contamination, the prevention mechanism relies mainly on liquidity monitoring of the spot market and filtering of anomalous quotes, the buffer mechanism reduces the influence of a single contamination source through the smoothing of the index price, but if the contamination persists, the recovery mechanism can only rely on the insurance fund to absorb the losses from erroneous liquidations. For aggregation distortion, the key to the prevention mechanism is adopting a more scientific aggregation algorithm (such as a volume-weighted median), the buffer mechanism can limit the magnitude of a single update by setting a change threshold on the aggregated result, and after a systemic error occurs it may be necessary to launch an emergency governance vote to correct the aggregation parameters manually. In responding to transmission latency, the prevention mechanism's means are to raise the oracle update frequency or adopt a pull-based architecture, the buffer mechanism can compensate for latency risk by widening the bid-ask spread or increasing trade slippage, and the recovery mechanism may need to restore balance by rolling back transactions or confiscating the profits of malicious arbitrageurs. Against feedback amplification, the prevention mechanism's focus is optimizing the mark price computation logic to avoid excessive lag in a one-directional move, the circuit breaker within the buffer mechanism can break the vicious cycle by forcibly pausing liquidation, and the recovery mechanism relies on the insurance fund and auto-deleveraging to digest the systemic risk. For governance rigidity, a tiered governance architecture separates routine parameter adjustments from core rule modifications and grants a security council the power for rapid intervention, a time-lock mechanism provides a buffer period for governance decisions to prevent malicious proposals from being executed instantly, and when governance fails completely, community consensus and forking mechanisms become the ultimate means of rebuilding trust.

Figure 14-15 summarizes, in matrix form, the cross-mapping between the five failure modes and the three lines of defense.

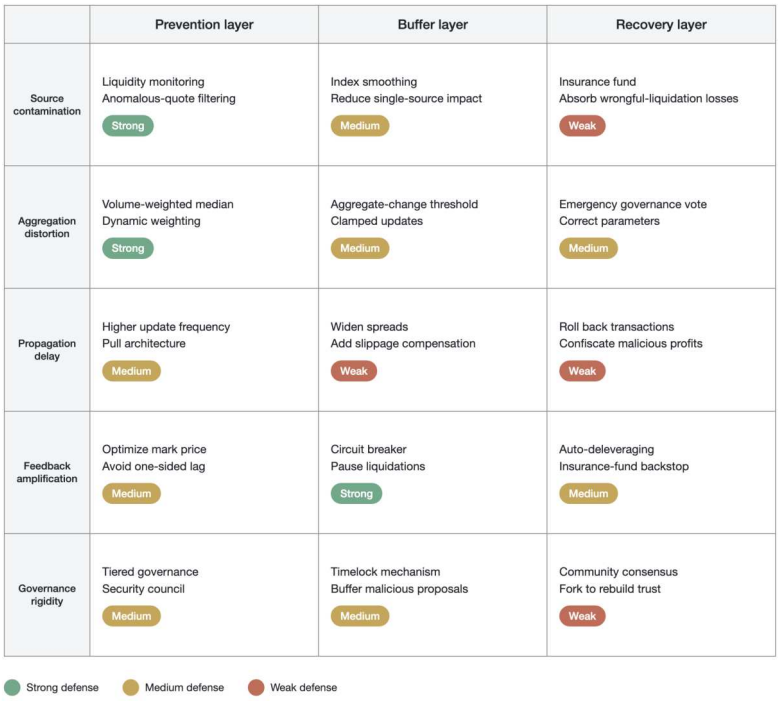


Figure 14-15. The attack-defense matrix of price-benchmark failure modes and the three lines of defense

The matrix shows that the prevention layer has strong coverage against source contamination and aggregation distortion but limited protective power against feedback amplification and governance rigidity. The circuit-breaker mechanism of the buffer layer plays a core role in responding to feedback amplification but is weaker in transmission-latency scenarios. The recovery layer, as a last resort, has an effectiveness that depends heavily on the adequacy of the insurance fund—and under extreme conditions, the insurance fund is precisely the resource most easily exhausted. This uneven coverage distribution indicates that the current defensive system has structural blind spots, especially in composite-attack scenarios where multiple failure modes are activated simultaneously.

14.7.5 Composite attacks and a case retrospective

A theoretical attack-defense matrix often appears structurally clear, but in the real on-chain world attackers frequently do not confine themselves to a single failure mode. The most dangerous and hardest to defend against is a composite attack that spans the index layer, the mark-price layer, and the oracle layer. In such an attack, multiple failure modes are activated simultaneously and the defenses are breached layer by layer.

The token-manipulation attack on a mainstream derivatives exchange in March 2025 is a classic case of a composite attack. The attacker first established a large short position

in a low-liquidity token on the derivatives market. Then, using their capital advantage, they sharply ramped up the token's price on the spot market. This behavior directly triggered source contamination at the index layer.

As the spot price surged, the oracle transmitted this distorted high price on-chain. Although the mark price algorithm tried to smooth this violent swing, under the attacker's sustained capital push the mark price was ultimately and irreversibly ramped up. This caused the liquidity pool that served as the protocol's insurance fund to incur an unrealized loss as large as $13.5 million. The attacker exploited the time lag of the mark-price layer and the mechanical execution of the oracle layer, converting spot-market manipulation into enormous profits in the derivatives market.

Facing this composite attack, the protocol's routine defenses proved unable to respond effectively. Had the system been allowed to run according to its established rules, the liquidity pool would have faced an insolvency risk of about $230 million (per Halborn; other estimates put it at about $290 million). In the end, the protocol team had no choice but to take an extreme governance intervention: forcibly delisting the token through a validator vote and forcibly settling all positions at a price favorable to the liquidity pool. Although this forced intervention averted the protocol's collapse, it also laid bare the system's fragility under an extreme composite attack, along with the residual risk of centralized intervention in price-benchmark governance. Viewed through the lens of operational risk management, the fundamental failure point of this attack was not the price-benchmark algorithm itself but the token-listing governance standards: a token with extremely thin spot liquidity (a market cap of only about tens of millions of dollars) was allowed to trade on a platform with enough depth to generate $13.5 million in risk exposure. Given that no mark price engineering can adequately protect an extremely illiquid asset from manipulation, asset-admission standards—particularly constraints based on minimum liquidity depth and a maximum ratio of risk exposure—constitute the first line of defense in price-benchmark governance. This episode shows that however sophisticated the technical design, the security of a price benchmark ultimately depends on dynamic risk management grounded in a thorough understanding of microstructure.

Figure 14-16 reconstructs the complete process of the composite attack in the form of a timeline, annotating the temporal relationship in which the index layer, the mark-price layer, and the oracle layer were separately activated at different stages.

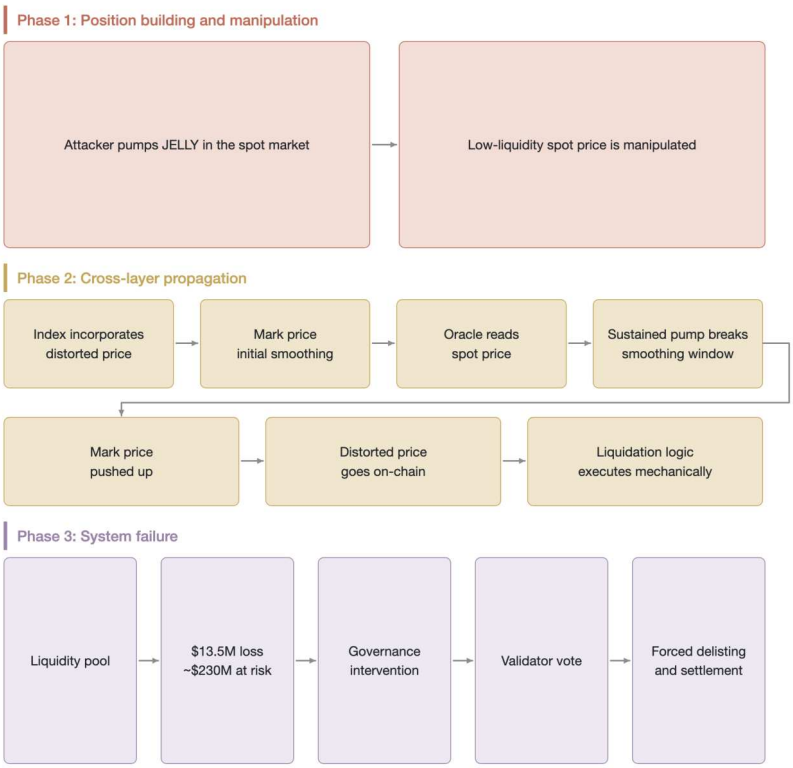


Figure 14-16. The timeline of a composite attack and the superposition of the three-layer attack surface

Note: The $13.5 million figure for the liquidity pool is the unrealized loss at the time of the event, corroborated by multiple reports; the $230 million figure is an estimate of the liquidity pool's potential risk from full exposure had the token continued to rise (per Halborn), not a realized loss.

The figure shows that within a window of about two hours the attacker exploited the vulnerabilities of the three layers in sequence: first contaminating the index-layer sources through spot-market manipulation (the position-building and price-ramping phase), then using the time lag of the smoothing algorithm to create a sustained deviation at the mark-price layer (the sustained-mark-price-deviation phase), and finally locking in the profit on-chain through the mechanical transmission of the oracle (the on-chain profit-locking phase). The superposition of the three attack surfaces made it impossible for any single-layer defensive measure to sever the complete transmission chain of the attack.

14.8 Case retrospective: the May 19, 2021 event

The effectiveness of price-benchmark governance must be tested under extreme market conditions. The cryptocurrency market crash of May 19, 2021 provides a highly complete study sample: a panic sell-off triggered by external news evolved, through a chain

reaction across the source layer, the index layer, the mark-price layer, and the oracle layer, into a market-wide liquidation cascade. This event was not the isolated failure of any single layer but a systemic validation of the four transmission paths analyzed in this chapter being activated simultaneously under extreme conditions.

14.8.1 Background and timeline reconstruction

In mid-May 2021, the cryptocurrency market was in the midst of a strong bull-market cycle; Bitcoin had touched an all-time high of about $64,000 in April of that year, and Ethereum set a record of about $4,360 on May 12 [45]. However, a series of negative news items were released densely within a short period, severely damaging market confidence. The National Internet Finance Association of China, the China Banking Association, and the Payment and Clearing Association of China jointly issued an announcement reiterating the prevention of the risks of speculative virtual-currency trading and prohibiting financial institutions and payment institutions from providing related services. A few days earlier, Elon Musk had expressed concern about the environmental impact of Bitcoin mining and announced the suspension of accepting Bitcoin payments [4]. These external shocks were rapidly amplified in the high-leverage derivatives market. The macro trigger of the May 19 event was a sustained fundamental-revaluation shock rather than a transient technical fault. The key difference between a shock of this nature and a flash crash is that the sustained one-directional selling pressure over several hours prevented the EMA smoothing mechanism from protecting traders by *waiting for a rebound*; the latency effect analyzed in Section 14.4.3 was, in this context, inverted from a protective mechanism into a catalyst for the backlog of delayed liquidations.

On May 19, market volatility peaked. Bitcoin's price plunged from about $43,000 to close to $30,000, a single-day drop of approximately 30%; Ethereum's fall was even more severe, with an intraday maximum drop of about 46%, breaking below $1,900 during the session [4] [46]. Driven by the collapse of spot prices, high-leverage positions in the derivatives market and in DeFi lending protocols faced enormous closing pressure, triggering a record wave of liquidations. By some counts, network-wide liquidations exceeded $8 billion within 24 hours [4]. On the centralized-exchange side, on May 19 alone the major platforms recorded more than $7 billion in long liquidations, of which Huobi accounted for approximately $2.95 billion [47]. In DeFi, the two major lending protocols Aave and Compound experienced approximately $310 million and $260 million in collateral liquidations, respectively, during May 17 to 23, of which the single-day liquidations on May 19 were approximately $170 million and $160 million, respectively—each an all-time high for its protocol [47].

Figure 14-17 compares liquidation amounts across major platforms in the May 19 event, showing the order-of-magnitude difference in liquidation scale between centralized exchanges and DeFi protocols.

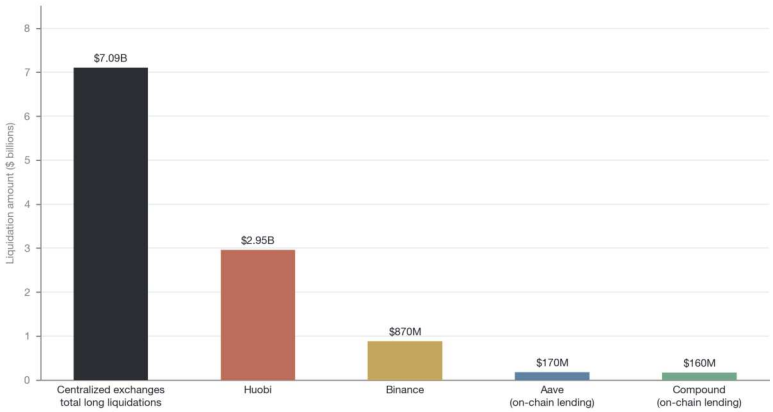


Figure 14-17. Comparison of liquidation amounts across platforms in the May 19, 2021 event

Note: The on-chain figures—approximately $170 million for Aave and approximately $160 million for Compound—are corroborated precisely by Gauntlet's liquidation retrospective; the centralized-exchange breakdown consists of third-party reconstructed values, including approximately $2.95 billion for Huobi and approximately $870 million for Binance, whereas Binance officially disclosed only approximately $24 million that day—a figure widely suspected of severe underreporting and therefore disputed.

This crash was not merely a violent price adjustment but a comprehensive stress test of the entire crypto financial infrastructure. From the matching engines of centralized exchanges to the price transmission of on-chain oracles, every link exposed design weaknesses under extreme load.

14.8.2 Mapping failure points across the layers

In the May 19 event, the transmission chain of the price benchmark showed clear failures at multiple layers, and these failures amplified market panic and intensified the liquidation cascade. At the source layer, extreme trading volume caused several mainstream centralized exchanges to experience outages or service degradation. Coinbase experienced serious connectivity issues from 5:50 a.m. to 7:38 a.m. Pacific Time, with its web servers largely unavailable because the Nginx routing connection count hit its ceiling, the error rate of the mobile-side GraphQL service climbing at one point to 10%, and the Coinbase Pro API partly paralyzed [48]. Binance likewise suffered serious technical failures, at one point suspending retail trading functions and halting the provision of some trading data [49]. Baumgartner and Güttler's study found that during the outage Binance's Bitcoin futures price diverged from those of other exchanges, with the spread widening to about seven times (relative to a 10-day reference period before the crash) [49]. These infrastructure failures caused the market to lose key price-discovery venues and the available price sources to plummet, directly damaging the quality of the price index's data inputs.

At the index layer, the absence and delay of price sources forced the index-construction

algorithm into degradation mode. When some exchanges' price data could not be obtained in time, the algorithm was forced to reallocate weights to the remaining, possibly poorly liquid, sources, causing the index's market representativeness to decline significantly. In some periods, some indices even had to use the last valid price as a substitute, effectively *freezing* the index update so that it no longer reflected the true market dynamics.

At the mark-price layer, the EMA smoothing algorithm produced a significant lag effect during the sustained one-directional fall. In the initial phase of the plunge, this lag actually provided a degree of protection, sparing some positions from erroneous liquidation caused by a momentary wick. However, when the mark price finally *caught up* to the true magnitude of the fall, the liquidations were not evenly distributed but erupted in a concentrated, pulse-like manner. Within a single time window, the liquidation volume far exceeded the market's liquidity-absorbing capacity, further intensifying the downward pressure on the price. This is precisely the classic manifestation of the mark price trap: in a one-sided move, the smoothing mechanism compresses liquidation pressure along the time dimension rather than eliminating it.

Figure 14-18 shows the price movements of Bitcoin and Ethereum on May 19 and the change in Ethereum network gas fees, presenting the synchronous relationship between the price plunge and network congestion.

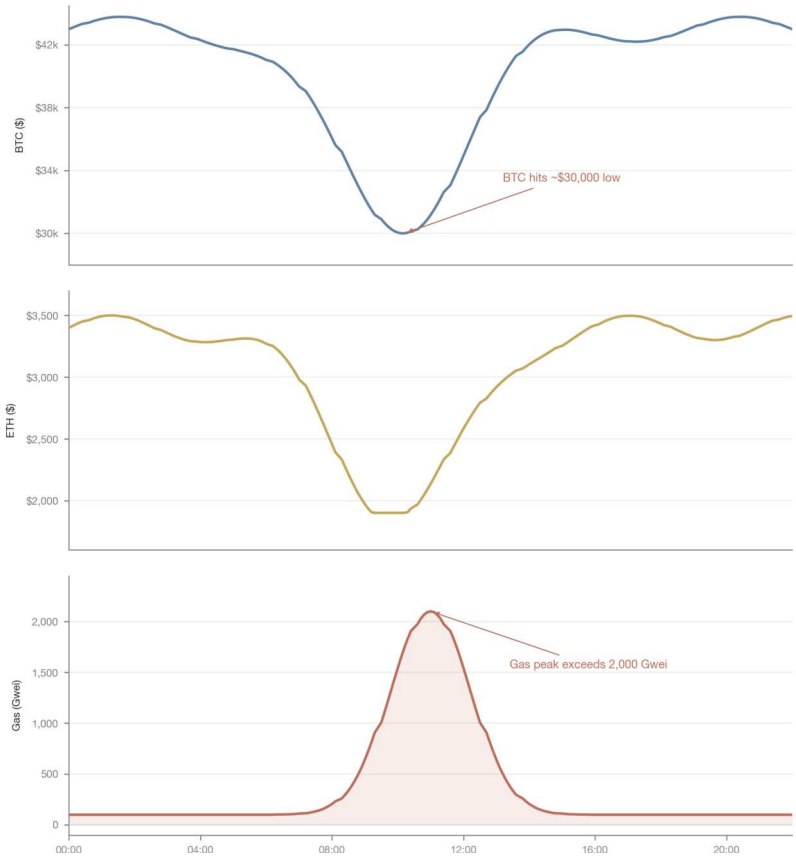


Figure 14-18. Bitcoin and Ethereum price movements and the change in Ethereum gas fees

Note: The Bitcoin and Ethereum price curves and the Ethereum gas-fee curve in the figure are illustrative reconstructions based on public records, not tick-by-tick measured series; the orders of magnitude—Bitcoin's intraday low of approximately $30,000, Ethereum declining from approximately $3,400, and gas briefly exceeding 1,000 Gwei that day—have been corroborated.

At the oracle layer, the extreme congestion of the Ethereum network became the critical bottleneck. Panicked investors, arbitrage bots, and liquidators competed to submit transactions, driving Ethereum's gas price at one point above 1,000 Gwei, with a peak exceeding 2,000 Gwei (the actual cost of complex transactions involving multiple contract calls was even higher), and some users reported that their transactions were still stuck after paying more than $1,000 in fees [50]. In such an extreme network environment, oracle nodes faced an economic constraint: submitting update transactions at a normal gas bid might leave them unpackaged for a long time, whereas frequent updates at a high gas bid would rapidly exhaust the operating budget. The price-update frequency of mainstream oracles such as Chainlink was therefore forced to decline, and on-chain prices lagged badly behind actual off-chain prices [51]. This lag not only prevented borrowers from adding margin in time but also exposed liquidators to enormous price-slippage risk when executing liquidations. At the same time, the high-gas environment entirely shut small and mid-sized traders out of on-chain operations: when the gas cost of a single swap ran to hundreds of dollars, a user holding a position worth a few thousand dollars effectively lost the ability to rescue themselves [50].

14.8.3 The simultaneous activation of the four paths

The fundamental reason the May 19 event evolved into a market-wide liquidation cascade is that the four transmission paths analyzed in this chapter were activated simultaneously under extreme conditions, forming a negative self-reinforcing loop.

Figure 14-19 shows the full-chain transmission path from external shock to liquidation cascade in the May 19 event, together with its feedback loop.

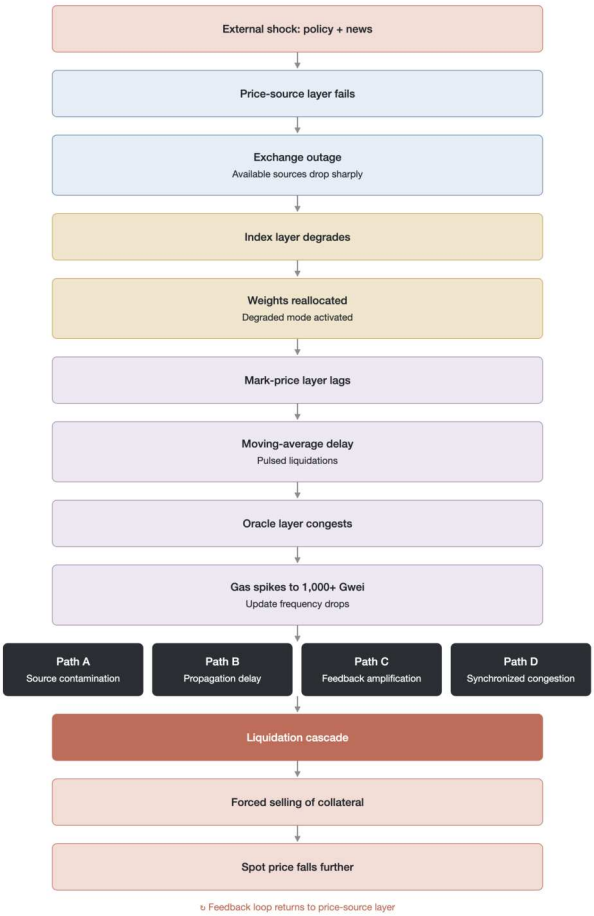


Figure 14-19. The full-chain transmission path and feedback loop of the May 19 event

At the source layer, when major exchanges such as Coinbase and Binance could not provide continuous trades and quotes because of extreme trading volume, the index-construction algorithm was forced to rely on the remaining, poorly liquid sources, giving extreme execution prices in a few markets a disproportionate weight and contaminating the overall index price. Under normal market conditions, the influence of these marginal sources could be diluted by the vast trading data of mainstream exchanges, but when the mainstream sources were absent at the same time, the index's manipulation resistance plummeted. At the same time, transmission latency struck at both the oracle layer and the mark-price layer: congestion on the Ethereum network delayed oracle quotes, while the mark price's smoothing algorithm added computational latency of its own, and this double latency decoupled the system states of DeFi protocols and derivatives exchanges from the real market, so that when the system finally executed liquidations based on the delayed price, it often forced positions closed at prices worse than the borrowers expected. At the liquidation-execution level, feedback amplification constituted the core driving force of the cascade: when high-leverage positions were liquidated, the liquidation engine forcibly dumped collateral assets on the market,

further depressing the spot price, and the fall in the spot price in turn transmitted back through the index and the oracle to the protocols, triggering the next round of liquidations of slightly-lower-leverage positions. In the May 19 event, this cycle of *fall–liquidation–dump–further fall* played out synchronously across centralized exchanges and DeFi protocols. Gauntlet's post-event analysis showed that the asset classes with the highest liquidated value in the Aave protocol were precisely those already flagged as high systemic risk in its risk model, indicating that the activation of feedback amplification is predictable [47]. At the infrastructure level, the synchronized liquidation caused by multiple protocols sharing an oracle instantly exhausted Ethereum's computational resources; borrowers could not send transactions to repay debt or add margin, oracles could not update prices in time, and the entire on-chain financial system fell into synchronized paralysis. The surge in gas fees was essentially an implicit admission barrier that shut resource-limited participants out of self-rescue actions, intensifying the irreversibility of the liquidations.

Table 14-3 summarizes the specific manifestations of the four transmission paths in the May 19 event and their failure points at each layer.

Transmission path	Failure layer	Specific manifestation	Consequence
Path A: Source contamination	Source layer, index layer	Exchange outages sharply reduced available sources; index weights tilted toward low-liquidity sources	Index representativeness declined; price deviations widened
Path B: Transmission latency	Oracle layer, mark-price layer	Surging gas fees delayed oracle updates, compounded by moving-average smoothing lag	On-chain prices decoupled from the real market; liquidation timing was misjudged
Path C: Feedback amplification	Entire chain	Liquidation selling depressed spot prices, triggering further liquidations	Cascade liquidations far exceeded the initial shock
Path D: Synchronized congestion	Oracle layer, infrastructure layer	Multiple protocols triggered liquidations simultaneously, congesting the entire Ethereum network	Small and mid-sized users lost the ability to self-rescue; liquidations became irreversible

Table 14-3. The manifestations and consequences of the four transmission paths in the May 19 event (Data source: compiled by the author from public materials on the May 19 event)

Comparing the actual behavior of the May 19 event with the predictions of this chapter's theoretical framework allows us to test the framework's explanatory power and its limits. Based on the EMA recursive formula, Section 14.4.3 predicts that in a sustained one-sided fall the lag of the mark price will cause liquidation events to be released in a concentrated burst along the time dimension; in the May 19 event, the mark price did indeed lag the last traded price in the early phase of the plunge, and the subsequent pulse-like concentrated liquidations validated this prediction. Section 14.6.3's analysis of oracle synchronization risk predicts that multiple protocols sharing the same ora-

cle feed would cause synchronized liquidation and network congestion under extreme conditions; the facts that gas fees surged above 1,000 Gwei and that multiple protocols triggered liquidations at the same time validated this prediction precisely. One dimension the chapter's framework failed to anticipate fully, however, is the synchronized failure of exchange infrastructure: the theoretical framework assumes that source failures are random and independent, but the simultaneous outages of major sources such as Coinbase and Binance in the May 19 event show that extreme trading volume itself constitutes a systemic common-cause failure, and this infrastructure-level correlation should be incorporated into future analytical frameworks. The overwhelming preponderance of long liquidations in the May 19 event calls for explanation. In the weeks before the crash, social-media amplification of the bull-market narrative and the overconfidence brought by successive all-time highs drove retail traders to go one-directionally long with extremely high leverage. This extreme directional concentration is itself a systemic-risk indicator that can be monitored in advance, but the market-governance framework of the time lacked an automated mechanism for dynamically adjusting margin or leverage limits based on such signals.

The simultaneous activation of the four paths reveals a key structural problem: the layers of a price benchmark do not operate independently but affect one another through complex coupling relationships. Under normal market conditions, this coupling is benign, and the multi-layer redundancy provides an extra margin of safety. But under extreme conditions, the coupling amplifies a local failure into a systemic collapse, and the defensive mechanisms of the layers not only fail to cover for one another but drag one another down because they come under pressure simultaneously.

14.8.4 Case comparison: systemic crash versus targeted attack

Comparing the May 19 event with the JELLY token attack of March 2025 analyzed in Section 14.7.5 reveals two entirely different modes of price-benchmark failure. The May 19 event is a systemic crash, its trigger an exogenous macro shock, with the four transmission paths activated simultaneously and spontaneously, absent any single attacker's subjective intent. The JELLY attack is targeted manipulation, in which the attacker, through a precise understanding of the mark price computation logic, actively and methodically exploited the vulnerabilities of the three layers. On the time scale, the liquidation cascade of the May 19 event lasted about 24 hours and involved hundreds of assets across the market; the JELLY attack was completed within about two hours and targeted only a single low-liquidity token. In terms of the mode of defensive failure, what the May 19 event exposed was insufficient infrastructure redundancy and the economic constraint on oracles in a high-gas environment, and the failure of the defenses was concurrent; what the JELLY attack exposed was the flaw in the mark price mechanism for low-liquidity assets, and the failure of the defenses was serial. This comparison shows that price-benchmark governance must not only cope with systemic tail risk but also guard against precise attacks that target specific parameter vulnerabilities—and the defensive strategies the two kinds of threat require differ fundamentally.

14.8.5 Post-event improvements and governance reflection

The May 19 event exposed the fragility of early crypto financial infrastructure but also spurred profound reflection and improvement across the industry in price-benchmark governance and system design. At the exchange level, centralized platforms substantially strengthened their systems' stress resilience and the redundancy of their price benchmarks. In its post-event analysis, Coinbase acknowledged that its auto-scaling rules had failed to adapt to the burst-traffic patterns unique to crypto markets and pledged to optimize its load-testing scheme specifically to simulate genuine extreme scenarios [48]. Several exchanges increased the number of price sources for their price indices, lowered the weight cap of any single source, and introduced liquidation cooling-off-period mechanisms during extreme volatility. Some platforms also optimized the moving-average parameters of their mark prices so that, while retaining the smoothing effect, they could respond more quickly to sustained one-sided moves.

At the oracle level, the industry accelerated its evolution toward more efficient architectures. Chainlink had already deployed its off-chain reporting protocol to mainnet in early 2021; by aggregating node signatures off-chain and submitting only a single transaction on-chain, this protocol reduced the gas cost of oracle updates by nearly 90% [52]. The May 19 event further validated the necessity of this upgrade and spurred continued optimization in subsequent versions. At the same time, a new generation of pull-based oracles such as Pyth Network began to gain adoption; their design philosophy shifted from the traditional *push* model to an *on-demand pull* model, allowing decentralized applications to actively fetch the latest price data when needed, thereby avoiding the update-latency problem of the push model during network congestion [53].

At the DeFi protocol level, mainstream lending platforms such as Aave and Compound introduced finer-grained risk-management mechanisms in their subsequent versions. A multi-oracle aggregation strategy became an industry standard: protocols no longer rely on a single oracle as their price source but simultaneously connect to multiple independent oracle networks, taking the median or weighted average as the reference price to reduce the risk of a single point of failure. Progressive liquidation mechanisms were also widely adopted, splitting a large liquidation into multiple small batches for execution to reduce the instantaneous impact on the spot market [47].

Viewed from a governance perspective, the core lesson of the May 19 event is that seemingly purely *technical parameters*—such as the length of the moving-average window, the update frequency of the oracle, and the degradation strategy of the index—directly determined the liquidation fate of billions of dollars under extreme conditions. The post-event improvements were not only technical optimizations but also an enhancement of governance transparency. A growing number of exchanges and protocols began to publish their index-construction methodologies and parameter-change logs, transforming the design of price benchmarks from a closed technical decision into an auditable governance action.

The far-reaching impact of the May 19 event was reflected not only in improvements to technical architecture but also, more directly, in an acceleration of regulatory action. In September 2021, the People's Bank of China (PBOC) and ten departments jointly issued a notice characterizing all virtual-currency-related business activities as illegal

financial activities, and over the same period a complete phase-out of cryptocurrency mining was implemented. At the broader global level, the event was cited by regulators in several jurisdictions as evidence of systemic risk in crypto markets, spurring the accelerated formation over subsequent years of regulatory frameworks targeting crypto-asset service providers. This trajectory shows that the ultimate binding force on price-benchmark governance may come not from the self-improvement of technical design but from external regulatory intervention triggered by systemic events.

14.9 Chapter summary

Having explored the construction, transmission, failure, and governance of price benchmarks, we must examine the essential nature of this infrastructure from a higher vantage point. In the context of crypto finance, a price benchmark is not merely a stream of objective data; it substantively wields broad financial power through technical and algorithmic means: defining what a *fair price* is, deciding who gets liquidated, shaping the distribution of liquidity, and setting the difficulty threshold for manipulation.

To constrain this implicit financial power and ensure the objectivity, fairness, and robustness of price benchmarks, traditional finance has accumulated rich experience. IOSCO's Principles for Financial Benchmarks, issued in 2013 after the LIBOR manipulation scandal, provided a systematic framework for the governance of global financial benchmarks [16]. Combining the characteristics of crypto markets with the analysis in this chapter, we can distill six core design principles for price-benchmark governance. Compilation independence requires that the benchmark's administrator be independent, in organizational structure and interests, from data providers and benchmark users, and that strict conflict-of-interest firewalls be established. As analyzed in Section 14.2, a crypto exchange simultaneously plays the *three-in-one* roles of index compiler, mark price setter, and liquidation executor, and its conflicts of interest exceed even those of the bank-quote system of the LIBOR era; breaking this structure is the core institutional need of crypto-market price-benchmark governance. Methodological transparency requires that the index's construction rules, weight-allocation logic, outlier-exclusion criteria, and degradation trigger conditions be fully public and subject to community and market scrutiny; parameter changes should come with advance notice and a complete change log, so that market participants can assess the potential impact of rule changes on their own positions. The principle of redundancy over optimization emphasizes that, while pursuing update frequency and precision, one must ensure multiple redundancy of price sources, computation nodes, and transmission channels; but the effectiveness of redundancy is premised on the independence of the components: if multiple price sources share the same underlying infrastructure (such as the same cloud provider or the same data-aggregation interface), they may fail simultaneously in a systemic event—that is, common-mode failure. The simultaneous outages of Coinbase and Binance in the May 19 event are a classic case of common-mode failure (see Section 14.8.2), showing that the value of redundancy depends on the diversity of components' failure modes, not merely on their number. Degradation over collapse requires that when some components fail, the system be able to transition smoothly to a preset degraded mode (such as lowering the update frequency or widen-

ing the spread tolerance) rather than simply stopping service or outputting erroneous data; this principle corresponds directly to the stress-degradation mechanism analyzed in Section 14.3.6, and degradation paths should be predefined and regularly validated through stress testing. Feedback awareness requires the benchmark's designers to fully understand the endogenous feedback loops between the price benchmark and the underlying market. As analyzed in Section 14.6, the moving-average parameters of the mark price affect the intensity of the liquidation cascade through a reflexivity equation, and this influence must be explicitly incorporated into design considerations—the liquidation spiral is not an unforeseeable *black swan* but an inevitable result of a specific parameter combination. Governance evolvability acknowledges that as market structure and underlying technology change, the methodology of a price benchmark cannot remain fixed; regular parameter reviews, community-participatory modification mechanisms, and forward-looking assessment of emerging risks together constitute the institutional foundation of governance evolution.

A price benchmark is regarded as a form of financial power because it directly defines the *fair price* in the eyes of a smart contract. This price determines whether a borrower is liquidated, sets the funding rate a derivatives trader bears, and shapes the liquidity distribution of cross-market arbitrageurs. Throughout this process, the benchmark's administrator constantly faces a series of difficult trade-offs.

Table 14-4 summarizes the trade-off spectrum of the six core dimensions of price-benchmark governance.

Governance dimension	Conservative end	Aggressive end	Core tension
Degree of smoothing	Long-window moving average, high robustness	Short window or no smoothing, high timeliness	Positioning within the mark price trilemma
Number of price sources	Minimal set, low complexity	Maximal set, high redundancy	Trade-off between operational cost and manipulation resistance
Degradation threshold	Low threshold, conservative suspension	High threshold, maximal continuity	Tension between user protection and market continuity
Liquidation speed	Immediate liquidation, protocol safety first	Delayed liquidation, user protection first	Allocation between protocol interests and user interests
Degree of transparency	Fully public, auditable	Partially confidential, gaming-resistant	Balance between trust-building and security defense
Oracle model	Push-based, always available	Pull-based, on-demand low latency	Trade-off between gas cost and timeliness

Table 14-4. The six-dimensional trade-off spectrum of price-benchmark governance (Data source: constructed by the author)

These trade-offs have no absolutely correct answer; they are essentially a redistribution of different types of risk. When a protocol chooses a longer-window smoothing algorithm, it is in effect transferring the risk of short-term price volatility from borrowers

to liquidity providers and the insurance fund: borrowers gain more time to add margin, but the protocol faces greater bankruptcy risk in an extreme one-sided move. When an oracle sets a specific deviation threshold, it is making a trade-off between guarding against minor manipulation and saving on network fees: a higher threshold lowers the update frequency and gas consumption but also widens the potential window of deviation between the on-chain price and the real market.

The core insights of this chapter can be condensed into four propositions. Price benchmarks equal implicit governance: designing a price benchmark is not a neutral technical activity but an exercise of power that defines *price reality*, and the lesson of LIBOR is replaying itself in a new form. The mark price trilemma shows that accuracy, robustness, and timeliness can be optimized only two at a time, and the choice of the moving-average window is not merely a technical parameter but also determines *who is protected and who is sacrificed*. Oracles constitute a cross-layer information bottleneck: the entire price reality of on-chain finance is obtained through the single conduit of the oracle, and the information compression along the three dimensions of time, precision, and source is irreversible. Attack-surface analysis reveals that the index, mark price, and oracle each have a quantifiable attack surface, the most dangerous being a multi-layer composite attack, and defending each layer carries a clear cost within the trilemma. The existence of flash loans fundamentally challenges the basic assumption of traditional market-microstructure theory that capital constraints serve as a deterrent to manipulation. In traditional finance, the cost of price manipulation under the Kyle (1985) [54] framework is proportional to the amount of capital an attacker must deploy, and this capital requirement constitutes a natural defensive barrier. Flash loans, however, allow an attacker to mobilize hundreds of millions of dollars of liquidity without collateral within a single transaction, effectively removing the capital constraint from the manipulation-cost function. This means that in the cost-benefit framework of manipulation analyzed in Section 14.5.8, the traditional cost term of *attack capital times slippage* must be redefined: in an on-chain environment that supports flash loans, the true constraint on manipulation is no longer the scale of capital but the depth of the target liquidity pool and the length of the oracle's time window.

Despite multiple rounds of iteration, price-benchmark governance in crypto financial markets still faces many unresolved challenges, and as the industry develops, some new technologies and paradigms are reshaping the boundaries of this field. The dilemma of decentralized benchmark construction is the most fundamental among them. Although the transmission layer of oracles has achieved a high degree of decentralization, the construction rules of the benchmark index are often still controlled by a small number of core teams or stakeholders. Can a cross-protocol, DAO-governed unified price benchmark be established? How can compilation independence be achieved in a decentralized environment? The answers to these questions will directly affect the institutional maturity of crypto financial markets. The timing and manner of regulatory intervention likewise remain undecided: is it necessary to establish a crypto version of a regulatory framework akin to IOSCO's Principles for Financial Benchmarks? Chapter 26 discusses this topic in depth.

At the technological frontier, the application of zero-knowledge proofs in the oracle domain shows enormous potential [55]. By computing the price index off-chain and

generating a zero-knowledge proof, an oracle can move complex computation logic off the main chain without sacrificing security, greatly lowering verification costs while guaranteeing the credibility of the computed result. The maturation of artificial intelligence and machine learning also offers new possibilities for the outlier detection and dynamic weight allocation of price benchmarks, but whether adaptive parameters will introduce new opacity still warrants careful evaluation; Chapter 31 explores this. In an era of coexisting multiple chains, how to build a secure, low-latency cross-chain price benchmark to support the integration of omnichain liquidity is another pressing engineering challenge; Chapters 28 to 30 develop the design from an engineering perspective.

Chapter 15 will use empirical data to test the theoretical framework established in this chapter and in Chapter 13, measuring the informational efficiency of crypto markets. Is the price-discovery leadership of perpetual futures empirically supported? How large is the magnitude of mechanical noise? How does efficiency vary across different market states? The price-manipulation attack surfaces analyzed in this chapter will be tested in Chapter 15 for whether they are detectable in the market through data.

References

[1] Ali, Z. (2025). Anatomy of the Oct 10-11, 2025 crypto liquidation cascade: Macroeconomic triggers, market microstructure, and systemic risk lessons. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5611392 https://doi.org/10.2139/ssrn.5611392

[2] HM Treasury. (2012). *The Wheatley Review of LIBOR: Final Report*.

[3] Solidus Labs. (2022). *The Mango Markets Exploit: An Order Book Analysis*.

[4] Deribit Insights. (2021). The (Flash) Crash of May 19, 2021. https://insights.deribit.com/exchange-updates/the-flash-crash-of-may-19-2021/

[5] Harris, L. (2003). *Trading and Exchanges: Market Microstructure for Practitioners*. Oxford University Press.

[6] Chainalysis. (2023). *Oracle Manipulation Attacks Rising: A Unique Concern for DeFi*.

[7] Hyperliquid. (2025). *Robust price indices*. Hyperliquid GitBook. https://hyperliquid.gitbook.io/hyperliquid-docs/trading/robust-price-indices

[8] Binance. (2025). *What Are Mark Price and Price Index in USD(S)-Margined Futures*. Binance Support.

[9] Alexander, C., Choi, D. F., Massie, H., & Sohn, J. (2020). Price discovery in Bitcoin: The impact of unregulated markets. *Journal of Financial Stability*, 50, 100776. https://doi.org/10.2139/ssrn.3583843

[10] BitMEX. (2025). *Fair price marking*. https://www.bitmex.com/app/fairPriceMarking

[11] Bybit. (2026). *Upcoming Improvement to Index Price Calculation*. Bybit Announcements.

[12] Shearer, M., Rauterberg, G. V., & Wellman, M. P. (2023). Learning to manipulate a financial benchmark. In *Proceedings of the Fourth ACM International Conference on AI in Finance* (ICAIF '23). Association for Computing Machinery. https://doi.org/10.1145/3604237.3626847

[13] Williamson, O. E. (1985). *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. Free Press.

[14] North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press. https://doi.org/10.1017/cbo9780511808678

[15] Duffie, D., & Stein, J. C. (2015). Reforming LIBOR and other financial market benchmarks. *Journal of Economic Perspectives*, 29(2), 191-212. https://doi.org/10.1257/jep.29.2.191

[16] International Organization of Securities Commissions (IOSCO). (2013). *Principles for Financial Benchmarks: Final Report*.

[17] Bitwise Asset Management. (2019). *Presentation to the U.S. Securities and Exchange Commission: Fake Volume on Cryptocurrency Exchanges*.

[18] European Union. (2023). *Regulation (EU) 2023/1114 on Markets in Crypto-Assets (MiCA)*, Article 72 (Conflicts of interest of crypto-asset service providers).

[19] International Organization of Securities Commissions (IOSCO). (2023). *Policy Recommendations for Crypto and Digital Asset Markets: Final Report* (IOSCOPD747).

[20] CoinDesk Indices. (2023). *CoinDesk Indices Update on Binance.US*.

[21] United States District Court, Southern District of New York. (2025). *United States v. Avraham Eisenberg, No. 1:23-cr-00010: Opinion and Order* (granting Rule 29 motion for judgment of acquittal, May 23, 2025).

[22] OKX. (2026). *Index components weightage adjustment FAQ*. OKX Help Center.

[23] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 135(2), 293-319. https://doi.org/10.1016/j.jfineco.2019.07.001

[24] Binance. (2025). *Binance futures will update the mark price calculation for all USD(S)-M and COIN-M perpetual contracts*. https://www.binance.com/en/support/announcement/detail/20c5bfbd9e084b6982c768516c316514

[25] OKX. (2025). *Mark price calculation*. https://www.okx.com/help/mark-price

[26] dYdX. (2025). *dYdX v4 documentation: Oracles*. https://docs.dydx.exchange/

[27] Brave New Coin. (2019). The Bitstamp flash crash: Why robust indices matter. *Brave New Coin*. https://bravenewcoin.com/insights/the-bitstamp-flash-crash-why-robust-indices-matter

[28] CoinDesk. (2025). HyperLiquid delists JELLY after vault squeezed in $13M tussle. *CoinDesk*. https://www.coindesk.com/markets/2025/03/26/hyperliquid-delists-jellyjelly-after-vault-squeezed-in-usd13m-tussle

[29] Messari. (2025). *Chainlink: A Full-Stack Institutional Platform*.

[30] Immunefi. (2025). *The Ecosystem Vulnerability Scoreboard: Six Years of DeFi*

Loss Data. Immunefi Research.

[31] Cong, L. W., Fox, L., Li, S., & Zhou, L. (2025). A primer on oracle economics. *Journal of Corporate Finance*. https://doi.org/10.2139/ssrn.5208421

[32] Chainlink Labs. (2021). *Chainlink 2.0: Next Steps in the Evolution of Decentralized Oracle Networks*.

[33] Pyth Network. (2022). *Pull, Don't Push: A New Price Oracle Architecture*.

[34] dYdX Foundation. (2023). *dYdX v4 Architecture Overview*.

[35] Rekt News. (2020). Harvest Finance – Rekt. https://rekt.news/harvest-finance-rekt/

[36] Humphry, R., Avramovic, P., Acevedo, F., & Alkhair, S. (2024). *Review of maximal extractable value & blockchain oracles*. Financial Conduct Authority.

[37] Benligiray, B., & Vänttinen, H. (2022). *Oracle extractable value (OEV) through order flow auctions* [Litepaper]. API3. https://github.com/api3dao/oev-litepaper

[38] Aspembitova, A. T., & Bentley, M. A. (2023). Oracles in decentralized finance: Attack costs, profits, and mitigation measures. *Entropy*, *25*(1), Article 60. https://doi.org/10.3390/e25010060

[39] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market Liquidity and Funding Liquidity. *The Review of Financial Studies*, 22(6), 2201-2238.

[40] Gromb, D., & Vayanos, D. (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics*, 66(2-3), 361-407. https://doi.org/10.1016/S0304-405X(02)00228-3

[41] FTI Consulting. (2025). Crypto Crash Oct 2025: Leverage Meets Liquidity. https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity

[42] Kaiko Research. (2025). Crypto's Pricing Problem Laid Bare. https://research.kaiko.com/insights/cryptos-pricing-problem-laid-bare

[43] Eskandari, S., Moosavi, S. M., & Clark, J. (2021). SoK: Oracles from the ground truth to market manipulation. In *Proceedings of the 3rd ACM Conference on Advances in Financial Technologies* (AFT '21). https://doi.org/10.1145/3479722.3480994

[44] CoinDesk. (2026). DeFi lending platform Aave sees a rare $27 million liquidations after a price glitch. https://www.coindesk.com/business/2026/03/10/defi-lending-platform-aave-sees-a-rare-usd27-million-liquidations-after-a-price-glitch

[45] Chainalysis. (2021). The May 2021 cryptocurrency crash explained: Selling by new investors. https://www.chainalysis.com/blog/cryptocurrency-price-crash-may-2021/

[46] CNBC. (2021, May 19). The crypto collapse: Here's what's behind bitcoin's sudden drop. https://www.cnbc.com/2021/05/19/the-crypto-collapse-heres-whats-behind-bitcoins-sudden-drop.html

[47] Gauntlet. (2021). Aave protocol liquidation retrospective: May 2021. https://www.gauntlet.xyz/resources/aave-protocol-liquidation-retrospective-may-2021

[48] Coinbase. (2021). Incident post mortem: May 19, 2021. https://www.coinbase.c

om/blog/incident-post-mortem-may-19-2021

[49] Baumgartner, T., & Güttler, A. (2022). Bitcoin flash crash on May 19, 2021: What did really happen on Binance? (IWH Discussion Papers No. 25/2022). Halle Institute for Economic Research. https://www.econstor.eu/handle/10419/266345

[50] Glassnode. (2021). DeFi uncovered: Navigating the crash. https://insights.glassnode.com/defi-uncovered-exploring-the-crash/

[51] Pandey, A. (2025). The oracle problem in DeFi: Chainlink and price feed architecture. Medium. https://medium.com/@arunpandeycdac/the-oracle-problem-in-defi-chainlink-and-price-feed-architecture-29cf1dd31342

[52] Chainlink. (2021). The year in Chainlink 2021: 7 pillars of momentum. https://blog.chain.link/the-year-in-chainlink-2021/

[53] Pyth Network. (2024). Pyth pull oracle launches on Solana. https://www.pyth.network/blog/pyth-network-pull-oracle-on-solana

[54] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315-1335. https://doi.org/10.2307/1913210

[55] Gu, B., & Nawab, F. (2023). zk-Oracle: Trusted off-chain compute and storage for decentralized applications. In *Proceedings of the 2023 IEEE International Conference on Blockchain and Cryptocurrency* (ICBC). IEEE. https://doi.org/10.1109/ICBC56567.2023.10174864

Informational Efficiency in Digital Asset Markets

On January 10, 2024, when news broke that the SEC had approved the first spot Bitcoin exchange-traded funds (ETFs), the price of Bitcoin swung violently within a single minute, as traders and algorithms around the world digested this historic decision at unprecedented speed. Yet after the initial spike, the market did not embark on the sustained rally that many had expected. Instead, within less than three days the price fell rapidly from a high near $49,000 to below $42,000, a textbook "buy the rumor, sell the news" move. During the same window in which the spot price gyrated, the perpetual futures market revealed a richer informational picture: the basis on BTC perpetual futures flipped from positive to negative within minutes; the funding rate ran through a full cycle from an extreme positive value back toward its mean over the settlement periods that followed; and open interest fell markedly within 48 hours. These dimensions, unique to perpetual futures, offer a microstructure perspective on informational efficiency that the spot market cannot provide. The episode crystallizes the complexity of informational efficiency in digital asset markets and raises a series of core questions. Why did the market react so quickly to a clear piece of regulatory approval, only to be dominated by behavioral biases over the following days? When a DeFi protocol is hacked, the attack is recorded publicly on the blockchain, so why do most investors learn what happened only hours later? Under the extreme stress of the FTX collapse, why did a price discovery mechanism that normally looked highly efficient come close to paralysis? Under the same shock, which reacts faster—perpetual futures or spot? How do design choices (funding rates, mark prices, liquidation mechanisms) systematically shape these differences? And do the funding rate and the basis themselves contain predictable information, constituting a structural gap in weak-form efficiency? The answers to all of these questions point to the same central issue: to what extent is the market "efficient"?

This question is not new. It runs through half a century of research in finance. Ever since Fama (1970) proposed the efficient market hypothesis, scholars and practitioners have continued to debate it. In the context of digital assets, however, this classic question takes on new analytical dimensions. The previous two chapters established the key foundations for understanding how markets operate. Chapter 13 dissected the micro-level process of price discovery, showing how information is gradually incorporated into prices through the interaction of order flow, market makers, and arbitrageurs. Chapter 14 examined the governance of price benchmarks, showing how the index price and the

mark price provide an anchoring reference for the market. Those two chapters answered the questions of "how prices form" and "how prices are measured." This chapter builds on them to ask a more fundamental question: to what extent do these price-formation and price-measurement mechanisms actually reflect information efficiently? In other words, we move from the microstructural "how it is done" to the macro-level "how well it is done."

This chapter takes perpetual futures as its primary data source and institutional sample, organizing the analysis around weak-form efficiency, semi-strong-form efficiency, institutional efficiency, and extreme-stress testing. The analysis begins from the classic efficient market hypothesis and its limitations in crypto markets, introduces the adaptive markets hypothesis, and constructs an "efficiency spectrum" framework built on five dimensions: time, asset, market tier, information type, and market state. On this foundation, the chapter uses statistical tools such as variance ratio tests and autocorrelation analysis to test weak-form efficiency; applies the event study method to compare how fast the market reacts to information with different processing costs; analyzes how the institutional design of perpetual futures creates predictable efficiency cycles; and, through the case of the FTX collapse, shows how market efficiency breaks down nonlinearly under extreme stress. Finally, the chapter takes a historical-evolutionary view of how crypto market efficiency has progressed from the frontier era to the regulated era, and summarizes the core findings and their practical implications.

15.1 The analytical toolkit

To judge whether a market is efficient, one first needs a precise set of conceptual tools and measurement standards. This section builds that analytical toolkit and lays the theoretical foundation for the empirical tests that follow. We begin from the efficient market hypothesis of Fama (1970), reviewing its classic three-part taxonomy and the structural limitations it exhibits in crypto markets; we then introduce the adaptive markets hypothesis of Lo (2004), which uses an evolutionary and behavioral lens to remedy the shortcomings of the static theory; and on that basis we propose this chapter's core analytical framework, the five-dimensional efficiency spectrum matrix, which turns efficiency from a binary label into a spectrum that can be measured continuously along five dimensions: time, asset, market tier, information type, and market state.

15.1.1 The efficient market hypothesis and crypto markets

Any serious discussion of market efficiency begins with the efficient market hypothesis (EMH), proposed by Fama (1970) in his landmark review [1]. The core proposition of the hypothesis is that in an informationally efficient market, asset prices fully and instantaneously reflect all available relevant information, so no market participant can consistently earn excess returns above the market average on the basis of existing information. EMH not only provided a theoretical cornerstone for asset pricing and portfolio theory but also shaped investment practice and financial regulation over the past half century. Fama (1970) divided market efficiency into three forms, from weak to strong, according to the breadth of the information set, and this classification remains the classic point of departure for testing market efficiency [1].

The first is weak-form efficiency, which assumes that current prices already fully reflect all historical trading data, including past price series, volumes, and other market-generated information. If a market is weak-form efficient, then any technical analysis based on historical price patterns—whether moving-average crossovers, the relative strength index, or candlestick patterns—will be unable to generate excess returns systematically. In that case, future price movements appear as an unpredictable "random walk," because the value of all historical information has already been fully exhausted in the current price. Tests of weak-form efficiency typically rely on statistical tools such as autocorrelation analysis and variance ratio tests, which are discussed in detail below.

The second is semi-strong-form efficiency, which goes further and assumes that prices reflect not only historical trading data but also all publicly available information. Here "public information" covers an extremely broad range, including corporate financial reports, macroeconomic data releases, changes in regulatory policy, news coverage, and even public discussion on social media. If this assumption holds, then even deep analysis of public information (that is, fundamental analysis) cannot consistently earn excess returns, because once information becomes public its value is instantly incorporated into the price. The classic test of semi-strong-form efficiency is the event study method, which assesses how fast and how completely the market reacts to new information by measuring the cumulative abnormal returns of an asset before and after a specific event. This method is applied systematically to crypto markets below.

The third and most stringent form is strong-form efficiency, which assumes that prices reflect all information, including inside information that has not yet been made public. In a strong-form efficient market, even insiders holding non-public information cannot gain an advantage. This assumption has long been regarded as too extreme even in traditional financial markets, because the very existence of insider-trading laws implies that inside information can indeed generate excess returns. In crypto markets, strong-form efficiency is likewise untenable: abnormal price movements before a token's listing and pre-emptive sell-offs by project teams both point to the significant role of inside information in price discovery.

In the early development of digital asset markets, the EMH framework, and especially its emphasis on informational transparency, seemed to align neatly with blockchain's "trustless" and "publicly verifiable" characteristics. As a public ledger, the blockchain in principle exposes the entire transaction history (the weak-form information set) and a great deal of protocol-state information (a kind of semi-strong-form information set) to all market participants. This unprecedented data transparency led many to expect that crypto markets could become an ideal testing ground for EMH, perhaps even reaching higher efficiency than traditional markets. As the market developed and academic research deepened, however, EMH's static, absolute assumptions proved to have clearly limited explanatory power for the many anomalies of crypto markets.

Empirical research has repeatedly shown that the efficiency of crypto markets is not a binary "yes" or "no" state but a continuous spectrum that varies with time, asset, and market conditions. Urquhart's (2016) early systematic tests of Bitcoin found significant evidence of weak-form inefficiency, but the inefficiency weakened in later subperiods,

suggesting that efficiency was improving [2]; this marked the beginning of systematic research on crypto market efficiency (details in Section 15.2.2). A large body of subsequent research confirmed the time-varying nature of efficiency: Tran and Leirvik (2020) found that the efficiency of five major cryptocurrencies was generally low before 2017 and improved gradually thereafter [3], while the cross-sectional tests of Kang, Lee, and Park (2022) found that crypto markets remained predominantly inefficient, with only about 6% of the sample satisfying the weak-form efficiency hypothesis [4]. These seemingly contradictory conclusions precisely illustrate the limits of describing crypto market efficiency with a single static label.

A deeper challenge comes from recognizing the gulf between "information being public" and "information being incorporated into price." Blockchain transparency means that information is "publicly available," but this does not mean that the information has been "fully digested" by the market and reflected in prices. Azar, Olivas, and Sinha (2025), in a study at the Federal Reserve Bank of New York, used the hacks of DeFi protocols as a unique natural experiment to quantify this gulf precisely: roughly 36% of the total price decline occurred between the moment the attack was recorded on the blockchain (information became public) and the moment the news spread widely on social media (information became common knowledge) [5]. This structural efficiency gap, created by information processing costs, is deconstructed in greater depth in Section 15.3.3.

Table 15-1 summarizes the academic literature's main test results for the three forms of EMH in crypto markets, organized by information type and time scale.

EMH form	Information set	Time scale	Asset class	Main test method	Representative findings	Representative literature
Weak-form	Historical prices and volume	Intraday (minute-level)	BTC/ETH	Variance ratio test, autocorrelation analysis	Major assets approximate a random walk at high frequency, with relatively high efficiency	Yi et al. (2023) [6]
Weak-form	Historical prices and volume	Daily	BTC	Variance ratio test, BDS test	Significantly inefficient early on; efficiency improved later but predictability persists	Urquhart (2016) [2]; Noda (2021) [7]
Weak-form	Historical prices and volume	Daily	Small- and mid-cap coins	Variance ratio test, calendar effect test	Pervasive significant weak-form inefficiency; calendar effects more pronounced	Tran & Leirvik (2020) [3]
Weak-form	Historical prices and volume	Weekly and above	BTC/ETH	Backtesting of momentum and reversal strategies	Medium-term momentum and long-term reversal effects present	Vidal-Tomás et al. (2019) [8]
Semi-strong-form	Macroeconomic and regulatory information	Event window (minutes to days)	BTC	Event study	Rapid reaction to standardized binary events; close to semi-strong-form efficient	ETF case in this chapter's introduction
Semi-strong-form	On-chain data and protocol events	Event window (hours to days)	DeFi tokens	Event study	Information processing costs cause significant lag; about 36% of the price move precedes the announcement	Azar et al. (2025) [5]
Strong-form	Including inside information	Event window	Various tokens	Detection of abnormal pre-listing returns	Widespread signs of insider trading; strong-form efficiency does not hold	Multiple industry reports

Table 15-1. Summary of empirical tests of the three forms of EMH in crypto markets (Data source: compiled by the author based on [2][3][5][6][7][8])

As the table shows, the efficiency of crypto markets is highly heterogeneous. At the

weak-form level, major assets are already close to efficient on high-frequency time scales, but significant predictability remains at lower frequencies and in small- and mid-cap assets. At the semi-strong-form level, the speed of reaction to different types of information varies enormously; standardized information is processed far faster than crypto-native on-chain information that requires complex interpretation. Strong-form efficiency scarcely holds in crypto markets at all: abnormal price movements before token listings and pre-emptive sell-offs by project teams both indicate the significant role of inside information in price discovery. Because testing strong-form efficiency already faces fundamental difficulties of data availability and legal definition even in traditional financial markets, and because crypto markets lack a standardized definition of insider trading and an enforcement framework, this chapter focuses on the systematic testing of weak-form and semi-strong-form efficiency, as these two levels both have more mature methodological tools and the most direct analytical value for understanding the institutional efficiency of perpetual futures. Strong-form efficiency is touched on indirectly in the analysis of the latent period in the FTX case (Section 15.5.1). This multilevel, multidimensional variation in efficiency is precisely what EMH's static framework struggles to capture in full, and it provides ample motivation for introducing a more explanatory theoretical framework.

15.1.2 The adaptive markets hypothesis

The challenge to EMH did not begin with crypto markets; it had been brewing within traditional finance for many years. The rise of behavioral finance, exemplified by the prospect theory of Kahneman and Tversky (1979), systematically revealed the cognitive biases pervasive in investor decision-making, such as loss aversion, overconfidence, and the anchoring effect. At the same time, a large body of empirical research uncovered anomalies in traditional markets that EMH struggled to explain, including the momentum effect, the value effect, calendar effects, and excess volatility. These findings created a seemingly irreconcilable gulf between EMH and behavioral finance: one side held that markets are rational, while the other held that markets are rife with irrational behavior.

The adaptive markets hypothesis (AMH) proposed by Lo (2004) was intended precisely to bridge this gulf [9]. AMH does not reject EMH wholesale; rather, it treats EMH as an extreme case that holds only under specific ideal conditions. Lo's (2004) core insight was to bring the evolutionary principles of biology—competition, adaptation, and natural selection—into the analysis of financial market behavior [9]. In the AMH framework, market participants are neither the perfectly rational economic agents assumed by EMH nor the systematically irrational individuals depicted by behavioral finance, but boundedly rational actors somewhere in between. They use heuristics to make decisions; these heuristics are effective in most situations but can produce systematic biases in specific environments. Crucially, these biases are not simply "errors" but adaptive strategies developed over the course of a long evolutionary process to cope with a complex environment.

The core propositions of AMH can be summarized as several interrelated claims. First, market efficiency is not a fixed attribute but a continuous variable that fluctuates with

time and environment. A market may be highly efficient in some periods and deviate significantly from efficiency in others. Second, the level of market efficiency is jointly determined by the composition of participants, the intensity of competition, and the available profit opportunities. When a great deal of capital and intellectual resource pours in to chase a particular inefficiency, that inefficiency gradually disappears and the market moves toward efficiency along that dimension. Conversely, when the market encounters structural change or an external shock, the old equilibrium is broken and new inefficiencies emerge. Third, the influence of behavioral biases is not constant but depends on the market environment. In calm markets, rational arbitrage forces may be sufficient to offset the influence of behavioral biases; but under extreme stress, fear and greed may overwhelm rational analysis, causing market efficiency to fall sharply.

AMH provides a highly explanatory lens for understanding digital asset markets, and its applicability may even exceed that in traditional markets. Crypto markets are themselves a rapidly evolving ecosystem: their participant base has evolved from the original cypherpunks and tech enthusiasts, through the mass influx of retail speculators, to today's deep participation by professional market makers, quant funds, and large financial institutions. Their infrastructure has evolved from the unstable centralized exchanges of the early days to today's sophisticated derivatives platforms and decentralized trading protocols that support high-frequency trading. Their regulatory environment has evolved from a blank slate to a growing focus of global regulators. In an environment where participants, technology, and rules are all changing dramatically, treating efficiency as a fixed attribute is unrealistic.

Academic research provides strong empirical support for the applicability of AMH to crypto markets. Noda (2021) used a time-varying model based on generalized least squares to test directly the dynamic evolution of Bitcoin and Ethereum market efficiency [7]. The study found that the efficiency of both cryptocurrencies did change significantly over time, exhibiting a clear pattern of cyclical fluctuation, which aligns closely with AMH's core prediction about the time-varying nature of efficiency. Phases of rising efficiency were typically accompanied by increased institutional participation, improved trading infrastructure, and deepening liquidity, whereas phases of declining efficiency were often associated with market bubbles, extreme volatility, or major negative events. Tran and Leirvik (2020) further confirmed this finding, observing that crypto market efficiency was generally low before 2017 and improved gradually as the market matured, though this process was not monotonic but accompanied by cyclical setbacks [3]. Together, these findings indicate that crypto market efficiency, just as AMH predicts, is in a dynamic process of continual adaptation, competition, and evolution.

Figure 15-1 traces the theoretical evolution from EMH to AMH and then to this chapter's efficiency spectrum framework.

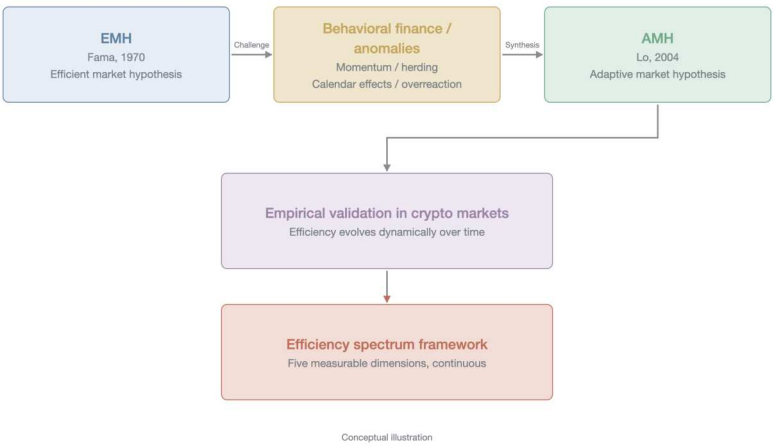


Figure 15-1. From EMH to AMH: the theoretical evolution

As the figure shows, EMH provided the classic classification framework for market efficiency; behavioral finance and market anomalies challenged it; Lo's (2004) AMH achieved a theoretical synthesis by introducing an evolutionary perspective; and empirical evidence from crypto markets further validated AMH's predictions. The efficiency spectrum framework proposed in this chapter operationalizes AMH's qualitative insights into a measurable, comparable, five-dimensional analytical tool.

15.1.3 The efficiency spectrum framework

Since EMH's static dualism is too rigid, and since AMH's dynamic evolutionary view offers the right way of thinking but leans toward macro narrative, we need a more operational analytical tool that turns AMH's deep insights into a measurable, comparable framework. To this end, this chapter proposes the efficiency spectrum as its core analytical framework. The central claim of the framework is that any statement about market efficiency is incomplete, or even misleading, if it is divorced from specific dimensional qualifications. We cannot assert in general terms that "the Bitcoin market is efficient" or that "crypto markets are inefficient"; we must specify precisely on which time scale, for which class of asset, for which type of information, and in which market state a given efficiency performance holds. The efficiency spectrum framework decomposes efficiency into five key dimensions—time, asset, market tier, information type, and market state—forming a multidimensional assessment matrix, as shown in Figure 15-2.

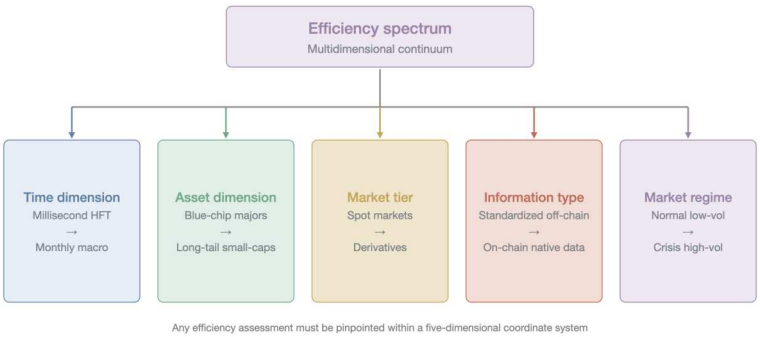


Figure 15-2. The five-dimensional efficiency spectrum matrix

As the figure shows, the five dimensions of the efficiency spectrum approach the assessment of efficiency from different angles and together constitute a complete analytical space. Any specific efficiency assessment can be located at a particular coordinate within this five-dimensional space. Each dimension is explained in turn below.

The time dimension: the efficiency gradient from milliseconds to months. Market efficiency can differ enormously across time frequencies, and this is the first key dimension of the efficiency spectrum. A market may approximate a random walk at the daily level, exhibiting relatively high weak-form efficiency, yet at the minute or even millisecond level, owing to microstructure factors—such as the limited depth of the order book, the friction created by trading fees, and delays in market makers' quote updates—it may contain fleeting predictable patterns that high-frequency strategies can capture. At the other end, on weekly and monthly time scales, medium- to long-term patterns such as momentum and mean reversion may also exist, although the persistence and economic significance of these patterns are themselves time-varying. In addition, crypto markets have their own institutional cycles, the most typical being the funding settlement cycle of perpetual futures, which creates predictable price behavior at specific points in time, as discussed in detail below. It is therefore essential to specify the time scale when discussing efficiency: intraday high frequency, daily, weekly, or a longer horizon.

The asset dimension: the efficiency gap from blue chips to the long tail. Different digital assets differ markedly in their level of efficiency, and this constitutes the second dimension of the efficiency spectrum. Bitcoin and Ethereum, as the two assets with the largest market capitalization, the deepest liquidity, the widest analyst coverage, and the deepest institutional participation, typically sit at the top of the crypto-asset efficiency spectrum. These two assets have the narrowest bid-ask spreads, the deepest order books, the most trading venues, and the most active arbitrage, all of which help improve price discovery. For mid-cap major coins, efficiency is typically lower than for Bitcoin and Ethereum because of relatively lower liquidity, insufficient information coverage, and fewer analytical tools. And for the vast universe of long-tail assets—including newly issued tokens, small-cap projects, and extremely illiquid trading pairs—markets often exhibit lower efficiency and more pronounced anomalies, such as wider spreads, slower reactions to information, and more frequent price manipulation. This efficiency gradient

along the asset dimension follows a logic similar to the phenomenon in traditional equity markets whereby large-cap stocks are usually more efficient than small-cap stocks, but in crypto markets the gradient may be even steeper, given the far greater number of assets, more severe information asymmetry, and more uneven regulatory coverage.

The market-tier dimension: efficiency divergence from spot to derivatives. Even for the same underlying asset (such as Bitcoin), efficiency differs systematically across market tiers—spot, perpetual futures, dated futures, and options—and this constitutes the third dimension of the efficiency spectrum. The difference stems mainly from heterogeneity in participant composition, transaction costs, and leverage structure. Because the perpetual futures market offers very high leverage, the convenience of short-selling, and greater capital efficiency, it systematically attracts a large number of informed traders with information advantages and advanced analytical capabilities. This concentration of trader types means that perpetual futures exhibit faster information reflection and higher pricing efficiency than the spot market on many time scales. Yet the institutional designs unique to perpetual futures (such as the funding settlement mechanism, the mark-price smoothing mechanism, and the oracle update frequency) may also introduce additional friction and predictability at specific moments, forming a dimension of "institutional efficiency" that goes beyond the classic three-part EMH taxonomy—these are not about the speed of reaction to external information but about efficiency fluctuations endogenously created by the market's own rules. We treat the market tier as an independent dimension rather than a subordinate of the asset dimension because the efficiency difference for the same asset across tiers is often larger than the difference for different assets within the same tier, and the analytical logic is entirely different. This dimension recurs throughout the chapter: in Section 15.2 we compare the weak-form efficiency of spot and perpetual futures; in the event studies of Section 15.3 we construct a dual-market comparison to examine semi-strong-form efficiency; in Section 15.4 we dissect how the institutional design of perpetual futures creates predictable efficiency cycles; and in the FTX case of Section 15.5 we show how the efficiency advantage of a market tier can flip into systemic risk under extreme stress.

The information-type dimension: processing costs from standardized signals to complex on-chain data. This is the core dimension of the semi-strong-form efficiency test and one of the most distinctive analytical angles in crypto markets. The speed and degree of the price reaction to different types of public information differ systematically, and the root of these differences lies in differences in information processing costs. Information in crypto markets can be divided roughly into several tiers. The first tier is standardized macroeconomic data and clear regulatory decisions, such as a Federal Reserve rate decision or an SEC approval or rejection of an ETF. Such information is binary and easy to interpret, and the market usually reacts to it fastest, often completing the price adjustment within minutes or even seconds, as in the ETF case in the introduction. The second tier is crypto-native information that requires some expertise to interpret, such as a major protocol upgrade, the passage of a governance proposal, or a change to a token's economic model. Such information has a higher processing cost than the first tier, and the market's reaction is correspondingly slower. The third tier is complex on-chain data, such as the movement of funds by large addresses, anomalous interactions with smart contracts, or the exploitation of a vulnerability in a DeFi protocol. As

Azar, Olivas, and Sinha (2025) reveal, although such information is publicly available on the blockchain, it requires highly specialized tools and knowledge to interpret; its processing cost is the highest, so the lag in the price reaction is also the most pronounced [5]. This dimension reveals a key insight: in crypto markets, "information being public" does not equal "information being incorporated into price," and information processing cost creates a structural hierarchy of information advantage, allowing participants with advanced analytical capabilities to capture information rents systematically.

The market-state dimension: the efficiency transition from normal conditions to crisis. The fifth dimension of the efficiency spectrum concerns the influence of the market's macro state on the level of efficiency. Market efficiency is not constant across all conditions but is highly dependent on the prevailing environment, exhibiting significant state dependence. Market states can be divided roughly into several typical modes. Under the "normal" state—ample liquidity and moderate volatility—competition is intense, arbitrage is active, and efficiency is typically at a high level. Under a "stressed" state induced by an external shock, such as the global market panic triggered by the COVID-19 pandemic in March 2020 (crypto's "Black Thursday"), liquidity contracts rapidly, market makers withdraw quotes, cross-exchange spreads widen sharply, and market efficiency declines significantly. More extreme still is the "decay" state induced by an internal collapse of trust: the collapse of the FTX exchange is a typical case, in which, once participants developed fundamental doubts about the solvency of their counterparties, liquidity did not merely contract but could evaporate entirely, the price discovery mechanism was nearly paralyzed, and market efficiency collapsed nonlinearly and sharply. This dimension is analyzed in depth in the FTX case study below. The process of recovery after a crisis deserves equal attention: the market usually does not return immediately to its pre-crisis efficiency level but passes through a "recovery" state of gradually rebuilding trust and liquidity, and the speed and completeness of this process are themselves important measures of market resilience.

In sum, the five dimensions of the efficiency spectrum—time, asset, market tier, information type, and market state—together constitute a complete toolkit for analyzing informational efficiency in digital asset markets. It discards the simple label of "efficient or inefficient" and treats efficiency as a continuous, multidimensional, dynamically changing spectrum. The value of this framework lies not only in its descriptive accuracy but also in its operational guidance: when we assess the feasibility of a trading strategy, we must specify the efficiency dimensions on which the strategy relies—on which time scale, for which class of asset, at which market tier, exploiting which type of information's inefficiency, and in which market state is it effective? Only by locating a strategy precisely within this five-dimensional coordinate system can we make meaningful judgments about its expected returns and risks. In the sections that follow, we use this framework to conduct systematic empirical tests of the weak-form and semi-strong-form efficiency of crypto markets, dissect the institutional efficiency of perpetual futures, and, through the FTX collapse, show the nonlinear behavior of efficiency under extreme conditions.

15.2 Testing weak-form efficiency

Within the efficiency spectrum framework established above, we begin with the most basic level of testing for digital asset markets: weak-form efficiency. The weak-form efficient market hypothesis defined by Fama (1970) holds that current asset prices already fully reflect all historical price and volume information [1]. If this hypothesis holds, then any technical analysis based on historical price patterns is futile, because the future path of prices appears as an unpredictable random walk. In other words, can past prices predict future prices? This question is the starting point of research on market efficiency and a basic yardstick for the maturity of a market.

The core argument of this section is that the weak-form efficiency of digital asset markets exhibits significant heterogeneity and dynamic evolution. For major assets such as Bitcoin and Ethereum, efficiency has strengthened significantly over time, approaching a random walk on high-frequency scales such as intraday; yet on longer time scales, momentum and reversal effects can still be detected, and for emerging long-tail assets, predictable price patterns are more prevalent. More importantly, these patterns are not fixed; their full life cycle—emergence, exploitation, and gradual weakening and even disappearance—provides direct empirical evidence for the adaptive markets hypothesis discussed in the previous section. The market is not statically stuck in a binary "efficient" or "inefficient" state but continually evolves in a dynamic process driven by participants' ongoing learning, competition, and adaptation. This section first introduces the core methodological tools for testing weak-form efficiency, then presents the empirical results for major assets, and, drawing on three concrete cases—perpetual-futures-native variables, calendar effects, and cross-exchange arbitrage—dissects the complex picture of weak-form efficiency in digital asset markets.

15.2.1 Testing methodology

The most direct way to test weak-form market efficiency is to analyze the statistical properties of the time series of asset returns. If the market is weak-form efficient, then the log-return series of an asset should exhibit no significant autocorrelation—that is, there should be no systematic linear relationship between returns at any lag and current returns. Positive autocorrelation implies a momentum effect: past rises foreshadow future rises, and past declines foreshadow future declines, which provides a theoretical basis for trend-following strategies. Negative autocorrelation implies a mean-reversion effect: past rises foreshadow future declines and vice versa, which creates opportunities for reversal strategies. Whether positive or negative, significant autocorrelation constitutes a departure from the random walk hypothesis and may offer traders systematic profit opportunities.

In practice, researchers typically compute the autocorrelation coefficients of a return series at multiple lags and use the Ljung-Box Q-statistic to jointly test whether these coefficients are collectively significantly different from zero. Simple autocorrelation analysis has limitations, however. The heteroskedasticity pervasive in financial time series—that is, volatility that itself varies over time—can bias the estimated standard errors of the autocorrelation coefficients and thereby undermine the reliability of the test. Moreover, the autocorrelation coefficient at a single lag captures only the linear

dependence at a specific frequency and struggles to characterize the predictability of the return series across different time scales.

It was precisely to overcome these limitations that Lo and MacKinlay (1988) proposed the variance ratio test, which became the classic method for testing the random walk hypothesis [10]. The core idea of the variance ratio test rests on a concise yet profound mathematical property: if a time series is a pure random walk, then the variance of its q-period return should be exactly q times the variance of the one-period return. This is because, under the random walk assumption, price increments in each period are independent and identically distributed, and the variance of a multi-period return equals the sum of the variances of the individual periods. The variance ratio VR(q) is therefore defined as:

$$\mathrm{VR}(q) = \frac{\mathrm{Var}(r_t(q))}{q \cdot \mathrm{Var}(r_t)}$$

where $\mathrm{Var}(r_t(q))$ is the variance of the q-period return and $\mathrm{Var}(r_t)$ is the variance of the one-period return. Under the null hypothesis of a random walk, the theoretical value of VR(q) should be 1. If VR(q) is significantly greater than 1, the return series exhibits positive autocorrelation—price changes are persistent, past trends tend to continue—which is the statistical fingerprint of a momentum effect. If VR(q) is significantly less than 1, there is negative autocorrelation—price changes are reversing, past extreme moves tend to be corrected—which is evidence of a mean-reversion effect.

The variance ratio test has several advantages over simple autocorrelation analysis. By comparing the variance relationships of returns across different time scales, it can capture predictable patterns at multiple frequencies simultaneously, rather than focusing only on a single lag. More importantly, Lo and MacKinlay (1988) provided test statistics under both homoskedasticity and heteroskedasticity assumptions, the latter (usually denoted $Z^*(q)$) being robust to the volatility clustering pervasive in financial time series [10]. In subsequent methodological developments, Chow and Denning (1993) proposed a joint variance ratio test that allows researchers to test the random walk hypothesis across multiple holding periods q simultaneously, avoiding the multiple-testing problem [11]. Wright (2000) proposed nonparametric variance ratio tests based on ranks and signs, further enhancing robustness to non-normal distributions [12]. These methodological advances have made the variance ratio test a powerful and broadly applicable toolkit for testing weak-form efficiency.

Even these improved methods, however, face distinctive challenges when applied to digital asset markets. The return distributions of crypto assets such as Bitcoin generally exhibit pronounced leptokurtosis and fat tails. Yi, Yang, Jeong, Sohn, and Ahn (2023), in their study of Bitcoin market efficiency, reported that the excess kurtosis of Bitcoin daily returns was as high as 42.5, far exceeding the assumption of a normal distribution [6]. This extreme non-Gaussian character means that even heteroskedasticity-robust test statistics may see the accuracy of their asymptotic distribution approximations degraded, reducing the statistical power of the test or causing the probability of a Type I error to deviate from its nominal level.

A more fundamental challenge is that, just as the adaptive markets hypothesis predicts, market efficiency is itself time-varying. Using a single test statistic covering the entire

sample period, however methodologically elegant, may mask the dynamic changes in efficiency across different phases of the market. A market that is "on average" close to a random walk over the full sample period may exhibit strong momentum during bull markets and mean reversion during bear markets. These two patterns may offset each other in a full-sample test, yielding a misleading conclusion of "market efficiency." Combining rolling-window analysis to examine the temporal evolution of metrics such as the variance ratio has therefore become a key methodological strategy for capturing the market's adaptive learning process. Noda (2021) made an important contribution in this direction, using a time-varying model based on generalized least squares (GLS) to measure the dynamic evolution of cryptocurrency market efficiency rather than delivering a static, black-and-white conclusion [7]. This methodological shift, from "is the market efficient?" to "when, and to what extent, is the market efficient?", is precisely the methodological reflection of the adaptive markets hypothesis in empirical research.

Table 15-2 summarizes the core features and applicable conditions of the methods commonly used in weak-form efficiency tests.

Test method	Core principle	Advantages	Limitations	Representative literature
Autocorrelation test and Ljung-Box Q	Tests whether the autocorrelation coefficients of the return series at each lag are zero	Intuitive and easy to compute	Sensitive to heteroskedasticity; captures only linear dependence	Box & Pierce (1970) [13]
Lo-MacKinlay variance ratio test	Compares the ratio of multi-period to single-period return variance	Provides a heteroskedasticity-robust statistic; multi-frequency information	Single holding-period test; asymptotic theory requires large samples	Lo & MacKinlay (1988) [10]
Chow-Denning joint variance ratio test	Tests the random walk across multiple holding periods simultaneously	Controls for the multiple-testing problem	Still relies on the asymptotic distribution	Chow & Denning (1993) [11]
Wright rank- and sign-based variance ratio test	Replaces raw returns with ranks or signs	Highly robust to non-normal distributions	Statistical power may be lower than parametric methods	Wright (2000) [12]
BDS test	Tests whether the return series is independently and identically distributed	Can capture nonlinear dependence	Sensitive to series length; does not indicate the direction of departure	Brock, Dechert, Scheinkman & LeBaron (1996) [14]
Rolling-window variance ratio	Recomputes the variance ratio within a moving time window	Captures the time-varying nature of efficiency	Window-size choice is subjective	Noda (2021) [7]

Table 15-2. Comparison of weak-form efficiency test methods (Data source: compiled by the author)

As the table shows, no single method can test weak-form efficiency perfectly. In practice, scholars typically use a combination of methods to strengthen the robustness of

their conclusions. For digital asset markets, given the extreme non-normality of return distributions and the time-varying nature of efficiency, a strategy that combines non-parametric methods (such as the Wright test) with rolling-window analysis is especially important.

15.2.2 The efficiency evolution of major assets

Digital asset markets, and Bitcoin in particular, offer a unique "natural experiment" for observing how an emerging market evolves from inefficient to efficient. Unlike traditional financial markets, which evolved slowly over decades or even centuries, the Bitcoin market underwent a dramatic transition from extreme inefficiency to near-weak-form efficiency in little more than a decade, a process compressed into a time window that can be observed and recorded in full.

Early academic research was almost unanimous that the Bitcoin market was inefficient. Urquhart (2016) conducted the first systematic weak-form efficiency test of Bitcoin, using a set of statistical tools including the autocorrelation test, the variance ratio test, and the BDS test [2]. Based on daily data from 2010 to 2016, his study found that Bitcoin returns departed significantly from a random walk and contained exploitable patterns. This conclusion was highly consistent with the microstructure of the market at the time: participants were mainly tech enthusiasts and early speculators, with professional market makers and institutional investors almost entirely absent; exchange infrastructure was weak, API stability poor, and order-book depth limited; and the lack of liquidity meant that large trades caused enormous price impact, while arbitrage was doubly constrained by technology and capital. In such a market environment, it is unsurprising that prices contained a great many predictable patterns.

Nonetheless, Urquhart's (2016) study already contained a noteworthy finding: when he split the sample into two subperiods, he found that the degree of inefficiency weakened in the later subperiod [2]. This subtle but important observation hinted that the market might be undergoing a process of improving efficiency. A large body of subsequent research confirmed this trend. Tran and Leirvik (2020) studied the five largest cryptocurrencies by market capitalization and found their efficiency levels to be highly time-varying, with the market generally inefficient before 2017 and efficiency improving thereafter [3]. Apopo and Phiri (2021) further distinguished efficiency across time frequencies, finding that Bitcoin daily returns largely satisfied the random walk hypothesis while weekly returns still exhibited significant predictability [15]. This finding aligns closely with the analysis of the "time dimension" in the efficiency spectrum framework proposed earlier: efficiency is not distributed uniformly across all frequencies but exhibits a gradient that declines from high to low frequency.

At high frequency, Aslan and Sensoy (2020) conducted a systematic study of the intraday efficiency of major cryptocurrencies and found that intraday efficiency exhibits a non-monotonic, U-shaped relationship with sampling frequency: efficiency is highest at the intermediate frequencies of 5 to 10 minutes and relatively lower at both the 1-minute high-frequency end and the 1-hour low-frequency end [16]. This means that microstructure noise at overly high frequencies and insufficient information aggregation at overly low frequencies can both create predictable patterns. Zhang, Chan, Chu,

and Sulieman (2020) found that during bull markets the hourly returns of Bitcoin, Ethereum, and Litecoin exhibited random-walk characteristics, but efficiency declined significantly during bear markets [17]. This state dependence of efficiency—the level of market efficiency changing with the market environment (bull or bear, high or low volatility)—is one of the core predictions of the adaptive markets hypothesis.

Figure 15-3 computes the rolling variance ratios VR(5) and VR(10) of Bitcoin daily returns from 2016 to 2025, using a rolling window of 252 trading days (about one year)—which both aligns with the traditional-market standard of "one trading year" and ensures the asymptotic properties of the variance ratio statistic. In interpreting departures of the variance ratio, we use the Lo-MacKinlay heteroskedasticity-robust statistic $Z^*(q)$ and judge, at the 5% significance level, whether VR(q) deviates significantly from 1.0. The VR evolution trend in this chapter's figure is consistent in direction with the full-sample test conclusions in the literature (such as Urquhart (2016) [2] and Noda (2021) [7]), but rolling-window analysis has some sensitivity to window size, and the magnitude of VR fluctuations may differ under different settings.

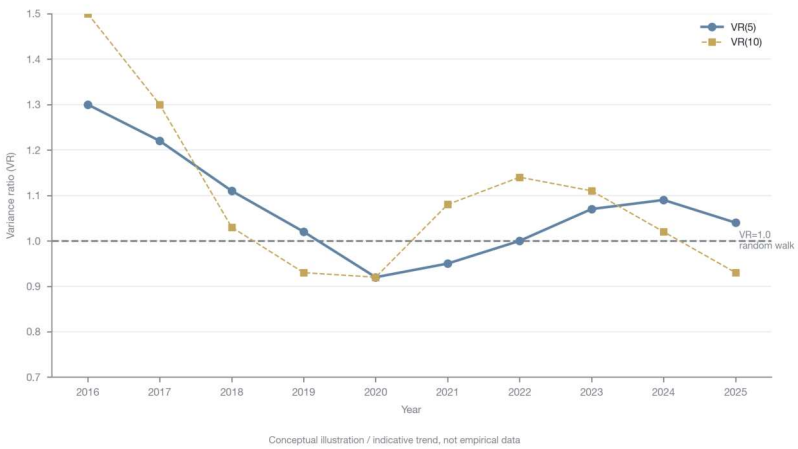


Figure 15-3. The temporal evolution of the BTC variance ratio

Figure 15-3, based on BTC/USD daily return data from major spot exchanges (Coinbase, Binance, and others), exhibits several key features. During the initial coin offering (ICO) bubble and bull market of 2017, VR values were persistently and significantly above 1, reflecting a strong momentum effect: "chasing the rally" dominated the market, and upward price trends had a self-reinforcing character. During this period, market participants were mainly retail speculators; information spread through social media and community word of mouth; and fear of missing out (FOMO) drove the continued climb in prices. The launch of CME Bitcoin futures in December 2017 was an important structural turning point: for the first time it gave institutional investors a standardized tool for shorting Bitcoin, allowing bearish views to be incorporated into prices more effectively and thereby dampening the one-sided momentum effect to some extent.

When the market suffered severe shocks, the behavior of the variance ratio also changed significantly. During the "Black Thursday" crash of March 2020, VR values briefly fell

below 1, showing the mean-reversion characteristic of a short-term oversold condition—that is, after panic selling the market quickly saw "bottom-fishing," and the extreme price decline was partly corrected. Notably, however, this crisis-period mean-reversion effect has been shrinking in both duration and magnitude, indicating that the market's self-correction mechanism is becoming more efficient.

The most striking long-term trend is the gradual narrowing of the range of variance ratio fluctuations. During 2016–2018, the range of VR(5) was roughly 0.7 to 1.5; by 2022–2025, this range had narrowed to about 0.85 to 1.15. The variance ratio increasingly falls within the "near-weak-form-efficient band" close to 1.0, indicating that consistently earning excess returns through simple historical price patterns is becoming increasingly difficult. The approval of U.S. spot Bitcoin ETFs in January 2024 further accelerated this trend: ETFs gave traditional financial institutions a convenient tool for Bitcoin exposure, and the influx of a large amount of professional capital significantly improved market depth and arbitrage efficiency.

As spot market efficiency improved, the derivative unique to digital asset markets—perpetual futures—exhibited a more complex efficiency profile. If we conduct rolling variance ratio analysis on the returns of BTC and ETH perpetual futures and form a "dual-track comparison" with the spot market, we find two distinct but internally consistent phenomena. On the one hand, at high frequency (minute-level), the variance ratio of perpetual futures is closer to 1.0 than that of spot—that is, more consistent with the random walk hypothesis. This is mainly because the perpetual futures market carries the vast majority of liquidity and price discovery functions, and informed traders and professional market makers are highly concentrated there, allowing new information to be digested more quickly. On the other hand, when we compute the variance ratio over the 8-hour cycle (the typical funding settlement cycle on major exchanges), perpetual futures exhibit a systematic departure from a random walk, whereas the spot market shows no comparable departure at this frequency. This departure is highly correlated with the periodic settlement of funding, suggesting that the distinctive institutional design of perpetual futures may endogenously create a new predictable pattern.

Noda's (2021) study provides a more rigorous econometric argument for this observation [7]. Using a GLS time-varying model that does not depend on sample size, he directly estimated the dynamic path of Bitcoin and Ethereum market efficiency. The study's three core findings align closely with our analysis. First, the efficiency of both cryptocurrencies did change significantly over time, exhibiting a clear pattern of cyclical fluctuation. Second, Bitcoin's efficiency level was higher than Ethereum's in most periods, consistent with Bitcoin's higher liquidity, longer trading history, and broader institutional participation. Third, more liquid markets evolved in efficiency faster, which provides empirical support for the proposition that "liquidity is the infrastructure of efficiency."

Nonetheless, even today, when efficiency has improved significantly, the Bitcoin market still exhibits detectable predictable patterns across days and on longer time scales. Zaremba, Bilgin, Long, Mercik, and Szczygielski (2021), using daily data on more than 3,600 cryptocurrencies, found a significant short-term reversal effect: the cryptocurrencies with the lowest returns on the previous day significantly outperformed those with

the highest returns on the previous day [18]. They further argued that this daily reversal effect stems mainly from the low liquidity of the vast majority of cryptocurrencies, since the reversal effect is most pronounced in the least liquid assets. This finding is entirely consistent with the analysis of the "asset dimension" in the efficiency spectrum framework: major assets are far more efficient than long-tail assets, and liquidity is the key variable explaining this difference. Caporale and Plastun (2020) focused on the momentum effect following abnormal single-day returns and found that in Bitcoin, Ethereum, and Litecoin, large single-day moves were followed by a significant next-day continuation effect, though the strength of this effect weakened over time [19]. Bianchi, Guidolin, and Pedio (2023) examined the dynamics of return predictability in cryptocurrencies from a more macro perspective and found that the time-series predictability of cryptocurrencies was generally lower than that of traditional asset classes but showed a time-varying yet significant sensitivity to investor-attention metrics [20].

These findings form a complex but internally consistent empirical picture: major assets are already close to a random walk at high frequency, but at low frequency and among long-tail assets, predictable patterns still exist. The gradual weakening of these patterns is precisely the process by which market participants continually "consume" inefficiency through arbitrage and learning, and it provides direct empirical support for AMH.

15.2.3 Perpetual-futures-native variables

In traditional asset pricing theory, weak-form efficiency focuses mainly on the ability of historical prices to predict future prices. The core derivative of digital asset markets, however—perpetual futures—introduces a set of entirely new native variables, the most central of which are the funding rate and the perpetual–spot basis. The existence of these variables provides a new dimension for testing weak-form efficiency: do the native signals of these derivatives markets contain predictability about future spot or contract returns?

The funding rate was designed to anchor the perpetual futures price to the spot index price. When the perpetual futures price is above spot, the funding rate is positive and longs pay shorts; when it is below, the rate is negative. He, Manela, Ross, and von Wachter (2024) conducted a systematic empirical study of the fundamentals of perpetual futures. They found that the magnitude of the perpetual futures price's deviation from the spot price is significant, with annualized mean absolute deviations as high as 60% to 90% across different crypto assets [21]. More importantly, this deviation not only exists but is highly predictable. They constructed a simple arbitrage strategy: open a position when the deviation exceeds a certain threshold and close it when the deviation reverts. The study showed that during their sample period (2020 to 2022), even after accounting for the high transaction costs faced by retail investors, this strategy for Bitcoin perpetual futures could still generate a Sharpe ratio as high as 1.8; for high-frequency, active market makers, the Sharpe ratio could even reach 3.5 [21]. As the study also notes, however, as arbitrage capital increased and market competition intensified, the average magnitude of these deviations has been shrinking at a rate of about 11% per year, so the Sharpe ratios above may already be significantly lower in the current (post-2024) market environment. Even so, the existence of this

predictability, even as its magnitude decays, still constitutes a structural challenge to the weak-form efficient market hypothesis.

The time-series characteristics of the basis (the difference between the perpetual futures price and the spot price) likewise exhibit significant predictable patterns. The empirical analysis of He et al. (2024) shows that past returns can significantly predict the future size of the basis, with high past returns often foreshadowing a larger perpetual futures premium, reflecting strong positive-feedback or momentum trading in the market [21]. This deviation does not expand without limit, however. Over time, the funding rate and basis deviations exhibit mean reversion, and in the later part of their sample, as arbitrage capital increased and market competition intensified, the average magnitude of these deviations showed a declining trend of about 11% per year [21].

Within the efficiency spectrum framework, how should we position this predictability arising from perpetual-futures-native variables? On the one hand, such high Sharpe ratios and persistent arbitrage opportunities seem to be compelling evidence of "weak-form inefficiency." They indicate that the market contains information that has not been fully digested, or that the arbitrage mechanism is severely obstructed. On the other hand, we can also understand it as an "institutional-friction premium." Perpetual futures have no expiration date, so when conducting basis arbitrage the arbitrageur cannot know when the position can be closed profitably; this "random maturity" character introduces additional liquidity risk and capital-occupation cost. In addition, differences in margin requirements and liquidation mechanisms across exchanges, as well as the tail risks unique to crypto markets (such as exchange failures), all constitute limits to arbitrage. This predictability may therefore not be a pure "free lunch" but compensation for the arbitrageurs who bear these complex frictions and tail risks. As market infrastructure improves (such as the spread of unified accounts and the establishment of cross-exchange clearing mechanisms), this "friction premium" is being gradually compressed, which again confirms the view of dynamically evolving efficiency in the adaptive markets hypothesis.

These seemingly generous arbitrage profits nonetheless require a more prudent risk assessment. The Sharpe ratios reported by He et al. (2024) are computed from the standard deviation of returns, but as the chapter will show later, the distribution of funding rates exhibits significant negative skew and excess kurtosis, which means the Sharpe ratio systematically understates tail-risk exposure. In practice, funding-rate arbitrage requires a trader to hold a spot long on one exchange and a perpetual futures short on another; the resulting cross-exchange margin lock-up typically means that the actual capital committed is two to three times the notional position, and a Sharpe ratio computed on actual capital committed rather than notional position would be significantly lower than the figures above. Moreover, exchange counterparty risk constitutes the tail risk hardest to hedge in this strategy. The collapse of FTX left many arbitrageurs' funds trapped in a bankrupt exchange, a loss entirely invisible in the conventional Sharpe ratio calculation. Zhivkov's (2026) empirical analysis of cross-exchange funding-rate arbitrage further confirms the severity of these frictions: although 17% of the observed sample exhibited an economically significant arbitrage spread, in a simulated delta-neutral portfolio as many as 95% of the trading opportunities could not be fully executed because of margin requirements and forced liquidations

[22]. The excess returns created by the predictability of the funding rate may therefore reflect compensation for bearing these complex frictions and tail risks rather than a costless exploitation of market inefficiency.

15.2.4 The discovery and disappearance of calendar effects

Calendar effects are one of the most classic illustrations of weak-form market inefficiency. They refer to the phenomenon whereby asset returns exhibit systematic regularities across specific time units (such as day of the week, month, or holiday). In traditional equity markets, researchers long documented anomalies such as the "Monday effect" (Monday returns significantly lower than other trading days) and the "January effect" (January returns significantly higher than other months). The existence of these effects directly violates the random walk hypothesis, because they offer simple trading strategies based on time information: if Monday returns are systematically lower than other days, an investor need only buy at Monday's close and sell on other days to earn excess returns.

For a digital asset market that trades continuously, seven days a week, 24 hours a day, year-round, the study of calendar effects has special significance. Some calendar effects in traditional markets stem from institutional factors (such as the accumulation of information over a weekend market closure and month-end fund-rebalancing needs), and the continuous-trading nature of crypto markets eliminates these institutional frictions. If calendar effects are still observed in such a market, their cause is more likely related to participants' behavioral patterns than to institutional arrangements. More importantly, the rapid evolution of crypto markets offers an ideal window for observing the full life cycle of a calendar effect (from discovery to disappearance), a process that in traditional markets might take decades to complete.

Early research on the Bitcoin market did find some significant calendar effects. The study by Caporale and Plastun (2019) was an important contribution to this area; using parametric and nonparametric methods to systematically test the day-of-the-week effect in Bitcoin, they found an abnormal positive return on Mondays [23]. This finding is striking because its direction is exactly opposite to the "Monday effect" (negative returns) observed in traditional equity markets. The researchers speculated that this might be related to the behavioral patterns of crypto market participants: retail investors tend to conduct research and make decisions over the weekend and execute their buys on Monday, thereby pushing up Monday prices. Aharon and Qadan (2019) found a significant "Monday effect" in Bitcoin in both returns and volatility [24].

According to the logic of the adaptive markets hypothesis, once such a predictable pattern is discovered by academic research and made public, the entry of arbitrageurs will gradually cause it to disappear. The logical chain of this prediction is clear and intuitive: after academic papers are published, traders and quant funds notice the pattern; they begin to buy on the expected low-return days and sell on the high-return days; these trades themselves flatten the return differences across days; and as more and more participants join the strategy, the profit space of the calendar effect is gradually compressed until it disappears entirely. This process is essentially a transmission chain in which information moves from "private knowledge" to "public knowledge" and then

to "priced in," and it is also the micro-mechanism by which market efficiency improves through competition and learning.

Figure 15-4 analyzes the average daily returns of Bitcoin by day of the week across different periods, illustrating this adaptive learning process.

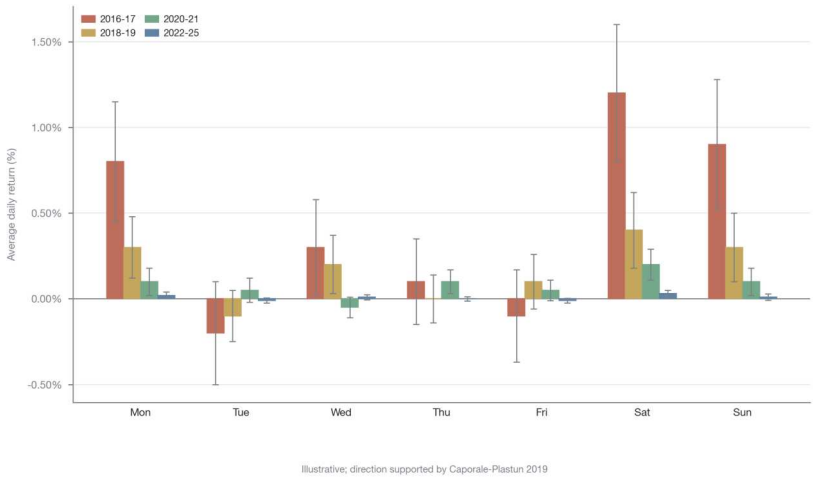


Figure 15-4. The discovery and disappearance of the BTC calendar effect (day-of-the-week effect)

Figure 15-4, based on BTC/USD daily return data from Coinbase and Binance, shows the average daily returns of Bitcoin by day of the week, along with their 95% confidence intervals (based on t-tests), across different periods. The figure reveals a clear pattern of temporal evolution. In the early stage of 2016–2017, there were significant differences in average returns across days of the week: weekend (Saturday and Sunday) average returns were clearly positive and sizable, while returns on some weekdays were relatively low or even negative. This difference constituted a potentially exploitable "weekend effect." Entering the bear-market cycle of 2018–2019, the return differences across days still existed but had clearly narrowed, and their statistical significance declined. By the bull-market cycle of 2020–2021, although overall market volatility increased, the systematic pattern of the day-of-the-week effect had become blurred. The most striking change appeared during 2022–2025: the average returns on all days were very close to zero, the differences across days were no longer statistically significant, and the 95% confidence intervals mostly covered zero. The day-of-the-week effect, once an exploitable "market anomaly," had essentially disappeared.

In crypto markets, however, beyond the day-of-the-week effect in the traditional sense, there is a distinctive "calendar effect" endogenous to the institutional design of perpetual futures: the "8-hour cycle effect" based on the funding settlement cycle. Major cryptocurrency exchanges (such as Binance, OKX, and others) typically settle funding once every 8 hours (for example, at 0:00, 8:00, and 16:00 Beijing time). Just before settlement, many traders choose to close positions briefly to avoid paying a high funding fee, causing prices to fluctuate in a regular way around settlement. This fluctuation pattern is not driven by fundamental information but is created purely by market

microstructure and institutional design.

Unlike the day-of-the-week effect, which has been almost entirely arbitraged away, this "8-hour cycle effect"—a perpetual-futures-native calendar effect—has not yet been fully arbitraged out. The reason is that arbitraging this effect requires high-frequency opening and closing of positions around settlement, which faces very high slippage costs and trading-fee friction. In addition, the moment of settlement is often accompanied by violent microstructure fluctuations, which increases the execution risk of the arbitrage strategy. The persistence of this native calendar effect therefore again confirms the key role of institutional friction in preventing the market from becoming fully weak-form efficient.

The timeline of this evolutionary process aligns closely with the structural changes in crypto markets. The years 2017–2018 were precisely when academic research on calendar effects in crypto markets was published intensively, and these findings spread rapidly to the trader community through academic conferences, financial media, and social networks. At the same time, professional quantitative trading firms and market makers began to enter crypto markets on a large scale, with the technical capability and capital to identify and exploit such statistical anomalies systematically. The combined effect of these forces meant that the calendar effect was effectively "arbitraged away" within a few years of being discovered.

Notably, the disappearance of the calendar effect does not mean that the market has become fully efficient across all time dimensions. More precisely, it illustrates that simple trading patterns—based on public information and easily replicated by a large number of participants—struggle to persist in an increasingly competitive market. This is precisely the core insight of the adaptive markets hypothesis: market efficiency is not a static attribute but a dynamic equilibrium continually shaped by the collective behavior of participants. Once an inefficiency is widely recognized, competition eliminates it; but new inefficiencies may reappear under new market conditions. For example, under extreme market stress, behavioral biases may temporarily overwhelm rational arbitrage, causing the calendar effect to "revive" briefly. This cyclical fluctuation in efficiency is precisely the essential characteristic of crypto markets as a rapidly evolving ecosystem.

15.2.5 The convergence of cross-exchange spreads

Another important dimension for measuring weak-form market efficiency is the spatial unity of price discovery—that is, whether the law of one price holds. In a fully efficient market, the prices of the same asset at different trading venues should be identical after deducting reasonable transaction and transfer costs. Any significant spread represents a near-riskless arbitrage opportunity, and the activity of arbitrageurs quickly flattens these spreads. The size of cross-exchange spreads and the speed of their convergence can therefore serve as a comprehensive measure of the efficiency of the entire market infrastructure and the speed of information transmission. If we regard the time-series predictability discussed earlier as the "time dimension" of weak-form efficiency, then cross-exchange spreads represent the "spatial dimension" of weak-form efficiency, testing whether prices are consistent at the same moment across different locations.

In the early stage of digital asset markets, cross-exchange spreads were not only perva-

sive but also sizable. The root of this phenomenon lay in the superposition of multiple structural barriers. At the technical level, early exchanges had poor API performance and low system stability, while blockchain networks had long confirmation times (Bitcoin required about 10–60 minutes) and volatile fees, making cross-exchange asset transfers both slow and expensive. At the capital level, the absence of professional market makers and arbitrage funds meant there was no sufficient "price-correcting" force in the market. At the institutional level, regulatory differences and capital controls across jurisdictions further impeded cross-border arbitrage.

The compounded effect of these barriers was most fully embodied in the "Kimchi premium" phenomenon. The Kimchi premium refers to the fact that the price of Bitcoin on Korean exchanges (such as Upbit and Bithumb) was persistently and significantly higher than on other major global exchanges. At the bull-market peak from late 2017 to early 2018, this premium once approached 50% [25]. This meant that if a trader could buy one Bitcoin in the United States for $10,000 and then sell it in Korea for $15,000, they could earn a 50% "riskless" profit. Korea's strict capital controls (restricting foreign capital inflows into the Korean crypto market), however, together with the banking system's restrictions on crypto trading, made this seemingly simple arbitrage strategy face enormous execution barriers in practice. Beyond capital controls, the exchange-rate risk of denominating arbitrage profits in Korean won, the compliance hurdles of Korean exchanges requiring local bank accounts and real-name verification, and the multi-day bank processing time needed to repatriate funds all further limited the participation of external arbitrageurs. The existence of the Kimchi premium was essentially an inefficiency "protected" by institutional friction.

As the adaptive markets hypothesis predicts, however, high profits incentivize market participants to keep finding ways to circumvent the barriers. Arbitrageurs developed more sophisticated strategies, including cross-border fund transfers through over-the-counter channels, the use of stablecoins (such as USDT) as an intermediary medium to circumvent fiat-conversion restrictions, and the establishment of pre-positioned fund pools across multiple exchanges to reduce the time cost of asset transfers. At the same time, the industry's infrastructure iterated rapidly: exchange API performance and stability improved dramatically, supporting higher-frequency automated trading; blockchain scaling solutions (such as the Lightning Network and Layer 2 solutions) reduced the cost and time of transfers; and professional quantitative trading firms entered with advanced technology and ample capital, raising arbitrage efficiency to a new level.

Figure 15-5 depicts the evolution of two types of cross-exchange BTC spreads since 2014.

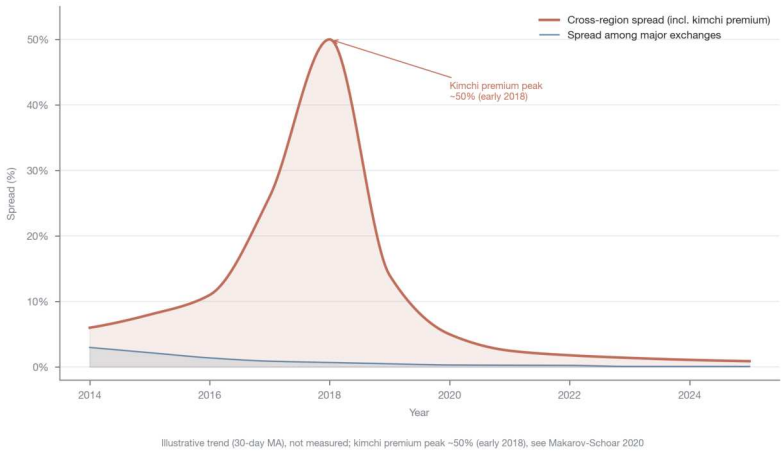


Figure 15-5. The convergence trend of cross-exchange BTC spreads

Figure 15-5 shows the long-run evolution of two types of cross-exchange BTC spreads, where the "spread between major exchanges" is defined as the absolute percentage of the daily average BTC/USD(T) spread between Coinbase and Binance, and the "cross-region spread" includes the premium between Korean exchanges (such as Upbit) and major global exchanges, with the data smoothed by a 30-day moving average. As the figure shows, both spreads exhibit a clear long-term downward trend, but the speed and path of convergence differ. The spread representing regional market segmentation (including factors such as the Kimchi premium), from a peak near 50% at the bull-market top of late 2017 to early 2018 (consistent with the Kimchi-premium measure above), underwent a relatively slow but sustained convergence and stabilized within 2% after 2022. The slowness of this convergence reflects the gradual nature of removing institutional barriers (such as capital controls), which technological progress alone cannot resolve, requiring also the evolution of the regulatory environment and the improvement of financial infrastructure.

By contrast, the spread between major global exchanges (such as between Coinbase and Binance) converged faster and more sharply. This type of spread converged rapidly from an early 2%–4% to below 0.5% after 2019, and even below 0.1% much of the time. This rapid convergence reflects the removal of purely technical and capital barriers: when exchange APIs are fast enough, network latency is low enough, and arbitrage capital is ample enough, the price differences of the same asset across major exchanges are eliminated almost instantaneously. Giudici and Pagnottoni (2019), analyzing high-frequency data from five major exchanges during the 2017 price surge and the 2018 correction, identified Bitfinex and Gemini as leaders in price discovery and found that inter-exchange connectivity was time-varying with the market state [26].

As spot-market spreads gradually converged, the spatial efficiency of derivatives markets also exhibited a distinctive structural feature. Are there significant funding-rate and basis differences between perpetual futures across different exchanges? Zhivkov (2026), analyzing 35.7 million high-frequency records from 26 cryptocurrency exchanges—11 centralized exchanges (CEXs) and 15 decentralized exchanges (DEXs)—

revealed a significant "two-tiered structure" in the cryptocurrency funding-rate market [22]. The study found that CEXs occupy an absolutely dominant position in price discovery, with market integration 61% higher than that of DEXs, and all significant information flows exhibiting unidirectional transmission from CEXs to DEXs. Notably, the funding-rate differences observed between CEXs and DEXs are not driven entirely by information asymmetry; structural differences in the funding-rate formulas across platforms are also an important explanatory factor. Binance uses a composite formula based on an interest rate component and a premium component, whereas dYdX and Hyperliquid use calculations based on time-weighted average price deviations; this formula difference produces structural rate deviations even under identical market conditions, so one must first control for this mechanical effect when assessing information-driven rate differences.

More crucially, although perpetual futures account for the vast majority of crypto derivatives volume, cross-exchange funding-rate arbitrage still faces numerous barriers. Zhivkov (2026) notes that as much as 17% of the observed sample exhibited an economically significant arbitrage spread (above 20 basis points), but after accounting for transaction costs and spread-reversal risk, only 40% of the top arbitrage opportunities could generate a positive return, and as many as 95% of arbitrage opportunities ultimately faced forced liquidation [22]. This indicates that although significant basis and funding-rate differences exist across platforms, the high transaction costs, margin requirements, and spread-reversal risk constitute powerful "limits of arbitrage" that prevent arbitrageurs from eliminating these large-scale mispricings. This resolves an apparent paradox: in a highly fragmented crypto derivatives market, significant price fragmentation and improving market efficiency in fact coexist.

The convergence of cross-exchange spreads is not a monotonically declining process, however. During periods of extreme market volatility, spreads spike briefly as liquidity evaporates instantaneously and the market panics. The "Black Thursday" of March 2020 is a typical case: Bitcoin plunged more than 40% within 24 hours, causing liquidity on several exchanges to evaporate almost entirely; market makers withdrew quotes, and cross-exchange spreads spiked from under 0.1% to several percentage points within minutes. The collapse of the FTX exchange in November 2022 triggered even more complex spread dynamics, involving not only a contraction of liquidity but also a repricing of counterparty risk, as market participants needed to assess which exchanges might be affected by the knock-on effects of the FTX collapse. These crisis-moment spread spikes are precisely a direct manifestation of the "market-state dimension" in the efficiency spectrum framework: market efficiency is highly dependent on the prevailing environment and may collapse nonlinearly under extreme stress.

Equally important, however, is observing the speed of spread recovery after a crisis. In the March 2020 crash, spreads between major exchanges recovered to normal levels within hours; and after the 2022 FTX event, although the recovery process was more protracted (involving the rebuilding of trust), spreads ultimately returned to their pre-crisis lows. This "shock–deviation–recovery" pattern indicates that digital asset markets have established fairly robust self-repair mechanisms. The resilience of the arbitrage infrastructure, which can resume operation relatively quickly even under extreme stress, is itself important evidence of improving market efficiency.

15.2.6 Overall assessment

Combining the analysis above, we can offer a multidimensional assessment of the weak-form efficiency of digital asset markets. Across the five dimensions of the efficiency spectrum framework, weak-form efficiency exhibits a clear gradient distribution and dynamic evolution.

Along the time dimension, efficiency declines in a gradient from high to low frequency. Major assets are already close to a random walk intraday (from minute to hour level), predictability at the daily level is weakening continuously, but momentum and reversal effects at the weekly and monthly levels can still be detected, although their economic significance is also declining. Along the asset dimension, efficiency falls steeply from blue chips to the long tail. The weak-form efficiency of Bitcoin and Ethereum is far higher than that of small- and mid-cap tokens, and liquidity is the core variable explaining this difference. The study by Zaremba et al. (2021) clearly shows that the daily reversal effect is concentrated mainly in the least liquid long-tail assets [18]. Along the market-state dimension, efficiency is at a high level under normal market conditions but may fall sharply under extreme stress; whether it is enhanced time-series predictability (such as a momentum crash or intensified mean reversion during a crisis) or a spike in cross-exchange spreads, both are manifestations of efficiency degrading under stress.

Along the market-tier dimension, the spot market and the derivatives market (especially perpetual futures) exhibit a complex interaction. On the one hand, because of its high liquidity and the concentration of professional institutions, the perpetual futures market shows higher efficiency than spot in high-frequency price discovery; on the other hand, its distinctive institutional design (such as the funding settlement cycle) and cross-platform arbitrage friction endogenously create new predictable patterns, such as the "8-hour cycle effect" and cross-exchange basis deviations. This dual character of "more efficient at high frequency, affected by institutional friction at low frequency" makes the weak-form efficiency of the derivatives market more subtle and multifaceted than that of the spot market.

From the perspective of dynamic evolution, the weak-form efficiency of digital asset markets has improved significantly over the past decade. This improvement is not accidental but driven by a series of structural factors: the entry of professional market makers and quantitative trading firms has provided continuous liquidity and arbitrage pressure; the development of derivatives markets (especially CME futures in 2017 and spot ETFs in 2024) has provided more channels for information to be incorporated into prices; technological iteration in trading infrastructure has reduced the cost and latency of arbitrage; and the gradual clarification of the regulatory framework has reduced institutional uncertainty. Each factor pushes the market toward a more efficient state, and their combined effect has clearly accelerated this evolution in recent years.

Nonetheless, simply labeling digital asset markets "weak-form efficient" remains inaccurate. A more precise description is this: major digital assets are already close to weak-form efficient on high-frequency time scales and under normal market conditions, but significant predictable patterns still exist on low-frequency time scales, in long-tail assets, and under extreme market conditions. The continued weakening of these patterns is evidence of the market's adaptive learning, but their complete disappearance

is neither inevitable nor imminent. This conclusion is entirely consistent with the core prediction of the adaptive markets hypothesis: efficiency is not an endpoint but a perpetually ongoing process, one driven by the competition, learning, and adaptation of participants.

Efficiency Spectrum Locator Card · Section 15.2

Dimension	Positioning	Quantitative proxy metric
Time	Efficient at high frequency; still predictable at low frequency	Rolling variance ratio VR(q)
Asset	High efficiency for blue-chip assets, low for long-tail assets	Autocorrelation coefficients across assets of different market caps
Information type	Historical price information (weak-form)	Ljung-Box Q-statistic / returns to momentum and reversal strategies
Market state	Efficient in the normal state; efficiency degrades under extreme stress	Magnitude of the variance-ratio shift during crises
Market tier	Derivatives discover prices at high frequency but are subject to institutional friction	Sharpe ratio of funding-rate forecasts / cross-exchange basis-arbitrage room

One-sentence takeaway: The weak-form efficiency of digital asset markets is not a static binary state but a dynamic process that exhibits a gradient across the dimensions of time, asset, state, and market tier, and that continuously evolves toward greater efficiency through participants' adaptive learning.

15.3 Testing semi-strong-form efficiency

Whereas weak-form efficiency tests focus on historical price information, semi-strong-form tests ask a broader question: can the market efficiently digest all newly public information? This is the core issue of the semi-strong-form efficiency test. In traditional financial markets, exploration of this question typically centers on standardized events such as corporate earnings reports and macroeconomic data releases. Digital asset markets, however, offer a distinctive environment with a more diverse range of information types and a more complex set of information sources. Here there are not only the macro and regulatory news common to traditional markets but also an entirely new category of information originating from the blockchain itself: on-chain events.

The core argument of this section is that the semi-strong-form efficiency of digital asset markets is not a constant but exhibits significant heterogeneity, whose key stratifying variable is information processing cost. Using the event study method, we systematically compare the speed of the market's reaction to three types of information:

standardized macroeconomic information, complex regulatory-game information, and highly technical crypto-native information (such as hacks). The research will show that the market reacts extremely fast to the first two types of information, close to the high-efficiency level of traditional markets; but for crypto-native information, even when its data is fully public on the blockchain, the high processing cost still creates a vast gulf between "being publicly available" and "being incorporated into price." We provide an in-depth reading of an important study from the Federal Reserve Bank of New York, which cleverly uses DeFi hacks as a natural experiment and, for the first time, quantifies the efficiency loss caused by information processing cost. Finally, we connect theory to reality by reconstructing the largest exchange theft in crypto history, the Bybit hack, to show concretely how, in a real crisis event, information is gradually processed, digested, and finally reflected in prices by different tiers of market participants.

15.3.1 The event study method

To measure the speed of the market's reaction to new information, the event study method is a long-established gold standard in financial economics. The method was pioneered and systematized by Fama, Fisher, Jensen, and Roll (1969) in their study of stock splits; its core logic is to use statistical methods to isolate the independent effect of a specific "event" on asset prices, thereby assessing the efficiency with which information is incorporated into prices [27]. A typical event study comprises several steps: first, precisely define the event itself and the "event window," a period before and after the event occurs; second, build a "normal" return model to predict the expected return of the asset had the event not occurred; third, compare the actual return during the event window with the expected return, the difference being the abnormal return, which represents the price impact brought by the event; and finally, cumulate the abnormal returns over the event window to obtain the cumulative abnormal return (CAR), which measures the total price impact of the event and the full trajectory of the information reaction.

In traditional financial markets, event studies typically use daily data, and the definition of the event is relatively standardized, such as an earnings announcement, a merger announcement, or a regulatory change. Digital asset markets, however, bring new opportunities and challenges to the application of this classic methodology. The greatest opportunity lies in the granularity of the data. Because crypto markets trade around the clock and the availability of high-frequency data (minute-level or even second-level) is far higher than in traditional markets, researchers can observe price reactions on a finer time scale, examining the process of information incorporation at the level of minutes or even seconds rather than days. This allows us to capture faster price discovery dynamics that were previously smoothed away in daily data.

At the same time, crypto markets pose new challenges. First, the definition of an event is more complex. Beyond news in the traditional sense, "events" in crypto markets also include on-chain activity such as protocol upgrades, hard forks, governance votes, and even hacks. The time these events occur can be pinned to the block timestamp, but interpreting their informational content requires expertise. Second, given the high volatility of the market and the absence of a widely accepted asset-pricing model, build-

ing a reliable "normal" return model becomes difficult. Early research typically used a simple market model or directly took Bitcoin as the market benchmark. For example, Saggu, Ante, and Kopiec (2025), in studying the SEC's regulatory interventions in crypto assets, computed abnormal returns by comparing the target asset's return with Bitcoin's return, thereby effectively isolating the regulatory shock specific to that asset [28].

Recent research, however, has achieved a key methodological advance, the most prominent example being the approach adopted by Azar, Olivas, and Sinha (2025) in analyzing DeFi hacks [5]. They recognized that in crypto markets the informational diffusion of an event is not instantaneous but involves a process from "occurring on-chain" to "being publicly known." They therefore creatively defined two distinct event time points: t_0, the time the first malicious transaction was confirmed on the blockchain, representing the moment information became "publicly available"; and t_{common}, the time mainstream social media (such as the project's official X account or a well-known security firm) first disclosed the event, representing the moment information became "common knowledge." By constructing event windows around these two time points separately, they were able to isolate precisely the price discovery that occurred within the "information asymmetry window" in which information processing cost played the dominant role. This dual-event-window design is a major innovation of the event study method in the crypto-market context; it deepens the analytical focus from "is the market efficient with respect to public information" to "how does the market react in stages to public information at different cognitive levels."

In addition, the existence of perpetual futures in crypto markets provides an entirely new dimension for the event study method. In a traditional event study, the analysis focuses mainly on the spot market. In crypto markets, however, an event study can be conducted simultaneously in both the spot and perpetual futures markets, constructing a "dual-market event study" methodology. By comparing the abnormal-return and cumulative-abnormal-return curves of the same asset in the spot and perpetual futures markets, researchers can reveal differences in the transmission speed of information across market tiers. For example, when facing sudden negative information, does the CAR curve of perpetual futures show a significant negative deviation earlier than the spot CAR curve? Can the dynamics of the perpetual futures funding rate and basis during the event window provide additional clues about how information is digested by leveraged traders? This dual-track comparison design allows us to locate the source of the information advantage more precisely and to assess differences in the efficiency with which different market microstructures process information. The analysis that follows in this section draws extensively on and builds upon these methodological innovations.

15.3.2 The stratification of information processing costs

The event study method provides a standardized tool for measuring the speed of the information reaction, and the diverse types of information in digital asset markets provide three test objects with distinct characteristics for this measurement. The semi-strong-form efficiency of the market is not a single number but a spectrum that depends on the characteristics of the information itself. Its core determinant is precisely the infor-

mation processing cost required to turn raw public data into a tradable signal. The lower the cost, the more standardized the information, the faster the market reacts, and the higher the efficiency; conversely, the higher the cost, the more complex the information, the slower the market reacts, and the deeper the efficiency gulf. This section examines, along this spectrum, the speed of the market's reaction to three typical types of events—macroeconomic information, regulatory and legal information, and crypto-native information—revealing how information processing cost systematically shapes the market's efficiency stratification.

The first type is standardized macroeconomic information. Such information—for example, a Federal Reserve rate decision or a release of the Consumer Price Index—is highly standardized, periodic, and predictable. The precise timing of the event is known in advance by market participants worldwide; the form in which the information is released is usually a clear number or a binary decision; and a broad consensus has already formed about the likely direction of its impact on asset prices. The processing cost of such information is therefore extremely low, and the main competition is over the speed of obtaining the information and executing trades. A series of reports published by LMAX Group in collaboration with Macro Hive, using millisecond-level trading data, offers an ideal window for observing this "speed race" [29]. According to the industry report, during major macro events such as a Federal Reserve meeting, the reaction speed of the Bitcoin market is already comparable to that of the most liquid foreign exchange market in the world, EUR/USD. On average, the first 2 seconds after an event absorbed about 12% of the price move over the following 10 minutes (BTC/USD, averaged across multiple macro events). This finding is of significant reference value: it indicates that for standardized, easy-to-interpret macro signals, the price discovery mechanism of crypto markets has already reached an efficiency level comparable to that of the most mature traditional financial markets. This owes to the deep participation of mature high-frequency trading infrastructure and algorithmic trading strategies, which can parse news feeds and execute trades at the fastest speed, compressing information processing cost to a minimum. In this process, the perpetual futures market often reacts more strongly than the spot market. For example, in the ETF approval event, the perpetual futures basis moved rapidly after the news broke, reflecting leveraged capital's rapid pricing of this standardized information. Comparable reaction speed does not imply comparable pricing quality, however. The "buy the rumor, sell the news" move described in this chapter's introduction—where the price fell rather than rose after the formal approval—suggests that the market's initial directional judgment may contain a systematic bias, and fast but inaccurate pricing itself constitutes a special form of efficiency defect. The intraday volatility triggered on January 9, 2024, when the SEC's social media account was hacked and a false approval message was posted, further exposed the fragility of crypto market participants in information verification—that is, a significant asymmetry between the market's ability to react instantly to unverified information and its ability to discern the authenticity of that information.

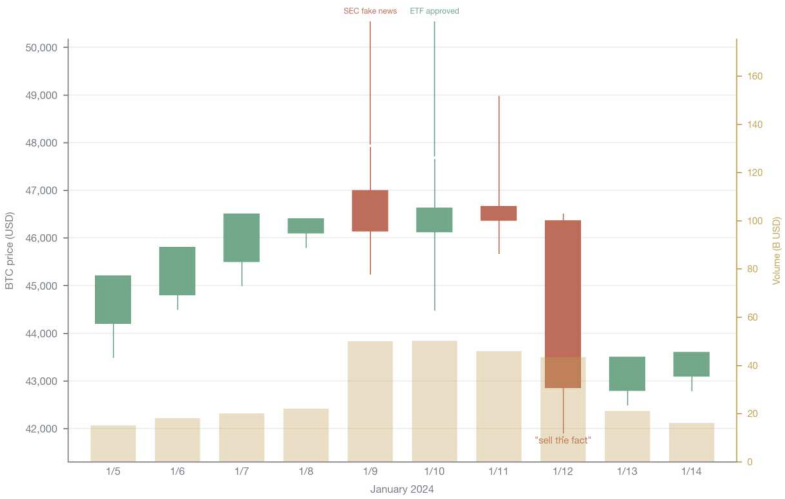


Figure 15-6. An event study of the Bitcoin ETF approval: BTC price movement and volume

Figure 15-6 shows the daily price movement and volume of BTC around the approval of the spot Bitcoin ETFs in January 2024. Two features are noteworthy: the false approval message posted after the SEC's social media account was hacked on January 9 triggered an intraday swing of about $1,500, highlighting the market's extremely fast reaction to unverified information; and after the ETFs' first trading day on January 11, when the price surged to about $49,000 and then fell back rapidly, breaking below $42,000 the next day, forming a textbook "buy the rumor, sell the news" move, while volume expanded within the window to a peak of about $50.1 billion. The daily open-high-low-close prices are given in Table 15-3 (Yahoo Finance, BTC-USD; verified); the candlesticks for January 5–8 and January 13–14 in the figure are illustrative extensions and are not part of the authoritative record in Table 15-3.

Date (2024)	Event	Open (USD)	High (USD)	Low (USD)	Close (USD)	Volume (billion USD)
January 9	SEC account hacked; false approval message posted	46,988	47,894	45,245	46,140	—
January 10	SEC formally approves spot Bitcoin ETFs	46,122	47,647	44,483	46,628	50.1
January 11	ETFs begin trading	46,656	48,969	45,679	46,369	45.8
January 12	“Sell the news” move continues	46,355	46,498	41,904	42,853	43.3

Table 15-3. Price snapshot on key dates around the Bitcoin ETF approval (*Data source: Yahoo Finance, BTC-USD*)

The second type is complex regulatory and legal information. Unlike standardized macro data, regulatory information has a significantly higher processing cost. When the SEC files a lawsuit against a particular crypto asset, or Congress proposes a new regulatory bill, such events are also public information, but their content is far more complex than a CPI figure. They involve interpreting legal texts, guessing regulators' intentions, and probabilistically assessing the outcome of a long legal battle ahead. The information is no longer a simple “bullish” or “bearish” but full of uncertainty and ambiguity. Saggu, Ante, and Kopiec (2025) conducted an event study of the SEC's enforcement actions unexpectedly classifying certain crypto assets as securities, and their results clearly show the lag effect brought by this high processing cost [28]. They found that after the SEC issued the relevant announcement, the prices of the affected assets did not adjust fully within seconds or minutes but began a slow decline lasting several weeks. On average, prices fell 5.2% in the first three days after the event and cumulatively fell 17.2% one month later. This slow, sustained decline reflects the time cost the market as a whole pays to digest and understand complex legal information. Traders need time to read the original announcement, wait for the analysis of legal experts, observe the response strategies of the projects, and continually reassess the asset's long-term risk. Here, although the information has been made public, its true meaning and impact require a process of gradually reaching consensus, and this process itself constitutes part of price discovery. For the perpetual futures market of the affected tokens, if it exists, the reaction of the funding rate and basis often leads the slow decline of the spot price, reflecting some informed traders expressing their pessimistic expectations about regulatory risk by shorting perpetual futures.

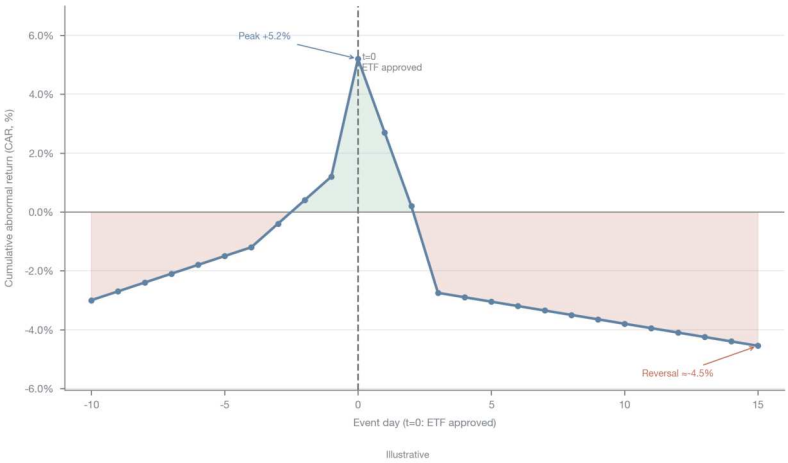


Figure 15-7. An event study of the Bitcoin ETF approval: the cumulative abnormal return curve

Figure 15-7 shows the cumulative abnormal return curve around the ETF approval event. The curve reveals the trajectory of the market's reaction to this standardized regulatory information: in the days before the event, the CAR had already begun to accumulate positively, reflecting the market's advance pricing based on probabilistic expectations; after peaking on the day of formal approval, the CAR quickly reversed and turned negative, indicating that "sell the news" behavior led to an excessive price correction. This pattern is highly consistent with the classic reaction to major policy events in traditional financial markets, indicating that crypto markets already possess an efficiency level comparable to that of mature markets when processing standardized binary information.

The third type is highly technical crypto-native information. This is the other end of the information-processing-cost spectrum and an efficiency challenge unique to digital asset markets. The typical representative of such information is a DeFi protocol being hacked. Technically, the moment the hacker's first malicious transaction is confirmed on the blockchain, the fact of the attack has already become "publicly available" information, recorded on an immutable global ledger. This "public" nature, however, is a highly misleading appearance. The raw on-chain transaction data is written in hexadecimal code, and understanding it requires deep smart-contract knowledge, a cryptographic background, and professional on-chain data analysis tools. For the vast majority of market participants, this data is technical data hard to interpret directly. Turning this raw data into a clear, tradable signal such as "protocol X was hacked for amount Y" requires a very high information processing cost. It is precisely this cost that creates an enormous time gap and cognitive gap between the information being "public on-chain" and "market consensus." Azar, Olivas, and Sinha (2025) precisely captured the price behavior within this time gap [5]. They found that between the on-chain time of a hack and the time the information was widely reported by the media, there is an "information asymmetry window" averaging several hours. Within this window, only a few informed traders with advanced on-chain analysis capabilities (such

as professional on-chain sleuths, security firms, and top quant funds) can decode the attack first and begin selling the relevant tokens. The study quantified that as much as 36% of the total price decline was completed within this window—that is, before ordinary investors even knew what had happened, the price had already fallen significantly. This finding is among the strongest evidence that semi-strong-form efficiency fails in crypto markets, and it starkly reveals the reality that "information being public does not equal information being incorporated into price." Here information processing cost plays the role of a "filter," determining who can extract value first from public but raw data and thereby gain a structural information advantage in the market.

Having explored the processing costs of these three types of information, we must introduce a crucial dimension: the information-amplification effect of perpetual futures. As the most active derivative in crypto markets, perpetual futures, through their distinctive institutional design, systematically accelerate the incorporation of information into prices. Compared with the spot market, perpetual futures offer high leverage and the convenience of short-selling. When new information appears in the market, especially negative information (such as a hack or a regulatory crackdown), a trader with an information advantage need not borrow the actual asset; they need only deposit a small amount of margin to express their view rapidly by shorting perpetual futures. This mechanism greatly reduces the friction cost of informed trading.

A large body of academic evidence supports the leadership of perpetual futures in price discovery. By computing metrics such as the Hasbrouck information share, research shows that for most of the time the perpetual futures market leads the spot market in reflecting new information. When the information processing cost is high, the few traders who can quickly parse the information preferentially choose the perpetual futures market to operate in, causing the perpetual futures price, funding rate, and basis to move first. Subsequently, arbitrageurs transmit these price signals to the spot market through cross-market arbitrage. Perpetual futures are therefore not merely a trading tool but the most important amplifier and conduit for information in crypto markets.

Taken together, these three types of information and the corresponding speeds of the market's reaction clearly delineate an efficiency spectrum determined by information processing cost. From high efficiency for macro information, to moderate efficiency for regulatory information, to significant inefficiency for on-chain native information, the market's semi-strong-form efficiency decays systematically as information complexity increases. This not only provides a unified explanatory framework for understanding the anomalies of crypto markets but also points the way to identifying potential trading opportunities and sources of risk. In the next section, we further dissect the study by Azar, Olivas, and Sinha (2025), deconstructing the composition of information processing cost in more detail and quantifying its specific impact on market efficiency.

15.3.3 A quantitative analysis of information processing costs

On the information-processing-cost spectrum, the report *The Price of Processing* by Azar, Olivas, and Sinha (2025), published at the Federal Reserve Bank of New York, is without doubt the most precise dissection to date of crypto-native events, the end of the spectrum with the highest technical barrier [5]. The study's core contribution is that it

uses the hacks that frequently occur in decentralized finance protocols as a series of ideal "natural experiments," thereby succeeding for the first time in quantifying the specific efficiency loss that information processing cost itself imposes on the market. This research provides decisive empirical evidence for the core argument that "information being public does not equal information being incorporated into price."

DeFi hacks make ideal natural experiments for several reasons. First, the attacks are exogenous—they are sudden events launched by external attackers, unrelated to other market activity, which effectively avoids the endogeneity problem. Second, the time an attack occurs can be recorded to the second by the blockchain ledger, providing unmatched precise timestamps for the event study. Finally, the outcome of an attack (the loss of funds) directly and unambiguously constitutes a negative shock to the value of the victim protocol's token. Using these properties, the research team constructed a detailed dataset of several hundred DeFi hacks and treated each attack as an independent event study.

The most central methodological innovation of the study lies in its fine-grained deconstruction of the concept of "the event occurring." Rather than adopting a single event time point, the researchers defined two key moments:

> t_0 (on-chain time): the time the attacker's first malicious transaction is confirmed by the blockchain network and recorded on a block. At this moment, the fact of the attack has already become publicly verifiable data on-chain, and the information is technically "publicly available." t_{common} (consensus time): the time mainstream information channels—such as the project's official social media account, a well-known blockchain security firm (such as PeckShield), or a mainstream crypto media outlet—first publicly disclose the attack. At this moment, the information is "translated" from raw on-chain data into human-readable language and becomes "common knowledge" for market participants.

By distinguishing these two time points, the researchers were able to divide the price discovery process into two stages: the "information asymmetry window" from t_0 to t_{common}, and the "information diffusion window" after t_{common}. The length of the former window, and the price movement that occurs within it, directly reflect the market-efficiency lag caused by information processing cost.

The study's quantitative results are striking. First, the average time gap from t_0 to t_{common} was as long as several hours, which itself shows the high time cost required to turn raw on-chain data into common knowledge. More important is the price reaction: the study found that before information became common knowledge, the price of the victim protocol's token had, on average, already fallen by 36% of its total decline. This means that more than one-third of the price shock occurred while ordinary investors did not even know an attack had taken place, completed first by a small group of informed traders with advanced on-chain monitoring and analysis capabilities. These informed traders monitor on-chain anomalies in real time through specialized tools; while ordinary people are still reading the news, they have already completed the entire process of analysis, decision, and trade.

This finding can be conceptualized through a "funnel model of how information is incorporated into price," as shown in Figure 15-8.

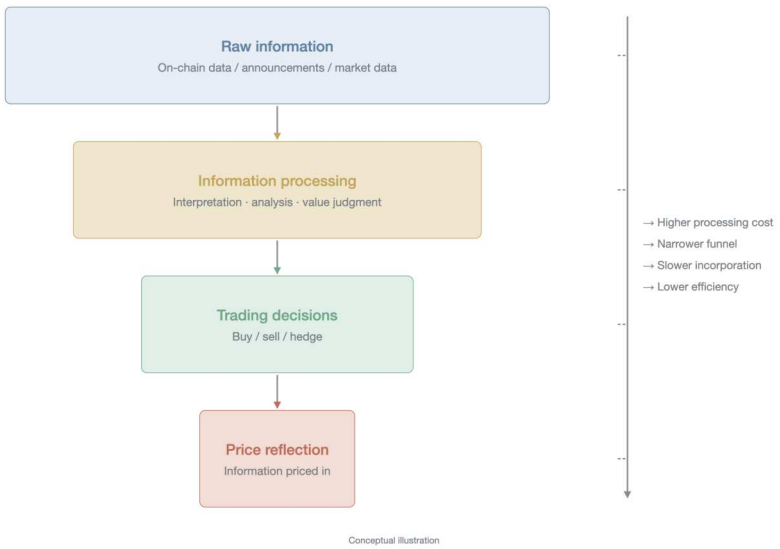


Figure 15-8. The funnel model of how information is incorporated into price

At the top of the funnel is raw, massive, noise-laden on-chain data, and information processing cost acts like each layer of a sieve: the first layer lets through only the very few top informed traders who can directly interpret raw transactions, forming the first wave of the price reaction; when security firms and the media "translate" and spread the event, the information penetrates the second layer and reaches a broader group of professional investors, triggering a second, sharper wave of decline; finally, through news and social media, it permeates the entire market. Information processing cost not only creates a time delay but also creates a hierarchy of profit—the earlier one penetrates the sieve, the higher the information rent one can capture.

For DeFi tokens listed simultaneously on a CEX, perpetual futures reflect attack information earlier because shorting requires no coin borrowing, and their price reaction often leads spot (for the mechanism, see Sections 15.3.2 and 15.4.4).

The study by Azar, Olivas, and Sinha (2025) therefore strongly demonstrates that blockchain transparency is a "raw transparency" that does not by itself guarantee market efficiency. On the contrary, in the absence of the corresponding interpretive tools and knowledge, this transparency may actually intensify information asymmetry. In digital asset markets, the true core competitive advantage is no longer whether one can obtain information (because information is open to everyone on-chain) but the speed and depth with which one processes and understands information. The study's conclusion has far-reaching implications for understanding the microstructure, risk management, and regulatory policy of crypto markets; it tells us that any assessment of crypto market efficiency must place information processing cost, the core variable, at the center of the analysis.

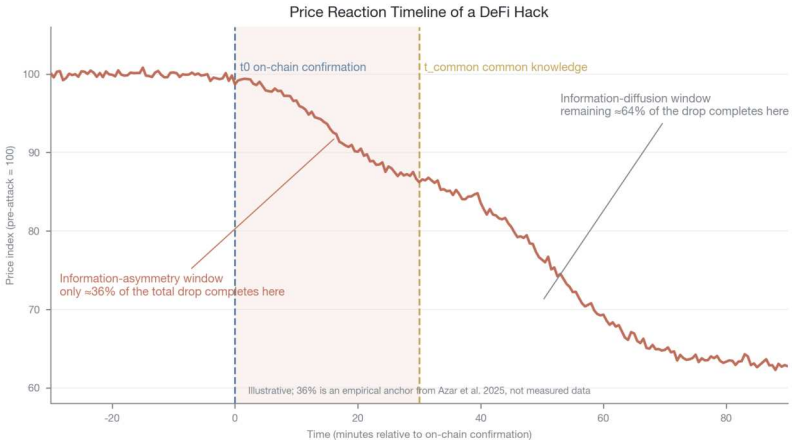


Figure 15-9. The timeline of the price reaction to a DeFi hack

Figure 15-9 uses a timeline to show the typical pattern of the price reaction to a DeFi hack: within the "information asymmetry window" between the attack occurring on-chain (t_0) and the information becoming common knowledge (t_{common}), about 36% of the price adjustment is already complete, while most market participants are still unaware. This staircase pattern of price discovery is highly consistent with the theory of information-processing-cost stratification proposed in Section 15.3.2.

15.3.4 A retrospective on the Bybit attack

The value of theory ultimately lies in its power to explain the real world. To make concrete the earlier abstract discussion of information processing cost and efficiency stratification, this section reconstructs the theft at the Bybit exchange on February 21, 2025. This largest single-exchange theft in crypto history offers a typical case for observing how information spreads, is processed, and finally affects prices in a real crisis. It clearly shows the time lag and cognitive gradient that exist between the birth of an on-chain fact and the formation of market consensus.

The starting point of the event can be traced to 2:13 p.m. UTC on February 21, 2025 (t_0). At that moment, through a carefully designed multi-stage attack, the attacker successfully breached the multi-signature security mechanism of one of Bybit's Ethereum cold wallets. The first step of the attack was a social engineering attack against the multi-signature authorizers, inducing them to sign a seemingly normal transaction request on a forged transaction interface; the second step exploited the proxy-contract architecture used by the Safe multi-signature wallet: using the obtained signature authorization, the attacker replaced the implementation contract behind the proxy contract with a contract containing malicious transfer logic, thereby bypassing the security assumptions of the multi-signature mechanism. This case shows that the weak link in multi-signature security often lies not at the cryptographic level but in the operational security of the signers and the upgradability of the smart-contract architecture. Over the following hour and more, the attacker transferred digital assets worth nearly $1.5 billion to addresses under their control without being detected. At this stage, the fact

of the attack had already been permanently recorded on the Ethereum blockchain as a series of transactions. For entities with advanced on-chain monitoring capabilities, these unusually large, unplanned transfers already constituted a strong danger signal. This is precisely the moment the "information asymmetry window" defined by Azar, Olivas, and Sinha (2025) opened. Within this window, a few top on-chain data analysis firms and quant funds that may have been monitoring exchange wallets became the first potential informed parties. Their actions (whether selling assets for risk-avoidance purposes or conducting more complex arbitrage trades) constituted the first driving force of price discovery, although this force may initially have been masked by the market's normal fluctuations.

The first diffusion of information occurred at 3:44 p.m. UTC (t_{common}), about 91 minutes after the on-chain attack began. Bybit's chief executive officer, Ben Zhou, posted a brief message on his personal social media account, publicly acknowledging for the first time that Bybit was "investigating a potential security incident." This moment can be regarded as the moment the information became "common knowledge." This message was itself vague and incomplete, however: it confirmed the existence of a problem but did not disclose the nature, scale, or specific loss of the attack. It therefore opened the second stage of the information's incorporation into prices. The market moved from "knowing nothing" to "knowing something had happened but not knowing how serious it was." Panic began to spread, but because of the lack of specific information, the price reaction remained limited and tentative. Traders faced classic Knightian uncertainty; not knowing the specific probability distribution of the risk, they could only trade on the basis of past experience and market sentiment. Within this 91-minute window, the proportion of the price move in both the spot and perpetual futures markets was significant, initially confirming the price discovery effect within the information asymmetry window.

As time passed, more information was gradually released and the funnel of information processing was penetrated further. In the hours after the CEO's first statement, through the tracking of on-chain analysts, follow-up media coverage, and Bybit's sporadic official updates, the estimated scale of the theft (initially estimated at several hundred million dollars, ultimately confirmed at nearly $1.5 billion) and the technical details of the attack gradually became clear. Each time a more specific, more definite piece of information was disclosed, it triggered a new round of price declines. Price movement was no longer an instantaneous reaction to a single event but a dynamic adjustment process driven by information flow and lasting several days. In the four days after the event, the price of Bitcoin fell from a level above $90,000 before the event to below the $90,000 mark. This process perfectly illustrates the true performance of the market's semi-strong-form efficiency when facing a complex, dynamic event: prices do not reflect all information in one step but converge toward a new equilibrium in an iterative, progressive manner as information processing cost gradually falls and consensus gradually coalesces.

In the Bybit event, the reaction of the perpetual futures market was especially striking and offers us a distinctive analytical dimension. Within the 91 minutes from the on-chain confirmation of the attack (t_0) to the CEO's public statement (t_{common}), although the spot market reacted relatively slowly, the perpetual futures market had already be-

gun to move anomalously. First, the funding rates of BTC and ETH perpetual futures rapidly flipped from positive to negative in the early stage of the event, indicating that short-side forces were quickly accumulating and that leveraged traders, anticipating a potential negative shock, were willing to pay a higher cost to maintain short positions. Second, the basis between perpetual futures and spot showed a significant negative deviation, with the perpetual futures price falling markedly faster than the spot price, further confirming the leadership of perpetual futures in digesting negative information. In addition, the open interest on major exchanges fluctuated violently after the attack was confirmed, and some long positions were forcibly liquidated as prices fell rapidly, this liquidation-cascade effect further amplifying the downward pressure on prices. These perpetual futures dimensions not only confirm informed traders' preference for leveraged tools but also show how the perpetual futures market accelerates the pricing of information through its distinctive microstructure during extreme events. The above description of anomalous moves in the perpetual futures market is based mainly on qualitative observation; limited by the availability of high-frequency data at the time of the event, this chapter cannot provide the precise basis-point figures of the funding-rate changes and basis deviations or a statistical significance test relative to the normal fluctuation range of the same period. Future research, after obtaining more complete high-frequency data, could conduct a more rigorous quantitative analysis of these anomalous moves.

Figure 15-10 uses a timeline to compare the relationship between key information nodes and the price reaction during the Bybit attack, showing the significant delay between the birth of an on-chain fact and its full reflection in prices.

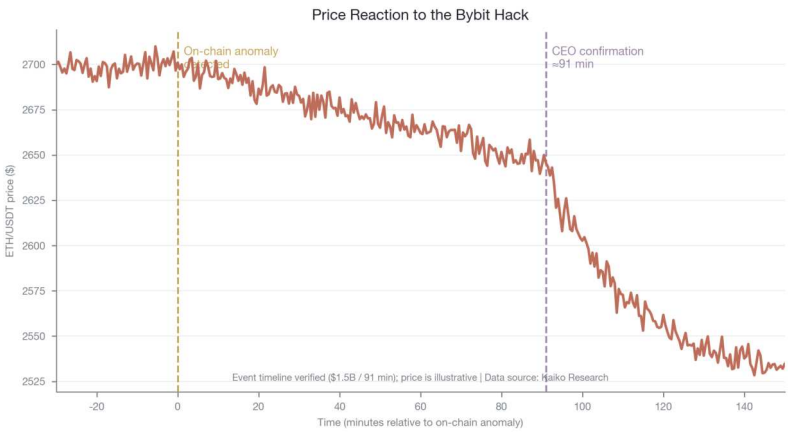


Figure 15-10. The price reaction to the Bybit hack

Figure 15-10 uses a timeline annotated with key nodes to show the price reaction of Bitcoin and Ethereum in the Bybit theft of February 21, 2025. Within the 91 minutes from the on-chain attack (14:13 UTC) to the CEO's first public statement (15:44 UTC), prices had already begun to move down, and over the following days, as the scale of the theft was gradually confirmed at nearly $1.5 billion, Bitcoin broke below the $90,000 mark; the funding rate and basis of perpetual futures deteriorated

continuously throughout the window, corroborating the leadership of the derivatives market in digesting negative information. The event nodes in the figure (the theft scale of about $1.5 billion, and the roughly 91 minutes between the on-chain anomaly at t0 and the CEO's first public statement) have been verified; the minute-by-minute curve for Ethereum is an illustrative trajectory (data source: Kaiko Research; not independently verified point by point) and is used only to present the direction and timing of the price reaction.

The Bybit case ultimately tells us that in digital asset markets, semi-strong-form efficiency is far from a simple "yes" or "no" question. It is a complex game about "who, at what time, at what cost, knows what degree of information." On-chain transparency provides raw material for everyone, but the ability to turn raw data into a tradable signal is held only by a few participants with professional tools and knowledge. This structural information advantage shaped by information processing cost is the key to understanding price behavior and the sources of risk in crypto markets.

Efficiency Spectrum Locator Card · Section 15.3

Dimension	Positioning	Quantitative proxy metric
Time	Significant information asymmetry and price lag within the event window (minutes to several hours)	Share of CAR within the t_0-to-t_{common} window (e.g., 36% in DeFi hack events)
Asset	Lower efficiency for tokens involved in a specific event (such as the token of a hacked protocol)	Abnormal return of the target asset relative to the market benchmark
Information type	High efficiency for macro information, moderate for regulatory information, lowest for crypto-native information (on-chain data)	Time to complete the event reaction for each information type (second-level vs. week-level vs. hour-level)
Market state	Under a sudden negative event, information asymmetry intensifies and market efficiency falls significantly	Widening bid-ask spreads and liquidity-contraction metrics during crises
Market tier	Perpetual futures lead spot in reflecting new information, playing an information-amplifying role	Magnitude of the negative deviation of the perpetual futures basis and the speed of funding-rate changes

One-sentence takeaway: The semi-strong-form efficiency of digital asset markets depends heavily on information processing cost, and perpetual futures, by virtue of their leverage advantage, play a leading pricing role in digesting complex crypto-native information.

15.4 Institutional efficiency

Having tested the reaction of digital asset markets to historical prices (weak-form efficiency) and external public information (semi-strong-form efficiency), our analysis now enters a deeper and more crypto-native layer. The earlier discussion centered mainly on how the market processes exogenous information that arrives at random, but in the digital asset domain, and especially in the perpetual futures market, market efficiency is also deeply shaped by its own internal institutional design. Perpetual futures, the cornerstone of the crypto derivatives market (their volume is estimated to account for more than 93% of all crypto futures volume [22]), are built around a complex set of price-anchoring and risk-management mechanisms involving the funding rate, the mark

price, and oracles. While anchoring the perpetual futures price to the spot price, these institutional designs also each create predictable price behavior patterns in their own way. Specifically, the mean-reverting nature of the funding rate mechanism creates structural arbitrage windows; the mark-price smoothing mechanism and the discrete updating of oracles cause systematic lags in information reflection; and high leverage combined with the liquidation mechanism triggers nonlinear price overshooting under extreme market conditions. These predictable efficiency fluctuations, actively shaped by institutional rules, constitute the core content of the "institutional efficiency" dimension of the efficiency spectrum.

This section dissects these institutional efficiency barriers through four progressive layers. We first examine the information content of the funding rate and its nature as a structural gap in weak-form efficiency; then analyze how the funding settlement cycle triggers a "U-shaped" fluctuation in efficiency at the microstructure level, and compare the impact of differences in settlement frequency across exchanges; next explore the information-reflection lag caused by the mark-price and oracle mechanisms; and finally, by analyzing the lead-lag relationship between spot and perpetual futures, reveal the state dependence of price discovery leadership.

15.4.1 Mean reversion in the funding rate

The most central institutional innovation of perpetual futures is the funding rate mechanism. Traditional futures contracts ensure that the futures price ultimately converges to the spot price by setting an expiration date and a delivery mechanism, but perpetual futures have no expiration date and therefore need an alternative means of anchoring the price. The funding rate was designed precisely for this: by periodically exchanging fees between holders of long and short positions, it continuously pulls the perpetual futures price toward the spot index price. When the perpetual futures price is above the spot index price, the funding rate is positive and longs pay shorts, which incentivizes traders to sell perpetual futures or buy spot, thereby narrowing the basis; conversely, when the perpetual futures price is below spot, the funding rate is negative and shorts pay longs, incentivizing the opposite arbitrage. On major exchanges, this settlement typically occurs once every 8 hours. Taking Binance as an example, the funding rate comprises two parts: an interest rate component reflecting the baseline borrowing cost, typically fixed at 0.01% per 8 hours; and a premium component computed as a time-weighted average of the deviation between the perpetual futures price and the mark price. The sum of the two is the funding rate for the period, subject to a cap and a floor; the cap is not a fixed percentage but is set according to the contract's maintenance margin rate (typically about $\pm 2\%$ per 8 hours for standard contracts, and dynamically set at 0.75 times the maintenance margin rate for high-leverage contracts). Exchanges differ in the details of the formula (the implementations of OKX and Bybit differ slightly from Binance's), but the core logic is the same: anchor the perpetual futures price to the spot index through economic incentives [21].

This design gives the funding rate a dual role. At the micro level, it is the direct financing cost of holding a leveraged long or short position, directly affecting traders' profit-and-loss calculations and position decisions. At the macro level, it is a sensitive

indicator of overall market sentiment and leverage demand. A high positive funding rate usually means the market is broadly bullish, with strong demand for leveraged longs and traders willing to pay a premium to maintain long positions; a persistently negative funding rate reflects market pessimism or strong hedging demand. The study by He, Manela, Ross, and von Wachter (2024) provides an important empirical foundation for understanding the information content of the funding rate [21]. They found that the price deviation between perpetual futures and spot can be largely explained by past return momentum: the momentum factor of past returns can explain more than 50% of the variation in the futures–spot spread. This means the funding rate is not a random variable but an information carrier that encodes the market's reaction to and expectations about recent price trends.

As we noted when discussing calendar effects in Section 15.2.3, any predictable price pattern constitutes a challenge to weak-form efficiency. The mean-reverting nature of the funding rate is precisely such a typical "institutional weak-form efficiency gap." When the funding rate deviates too far from its long-run mean, the economic incentive mechanism creates powerful countervailing forces. Take the positive extreme: when the funding rate is abnormally high, the "spot-perpetual arbitrage" strategy of shorting perpetual futures while buying an equal amount of spot becomes highly attractive. Traders executing this strategy can not only profit from the future convergence of the basis but also steadily collect the high funding fee while holding the position. This low-risk arbitrage opportunity attracts a large amount of capital, increases the supply of short positions, and thereby drives the funding rate down toward its mean. The scale of this arbitrage quantified by He et al. (2024) (the mean absolute deviation, the Sharpe ratios achievable by retail investors and market makers, and their year-over-year decay) was given in detail in Section 15.2.3 [21] and is not repeated here.

It must be emphasized that this weak-form efficiency gap does not arise from participants' irrational behavior or cognitive biases but is an inevitable result of the funding rate's institutional design. To achieve the core goal of price anchoring, the system must offer sufficiently high economic compensation to attract arbitrageurs, and this compensation itself manifests as price predictability. These seemingly generous arbitrage profits are not a riskless "free lunch," however: as described in Section 15.2.3, Zhivkov (2026) found that a high proportion of economically significant arbitrage spreads ultimately could not be fully executed because of margin requirements and forced liquidations [22]. This reveals an important friction in the mean reversion of the funding rate—arbitrage forces exist and are powerful, but they are constrained by capital limits, the margin system, and liquidity risk, so their correction speed is far slower than theory predicts.

Figure 15-11 uses four panels to show these key statistical properties of the funding rate.

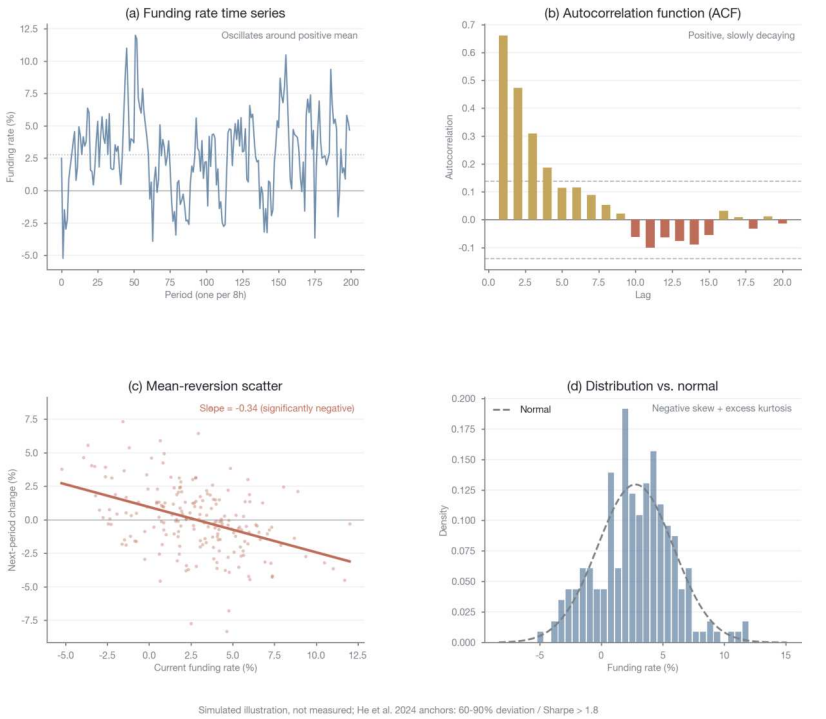


Figure 15-11. The information content and mean-reverting properties of the funding rate

The four panels of Figure 15-11 together characterize the key statistical features of the funding rate: the autocorrelation is significantly positive and decays slowly, indicating short-term persistence; the regression slope of the current rate on the next period's change is significantly negative, confirming mean reversion; and the probability distribution shows negative skew and excess kurtosis relative to the normal, revealing fat tails in which extreme negative-rate events (market panic) exceed what the normal distribution predicts. Within an Ornstein-Uhlenbeck process framework, this negative slope directly yields a half-life estimate: after the funding rate of BTC perpetual futures deviates from its mean, it typically reverts about half of the deviation within one to two settlement cycles. This half-life is not constant: in extreme bull markets, crowded longs slow the reversion of positive rates, and in panic selling the persistence of negative rates may also exceed the historical mean; this asymmetry has direct implications for the holding period and capital planning of an arbitrage strategy.

Taken together, the statistical properties of the funding rate (short-term persistence, long-run mean reversion, and a fat-tailed distribution) make it a predictable economic variable. This predictability directly violates the basic assumption of weak-form market efficiency and constitutes an intrinsic efficiency gap in the perpetual futures market, created by institutional design. This gap is not "free profit," however, but is protected by arbitrage friction, capital constraints, and execution risk. The tension between the absence of efficiency and the imperfection of arbitrage is precisely the key to understanding the institutional efficiency of the perpetual futures market.

15.4.2 Efficiency fluctuations within the settlement cycle

The mean-reverting nature of the funding rate describes the macro dynamics spanning multiple settlement cycles, but within each settlement cycle, the market microstructure likewise exhibits a finer pattern of efficiency fluctuation driven by institutional arrangements. Ruan and Streltsov (2022), in a study based on high-frequency order-book data from 2017 to 2023, found a striking phenomenon: within each 8-hour funding settlement cycle, the microstructure quality metrics of the spot market exhibit a significant "U-shaped" pattern [30]. This finding reveals an hourly efficiency cycle, which we call the pre-settlement effect.

Understanding the mechanism that generates this U-shaped pattern requires starting from traders' strategic behavior around funding settlement. At the moment of settlement, the exchange exchanges cash flows between all outstanding long and short positions according to the prevailing funding rate. If the expected funding rate is high (say 0.1%), then a trader holding a $1 million long position will pay a $1,000 fee at settlement. To avoid this cost, some long holders choose to close their positions before settlement and immediately reopen them afterward. At the same time, arbitrageurs may establish short positions before settlement to collect the funding fee and close them afterward. These strategic behaviors burst out around the moment of settlement, causing sharp changes in order flow and a temporary rise in the proportion of informed trading.

Market makers respond rationally to this foreseeable behavior pattern. Facing the rise in the proportion of informed trading and the intensified adverse selection risk around settlement, market makers strategically widen their bid-ask spreads and reduce quote depth to protect themselves from losses due to an information disadvantage. This behavior produces a clear cyclical pattern: within the window approaching settlement, bid-ask spreads widen, price impact increases, market liquidity deteriorates, transaction costs rise, and market efficiency declines accordingly. After settlement is complete, uncertainty is removed, strategic traders leave, market makers restore normal quotes, market liquidity improves, and efficiency recovers. As time proceeds to the midpoint of the next settlement cycle, the market is in its calmest state and efficiency reaches its peak. Thereafter, as the next settlement point approaches, the entire cycle plays out again.

Figure 15-12 shows the efficiency-fluctuation pattern of four key market-microstructure metrics within a typical 8-hour settlement cycle.

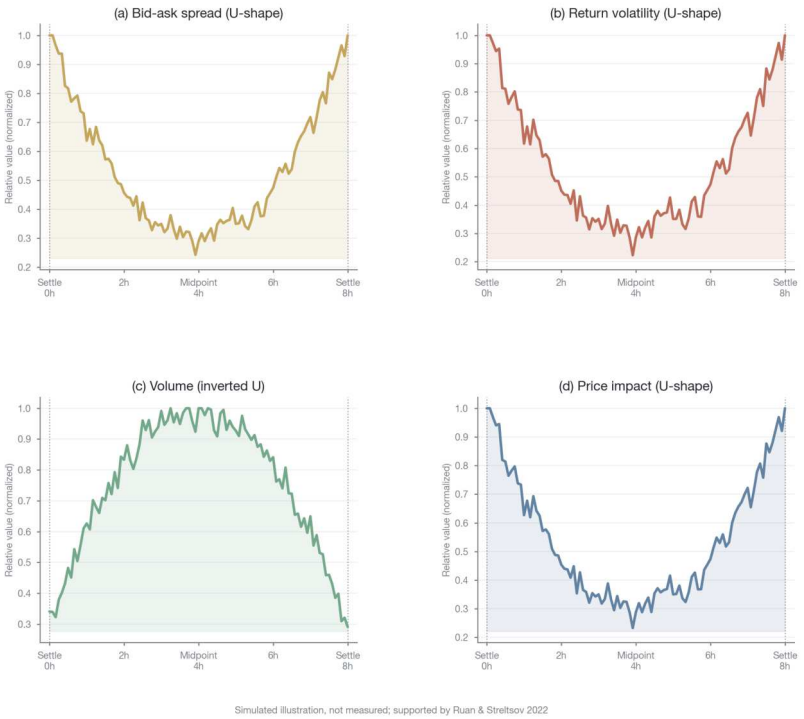


Figure 15-12. Market-efficiency fluctuations within the funding settlement cycle (the U-shaped pattern)

As the figure shows, the bid-ask spread, the price-impact coefficient, and return volatility all peak near the settlement points (0 and 8 hours) and fall to a trough at the midpoint of the cycle (around the 4-hour mark), forming a clear U-shape; trading volume, by contrast, exhibits a complementary inverted U-shape, most active in the middle of the cycle when the market is calmest and transaction costs are lowest. The quantitative analysis of Ruan and Streltsov (2022) shows that the higher the absolute value of the funding rate in a cycle, the more pronounced the increase in the bid-ask spread and price impact around settlement [30], indicating that the pre-settlement effect is not merely statistical but carries real economic costs that substantially affect the execution quality and strategy design of high-frequency traders.

The study by Ruan and Streltsov (2022) also revealed an important moderating factor: the larger the absolute value of the funding rate, the more pronounced this U-shaped pattern [30]. This finding is intuitively reasonable: a larger funding rate means a stronger incentive to close positions or arbitrage, leading to more intense trading around settlement and greater microstructure disturbance. The study also used the natural experiment of Huobi's termination of perpetual futures trading, and 95 staggered contract-listing events as an identification strategy, to further confirm the causal effect of perpetual futures on spot market quality: the introduction of perpetual futures increased spot trading volume but also widened quoted spreads, reflecting the increase in informed trading and market makers' compensation for adverse selection risk [30].

Notably, the magnitude of this microstructure-level efficiency fluctuation depends heav-

ily on the settlement frequency set by the exchange. As Chapter 10 noted in discussing the evolution of the crypto derivatives market, major centralized exchanges (such as Binance and OKX) have long used discrete settlement once every 8 hours as the most common default setting, although in recent years some exchanges have begun to introduce more flexible settlement-cycle options for specific contracts. This relatively long settlement cycle means that the funding fee accumulated in a single settlement is relatively high, giving traders a strong incentive to avoid it and producing the significant U-shaped fluctuation described above. As market competition has intensified and technology has advanced, however, many emerging platforms (such as the DEXs dYdX and Hyperliquid) have begun to adopt 1-hour or even continuous settlement mechanisms.

Higher settlement frequency fundamentally changes the microdynamics of the pre-settlement effect. When the settlement cycle is shortened from 8 hours to 1 hour, the funding fee for a single settlement is greatly diluted, and the transaction cost (fees and slippage) of frequently closing and reopening positions to avoid a tiny single fee far exceeds the funding fee saved. Strategic position-closing behavior therefore decreases significantly, the adverse selection risk faced by market makers falls, and the cyclical widening of the bid-ask spread weakens sharply as a result. This shows that by optimizing institutional parameters (such as raising the settlement frequency), a market platform can effectively smooth this rule-driven efficiency pulse and improve overall market microstructure quality. That said, the above inference that high-frequency settlement smooths the pre-settlement effect is currently based mainly on theoretical reasoning and lacks a systematic empirical comparison of microstructure metrics across platforms with different settlement frequencies, which constitutes an important direction for future research.

15.4.3 Mark price and oracle latency

The third key institutional design of perpetual futures concerns their valuation and liquidation mechanisms. In centralized exchanges and decentralized exchanges, this mechanism is implemented in different ways, but both introduce a specific form of information-reflection lag. This constitutes another institutional efficiency barrier and directly echoes the price-benchmark governance issues discussed in Chapter 14.

In centralized exchanges, to prevent unnecessary forced liquidations caused by the manipulation of a single exchange's price or by a flash crash, the valuation and liquidation judgment of a contract do not directly use that exchange's latest traded price but rather a computed mark price. Taking Binance as an example, the mark price is computed as the index price multiplied by an adjustment factor of one plus the funding basis, where the funding basis is estimated from the linear decay, over the current settlement cycle, of the previous cycle's funding rate. OKX's implementation differs slightly: its mark price takes the median of two values—the exponential moving average of "the index price plus the basis" and "the mid-price of the current best bid and ask." Although the specific formulas differ across exchanges, their common design philosophy is to create a more robust, harder-to-manipulate fair-value benchmark through multi-source data aggregation and basis adjustment. In most exchanges' implementations, the exponential-moving-average smoothing is applied mainly to the computation of the

premium component of the funding rate rather than directly to the mark price itself—a distinction that matters for understanding the mark price's response speed under extreme conditions.

This robustness comes at the cost of information immediacy, however. When the market price jumps sharply and rapidly (for example, when major news is released or a large liquidation cascade is triggered), the mark price, because of its inherent multi-source aggregation and smoothing mechanism, moves with a significant lag behind the spot price and the latest traded price. This lag may be only a few seconds under normal market conditions but can extend to tens of seconds or longer under extreme volatility. The consequences of the lag are double-edged: on the one hand, it protects traders from improper liquidation caused by instantaneous price manipulation; on the other hand, it means that when the market moves rapidly, the mark price fails to reflect true market conditions in time, potentially allowing high-risk positions that should have been liquidated to persist, or preventing arbitrageurs from exploiting the true price deviation in time.

In the perpetual futures protocols of decentralized exchanges, the problem of information latency is more acute. Because on-chain protocols cannot directly access real-time off-chain market data, they rely on oracle networks to obtain external price information. Oracle networks (such as Chainlink and Pyth) aggregate prices from multiple off-chain data sources and submit the updated prices to the blockchain. This process introduces multiple layers of latency. At the physical level, the whole process—oracle nodes obtaining prices from data sources, reaching consensus among nodes, broadcasting the transaction to the blockchain, and waiting for it to be packed and confirmed—has inherent network and consensus latency. At the economic level, to save on-chain transaction costs, oracle updates are not continuous but discrete, triggered only when the price change exceeds a preset deviation threshold (such as 0.5%) or a time interval is reached. This means that between two updates, the price used by the on-chain protocol is "stale," with a deviation from the true market price that accumulates over time. Zhivkov's (2026) study quantified the information gap between CEXs and DEXs: centralized exchanges have 61% higher integration in price discovery than decentralized exchanges, and all statistically significant information flows run from CEXs to DEXs, with no reverse causality [22].

Table 15-4 compares the key differences between the CEX mark-price mechanism and the DEX oracle mechanism in reflecting information.

Dimension	CEX mark price	DEX oracle price
Data source	Weighted average across multiple spot exchanges	Off-chain data aggregated by an oracle network
Update frequency	Near real-time (millisecond-level)	Varies widely by architecture: push-based oracles (e.g., Chainlink) update at minute-level, pull-based oracles (e.g., Pyth) at sub-second level, and native order books (e.g., Hyperliquid) in near real-time
Smoothing mechanism	EMA smoothing, moving average of the basis	Median aggregation, deviation-threshold triggering
Typical latency	Normally <1 second; seconds to tens of seconds under extreme conditions	Normally 3–15 seconds; up to minute-level under extreme conditions
Manipulation resistance	High (multi-source aggregation, though it depends on the number of constituent exchanges, weight concentration, and outlier-detection mechanisms; see Chapter 14)	Medium to high (depends on oracle design, number of data sources, and consensus mechanism)
Information completeness	High (close to the real-time state of the market)	Medium (discrete snapshots that may miss rapid moves)
Cost structure	Computed internally by the exchange, at no additional cost	On-chain transaction costs limit update frequency

Table 15-4. Comparison of the information-reflection mechanisms of the CEX mark price and the DEX oracle (Data source: compiled by the author)

Figure 15-13 offers an intuitive comparison of how these two mechanisms reflect information during a sharp price-jump event.

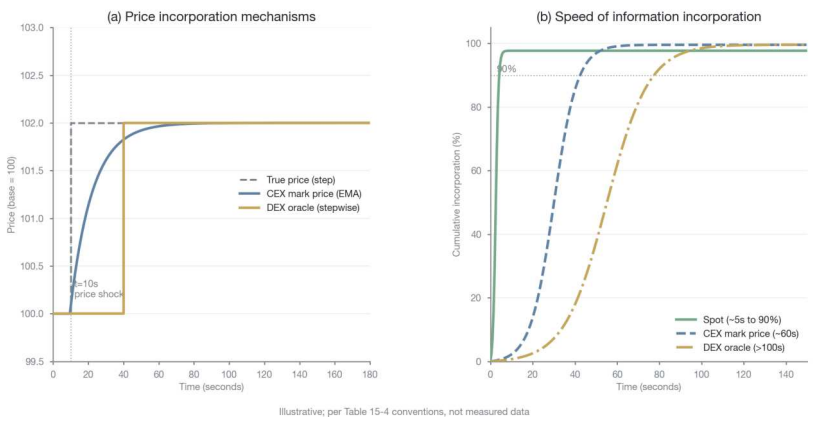


Figure 15-13. The effect of mark-price smoothing and discrete oracle updates on information-reflection speed

Panel (a) shows that when the spot price jumps instantaneously at the 10th second, the CEX mark price catches up slowly along a gentle S-curve because of EMA smoothing, forming a "mark-price lag," while the DEX oracle price updates in discrete steps, with

several seconds of "information vacuum" between updates. Panel (b) quantifies this gap in terms of cumulative information incorporation: to reach 90% incorporation, spot needs only about 5 seconds, while under an extreme flash crash the CEX mark price needs about 60 seconds and the DEX oracle may need more than 100 seconds (in day-to-day volatility, the CEX mark price's typical tracking lag is on the order of 5 to 15 seconds). The time gap between the spot price and the oracle price is precisely the "efficiency gap" created by the market's institutions themselves.

The real-world impact of this institutional latency is far from an abstract theoretical discussion. During sharp market volatility, the lag of the mark price can cause the liquidation mechanism to fail: high-leverage positions that should have been liquidated in time survive because the mark price has not yet caught up, accumulate larger losses, and finally trigger a larger-scale cascade of liquidations when the mark price catches up. For DEXs, oracle latency also creates an arbitrage opportunity known as oracle extractable value (OEV), in which informed traders exploit the known deviation between the on-chain oracle price and the true market price to execute profitable trades on the DEX—essentially a systematic extraction of value from liquidity providers. The scale of OEV is not trivial; industry estimates suggest it can reach the order of hundreds of millions of dollars per year in on-chain perpetual futures, directly eroding the returns of liquidity providers and distorting market-making incentives. Several protocols are currently developing auction mechanisms to reclaim this value—for example, by handing the priority ordering of oracle-update rights to a competitive auction, so that the extracted value can flow partly back to the protocol or to liquidity providers. Moreover, unlike front-running in traditional finance, the source of OEV lies in the discreteness and predictability of on-chain price information, which means that as long as the oracle-update mechanism retains its discrete character, this structural value extraction is hard to eliminate entirely. Beyond update latency, oracle networks also face accuracy risks, including manipulating the oracle's input price through flash loan attacks on low-liquidity exchanges, common-mode failure risk when oracle nodes use the same infrastructure, and blockchain congestion that prevents update transactions from being confirmed in time. These risks can cause on-chain protocols to execute liquidations or settlements based on erroneous price signals, damaging market efficiency far more severely than mere latency. Together, these phenomena show that the price-benchmark design of perpetual futures, in pursuing robustness and manipulation resistance, inevitably introduces a loss of informational efficiency. How to strike the optimal balance between these two goals is precisely the core issue of price-benchmark governance discussed in Chapter 14.

15.4.4 Price discovery leadership

Having dissected the institutional efficiency barriers within perpetual futures, a naturally following question is: what role does this vast, active derivatives market play in the price discovery of the entire crypto ecosystem? Is it merely a passive follower of the spot market, or has it become the main force leading price movements? Analysis of this lead-lag relationship is the final step in assessing the informational efficiency of the perpetual futures market.

In financial econometrics, there is a mature methodological system for quantifying the contribution of different markets to price discovery. The information share model proposed by Hasbrouck (1995) is one of the most widely applied tools [31]. The model is based on a vector error correction model, and its core logic is this: if two markets trade the same underlying asset (or an economic equivalent), their prices should converge in the long run (that is, a cointegration relationship exists), but they may deviate in the short run. The information share measures the proportion of each market's contribution to the permanent movement of this common efficient price. If a market's information share is 70%, it means that market contributes 70% of the variance of the permanent movement of the efficient price—that is, it plays a dominant role in price discovery. The complementary Gonzalo-Granger component share model measures each market's contribution to the long-run equilibrium price from the perspective of error correction. The Hasbrouck information share is somewhat sensitive to the ordering of variables in the vector autoregression model, and what is usually reported is the average of the upper and lower bounds. In crypto markets, because the price-update frequency between perpetual futures and spot is often asymmetric, and because the data frequency used in different studies (ranging from tick-by-tick to minute-level) also affects the estimates, the information-share values reported in different studies should not be compared simply side by side.

A large body of academic research has applied these methods to the spot–futures price discovery analysis of crypto markets and reached a fairly consistent conclusion: the futures market, and especially the perpetual futures market, dominates price discovery for major assets such as Bitcoin and Ethereum. The early study by Kapar and Olmo (2019), using the Hasbrouck information share method, found that the Bitcoin futures market's information share was as high as 88.7%, while the spot market's was only 11.3% [32]. The dedicated study of BitMEX perpetual futures by Alexander, Choi, Park, and Sohn (2020) confirmed that BitMEX's derivative price movements led those of major spot exchanges, with a positive net spillover from derivatives to spot [33]. The most recent study by Robertson and Zhang (2025) further confirmed the leading role of the futures market in Bitcoin price formation and noted that trade size is the key factor determining the contribution to price discovery [34].

The dominance of perpetual futures in price discovery is not accidental but is determined by their institutional features. Compared with the spot market, the perpetual futures market offers lower transaction costs (no need to actually deliver the underlying asset), higher leverage (typically up to 20x to 125x), and a more convenient short-selling mechanism (no need to borrow the asset). These features allow informed traders to inject their views into the market at lower capital cost and higher speed. A trader wishing to go long Bitcoin on the basis of a piece of news need only deposit a small amount of margin to establish a large position in the perpetual futures market, whereas in the spot market they would need to mobilize the full amount of capital. This leverage effect makes perpetual futures the venue of choice for informed traders, giving them the first-mover advantage in price discovery.

This leading position is not fixed, however, but exhibits strong state dependence: as the adaptive markets hypothesis holds, the market's efficiency structure changes dynamically with the evolution of participants, technology, and the external environment.

Figure 15-14, based on multiple studies, depicts the evolution of the information share of perpetual futures and the spot market in price discovery from 2019 to 2025.

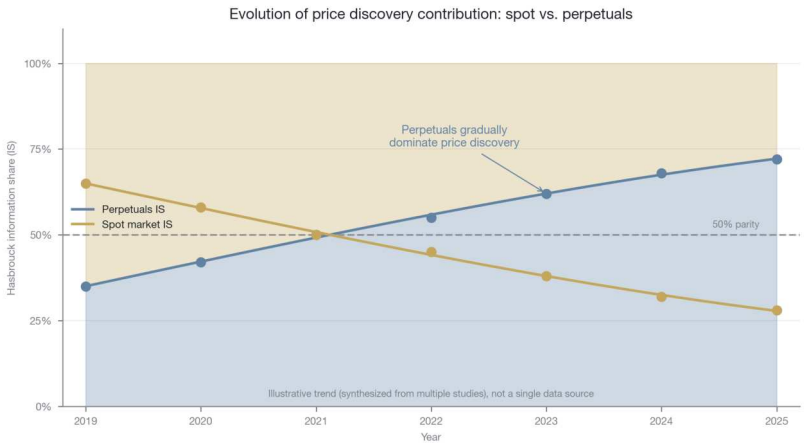


Figure 15-14. The price discovery contribution of spot and perpetual futures: the evolution of the Hasbrouck information share

Figure 15-14 synthesizes multiple studies (Kapar and Olmo (2019) [32], Alexander et al. (2020) [33], Robertson and Zhang (2025) [34]) to depict the evolution of the Hasbrouck information share of perpetual futures and the spot market from 2019 to 2025 (the moving average is an illustrative trend reflecting the direction reported in the literature rather than the precise calculation of a single continuous data source). The information share of perpetual futures is above the 50% parity line most of the time and has risen over time from about one-third early on to the majority share in the mature period (the specific values differ significantly by trading venue, estimation method, and sample period, and the literature is not unanimous), reflecting the deepening of liquidity and the increase in institutional participation; but during major market shocks, this dominance undergoes a brief and significant reversal.

Under extreme stress, the institutional advantage of perpetual futures often quickly turns into institutional risk. As Chapter 11 analyzed in detail in its discussion of liquidation mechanisms, when the market encounters a black-swan event that triggers a sharp decline, the high-leverage nature of perpetual futures triggers large-scale forced liquidations. These forced liquidations are essentially market sell orders executed at any cost, which push prices lower still, trigger more liquidations, and form a destructive liquidation cascade. In this extreme environment, the liquidity of the perpetual futures market can evaporate instantly, counterparty risk rises sharply, and the once-leading price discovery mechanism is nearly paralyzed. At such times, traders tend to return to the most basic spot market, which carries no liquidation risk or leverage-collapse risk. During a crisis, therefore, the spot market often temporarily recaptures the leadership of price discovery.

In the COVID crash of March 2020, the information share of perpetual futures fell sharply. The extreme market panic of this period caused large-scale cascading liquidations, and the spot market briefly dominated price discovery. A similar pattern played

out again in the FTX collapse of November 2022. As one of the largest perpetual futures trading platforms at the time, FTX's collapse directly undermined the overall credibility and liquidity of the perpetual futures market. De Blasis and Webb (2022) likewise observed that during market dislocations, significant spillover effects and arbitrage opportunities emerge between different types of futures contracts [35]. The approval of spot Bitcoin ETFs in January 2024 represented a different type of shock: this event directly benefited the spot market, and a large amount of capital poured into the spot market through the ETF channel, temporarily raising the information contribution of spot trading.

Taken together, the perpetual futures market plays a central but not absolute role in the price discovery of digital assets. Although it contains predictable efficiency barriers triggered by institutional designs such as the funding settlement cycle, mark-price smoothing, and oracle latency, its efficient trading mechanism and enormous liquidity make it the main venue for information to be incorporated into prices under the vast majority of market conditions. This reveals a striking paradox of market efficiency: a market that contains multiple institutional "inefficiencies" at the micro level nonetheless becomes the main price discovery engine of the entire ecosystem at the macro level. Understanding this multilevel, seemingly contradictory efficiency structure is the key to grasping the operating logic of digital asset markets, and it lays the necessary analytical foundation for the next section's in-depth analysis of the FTX collapse, in which the institutional efficiency of the perpetual futures market underwent an extreme stress test.

Efficiency Spectrum Locator Card · Section 15.4

Dimension	Positioning	Quantitative proxy metric
Time	Predictable microstructure fluctuations around funding settlement	U-shaped cycle in bid-ask spread and volatility
Asset	Major-asset (BTC/ETH) perpetual futures dominate price discovery	Information share of perpetual futures (typically 70%–80%)
Information type	Institutional signals (such as extreme funding-rate values) have weak-form predictability	Funding-rate autocorrelation, arbitrage Sharpe ratio
Market state	Under extreme stress, perpetual futures efficiency falls sharply and spot recaptures leadership	Sharp drop in the perpetual futures information share during crises
Market tier	In the normal state, perpetual futures lead spot but are subject to mark-price latency	Reflection-time gap between the oracle/mark price and the spot price

One-sentence takeaway: The institutional design of perpetual futures (funding rate, mark price, liquidation mechanism) creates predictable weak-form efficiency gaps, and although perpetual futures dominate price discovery in the normal state, their institutional advantage quickly flips into liquidity risk under extreme market conditions.

15.5 The FTX collapse: a case study

The preceding sections provided theoretical tools and empirical tests for assessing the informational efficiency of digital asset markets, and the collapse of the FTX exchange in November 2022 provides an ideal ultimate stress test for these tools and theories. FTX's rapid collapse was not only one of the most significant bankruptcies in cryptocurrency history (its founder, Sam Bankman-Fried, was ultimately sentenced to 25 years in prison for fraud) but also a comprehensive test of the multidimensionality, complexity, and state dependence of market efficiency. It clearly revealed the market's starkly different efficiency performance in processing different types of information: in processing simple, observable on-chain information, the market exhibited extremely high semi-strong-form efficiency; in the face of complex, hidden structural fraud, the market exhibited long-lasting, significant inefficiency; and in the process of crisis transmission, the market exhibited a destructive "over-efficiency"—an irrational contagion effect.

This section dissects the event in depth by reconstructing the key timeline and analyzing it systematically along three core dimensions of the efficiency spectrum (the high efficiency of on-chain information, the low efficiency of complex information, and

the over-efficiency of the contagion effect). The earlier sections established the five-dimensional analytical matrix of the efficiency spectrum and separately conducted empirical tests of weak-form efficiency, semi-strong-form efficiency, and the institutional efficiency of perpetual futures. The FTX event happens to span multiple quadrants of these dimensions, providing an ideal opportunity to apply all the analytical tools together. Just as the adaptive markets hypothesis predicts, market efficiency collapses nonlinearly and sharply under extreme stress, and the FTX event is the most representative validation of this prediction. Through an in-depth dissection of this case, we ultimately demonstrate that informational efficiency is not a black-and-white binary state but a continuous spectrum dependent on information type, market state, asset characteristics, and time scale. The FTX event occurred in 2022, after the crypto market had already experienced a series of major shocks such as the Terra/LUNA collapse and the bankruptcy of Three Arrows Capital, so the market should have learned from those earlier lessons and remained more vigilant against similar structural risks. Yet the collapse of FTX shows that the market's process of learning and adaptation is far slower and more uneven than theory predicts, and this observation is itself an important test of the concept of "adaptation" in AMH.

15.5.1 Reconstructing the event timeline

FTX's collapse did not happen in an instant but was a continuous process of information gradually fermenting, market sentiment shifting sharply, and finally liquidity drying up and trust collapsing. Understanding this timeline is essential for analyzing market efficiency, because it reveals how different types of information were digested (or ignored) by the market at each stage. We divide this process into four key stages: the latent period, the information diffusion period, the rapid collapse period, and the post-bankruptcy period. Figure 15-15 uses the changes in FTT's price and volume to present intuitively the rhythm and intensity of the market's reaction during this process.

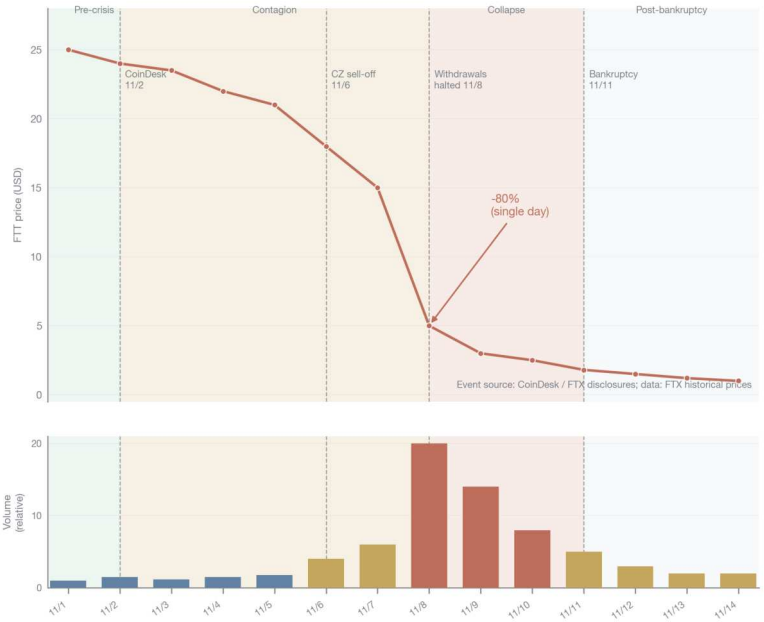


Figure 15-15. The FTX collapse timeline and the FTT price

Figure 15-15 shows the daily price (upper panel) and volume (lower panel) of FTT from November 1 to 14, 2022, with vertical dashed lines marking the four key events that triggered violent reactions, and background shading dividing the crisis into four stages: pre-crisis (green), the information diffusion period (yellow), the rapid collapse period (red), and post-bankruptcy (gray). Arrows in the price panel mark FTT's single-day decline of more than 80% on November 8, while the volume panel shows the abnormal amplification of trading activity at the height of the crisis.

The first stage, the latent period before the crisis (before November 2), the market appeared calm on the surface. FTT's price was stable around $25, average daily volume was at normal levels, and implied volatility showed no anomaly. Behind the surface data, however, the circular leverage and balance-sheet problems centered on FTT that had long existed between FTX and its affiliated trading firm, Alameda Research, had already planted the seeds of the collapse. From the perspective of the efficiency spectrum, the market's performance in this stage is precisely a typical manifestation of the low efficiency of complex information: structural risk already objectively existed, and warning signals were not lacking in the on-chain data, but the market failed to integrate these scattered signals into a coherent risk judgment. From the perspective of strong-form efficiency, the insiders of FTX and Alameda undoubtedly held private information about the true state of the balance sheet, and this information went unpriced by the external market for years, constituting an extreme case of the failure of strong-form efficiency. As the study by Vidal-Tomás et al. (2023) summarizes, FTX and Alameda used the vast quantity of FTT tokens they controlled as a leverage tool, creating a financial conglomerate that looked enormous but was extremely fragile, and this fragility was almost unpriced by the market before the collapse [36]. Notably, this

prolonged collective oversight was not because the information was completely unavailable (the public nature of the blockchain ensured that fund flows were traceable) but because the cognitive cost and analytical capability required to turn this raw data into a risk judgment far exceeded the capacity of most market participants.

Within this latent period, if we introduce the perpetual futures perspective, we find that the derivatives market likewise failed to price this structural risk in advance. The funding rate of FTT perpetual futures stayed within a normal fluctuation range near the zero axis before November 2, showing no systematic short-side tilt; the basis between perpetual futures and spot also stayed within a reasonable arbitrage range, and open interest remained stable. This shows that even the perpetual futures market participants, usually thought to have an information advantage and to be more adept at price discovery, failed to see through FTT's "market-cap illusion."

The second stage, the information diffusion period (November 2 to November 6), began with a key media report. On November 2, 2022, the crypto media outlet CoinDesk published an investigative report disclosing details of Alameda Research's balance sheet, noting that of its $14.6 billion in assets, about $5.8 billion was in FTT tokens (of which about $3.66 billion was unlocked FTT and about $2.16 billion was FTT posted as collateral) [37]. This report exposed for the first time the blurred boundary between the two companies and their excessive reliance on their own issued token. From the perspective of informational efficiency, the CoinDesk report played a key role: it acted as an information intermediary, turning high-processing-cost raw on-chain data and financial documents into a simple, easily transmissible narrative signal. This is precisely a direct manifestation of the information-processing-cost theory discussed earlier in the semi-strong-form efficiency tests. Between information "being publicly available" and being "widely understood" by the market lies a gulf constituted by processing cost, and media reporting is precisely the bridge across that gulf.

The market's initial reaction to the CoinDesk report was relatively mild, with FTT's price falling only about 3% to near $24. This sluggish reaction is itself worth pondering: it shows that even after the information had been processed and disseminated by the media, the market still needed time to digest its meaning and assess its actual impact on FTX's solvency. During these few days of the information diffusion period, discussion on social media gradually heated up, and opinion leaders and analysts in the crypto community began to interpret Alameda's balance sheet more deeply. The market's overall reaction remained gradual rather than explosive, however, consistent with the reaction pattern to complex crypto-native information observed earlier in the semi-strong-form efficiency tests: information needs to go through multiple rounds of dissemination and interpretation before market consensus can gradually form.

The key turning point came on November 6. Changpeng Zhao, the chief executive officer of Binance—the world's largest cryptocurrency exchange—publicly announced on a social media platform that, out of risk-management considerations, Binance would liquidate all of its remaining FTT tokens on its books (about 23 million, worth about $530 million at the time) [38]. This public statement from a core market participant carried far more information than any media report, because it not only confirmed the existence of the risk but also directly foreshadowed the imminent arrival of a massive

wave of selling. Zhao's statement turned an abstract discussion about balance-sheet structure into a concrete threat about immediate market supply and demand, greatly reducing the information processing cost so that even the least sophisticated market participants could understand its meaning. From the perspective of semi-strong-form efficiency, the effect of this statement was similar to the standardized binary signal discussed earlier (such as a regulatory approval or rejection); its information content was clear and its interpretation cost low, so the market reacted quickly and violently.

Accompanying Zhao's statement, the perpetual futures market was the first to issue a clear warning signal. On the evening of November 6, the FTT perpetual futures funding rate on major exchanges began to turn sharply negative, reflecting the rapid massing of short-side forces; at the same time, the perpetual futures price began to fall significantly below the spot price, and the basis went into a deep inversion. Over the following 48 hours, as panic spread, the open interest of FTT perpetual futures began to fall sharply, marking the panic-driven closing or forced liquidation of a large number of leveraged positions. These synchronous and even slightly leading violent reactions of the perpetual futures market provide a microstructure perspective on the crisis that the spot market cannot offer.

The third stage, the rapid collapse period (November 6 to November 11), the market's efficiency entered the "decay state" defined by the efficiency spectrum framework—a nonlinear efficiency collapse driven by a negative feedback loop. Zhao's statement triggered a wave of FTT selling, and its price began to fall significantly on November 7, from about \$23 to about \$21. The true disaster occurred on November 8, however. That day, the FTT price plunged from about \$22 to less than \$5 within 24 hours, a decline of more than 80%, and volume surged to dozens of times its normal level. The speed and magnitude of this price collapse far exceeded any reasonable adjustment range based on fundamental analysis; what it reflected was a liquidity-driven negative-feedback spiral: the price decline triggered margin calls on loans collateralized by FTT, forced liquidations pushed the price down further, and the further price decline triggered still more liquidations, a process that self-reinforced in the absence of an external liquidity injection until the price approached zero. This mechanism has dynamics similar to the forced-liquidation cascade in the perpetual futures market discussed earlier, except that in the FTT case the trigger was not insufficient margin but the collapse of collateral value.

At the same time, a large number of users rushed to the FTX exchange to withdraw funds, triggering a classic bank run. According to bankruptcy filings later disclosed, FTX faced about \$6 billion in withdrawal requests between November 6 and 8, far exceeding its available liquid assets [39]. Unlike a traditional bank run, withdrawals from a crypto exchange face distinctive operational bottlenecks: congestion on the blockchain network and surging gas fees slow the on-chain confirmation of withdrawals, and the exchange's cold-wallet–hot-wallet architecture means that large-scale withdrawals require retrieving assets from cold storage, a process with inherent operational latency. More crucially, the withdrawal delays were themselves interpreted by market participants as a signal of insufficient solvency, further accelerating the flood of withdrawal requests and forming an operational-level positive feedback loop. Facing dried-up liquidity, FTX suspended user withdrawals on November 8. Although technically intended to protect

the remaining assets, this decision released the most critical negative signal at the informational level: it confirmed to the market that FTX was already effectively insolvent. In traditional banking, suspending withdrawals is usually the harbinger of a bank failure; in the context of a crypto exchange, this signal has an equally serious effect. Notably, Binance had that day announced that it had signed a non-binding letter of intent to acquire FTX in order to resolve its liquidity crisis. This news briefly stabilized market sentiment, and the FTT price rebounded modestly after it broke. Just one day later (November 9), however, Binance announced it was abandoning the acquisition, citing that after preliminary due diligence it had found FTX's problems "beyond our control or ability to help" [40]. The disappearance of this acquisition possibility marked the loss of FTX's last avenue of external rescue and the market's transition from the "decay state" into a complete collapse of trust.

The fourth stage, the post-bankruptcy period (after November 11), the situation became largely clear. On November 11, 2022, FTX, Alameda Research, and about 130 affiliated companies formally filed for Chapter 11 bankruptcy protection in the United States. In his sworn declaration, the new chief executive officer, John J. Ray III, wrote that this was the "most serious corporate governance failure" he had seen in his career, its chaos exceeding even the Enron bankruptcy he had once handled [39]. The price of the FTT token soon fell to the $1-to-$2 range and long lingered there; its market capitalization (by circulating supply) shrank by more than 90%, from about $3 billion before the collapse to roughly $200 million. This unprecedented collapse in crypto history took less than 10 days from the disclosure of the key information to the final bankruptcy, and this speed is itself strong proof of how fast information is transmitted in digital asset markets under crisis conditions; it also reveals how severe the fragility of crypto markets is in the absence of the safety nets of the traditional financial system (such as the central bank's lender-of-last-resort function and deposit insurance).

Reviewing this complete timeline from the time dimension of the efficiency spectrum, a striking pattern emerges: the market's reaction to FTX risk exhibited an extreme nonlinearity. During the years-long latent period, although on-chain data continually issued warning signals, the market's reaction was nearly zero; during the mere few days of the information diffusion period, the market's reaction warmed gradually from mild; and during the final collapse period, the market's reaction was explosive and irreversible. This sharp leap from "zero reaction" to "overreaction" exhibits the characteristic features of a typical nonlinear transition: the system maintains surface stability under long-accumulated pressure and undergoes a sharp state transition once a critical threshold is breached. In financial markets, the trigger of such a nonlinear transition is often not a change in the information itself but the critical moment when information turns from "private knowledge" into "common knowledge." Zhao's tweet played precisely the role of triggering this consensus shift.

15.5.2 High efficiency in processing on-chain information

Reviewing the FTX event from the perspective of informational efficiency, the first conclusion to emerge is seemingly paradoxical but revealing: in processing simple, directly observable on-chain information, the market exhibited extremely high semi-strong-form

efficiency. This "high efficiency" needs to be defined precisely. It does not mean that the market reacted to all on-chain signals immediately; rather, it means that once the on-chain evidence of fund flows between Alameda and FTX was turned by professional information intermediaries into an easily understood signal, the market could rapidly and violently price it in. This distinction matters because it corresponds precisely to the information-processing-cost framework established earlier: the bottleneck of efficiency lies not in the availability of information but in its interpretability.

Blockchain transparency meant that Alameda Research's fund movements were, to a large extent, an "open secret." Professional on-chain analysts and data platforms, such as Nansen and Arkham Intelligence, could monitor the fund flows between them in real time by labeling and tracking known Alameda and FTX wallet addresses. In fact, months or even years before the CoinDesk report, on-chain data had already clearly revealed several key facts. Nansen's post-event analysis documented these on-chain traces in detail [41]. Of FTT's total token supply of about 328 million, FTX and Alameda themselves controlled as much as roughly 80%, meaning that its true external free float was extremely low. More telling still, during the Terra/LUNA and Three Arrows Capital collapses of mid-2022, on-chain data showed large-scale fund flows between Alameda and FTX: Alameda transferred about $4 billion worth of FTT tokens to FTX, while FTX provided Alameda with loans sourced from user assets. These transactions were fully public on the blockchain, and any observer with basic on-chain analysis tools could trace them. In addition, in the weeks before the crisis erupted, on-chain data also showed an anomalous outflow pattern from Alameda's wallets, including the transfer of large amounts of stablecoins and ETH to multiple exchanges, behavior later interpreted as Alameda's attempt to move assets and hedge risk before the crisis fully broke.

Nonetheless, although this on-chain information was publicly available, it was raw and unprocessed, mixed into the massive daily data stream of millions of transactions on blockchains such as Ethereum and Solana. As discussed earlier with respect to semi-strong-form efficiency, the high information processing cost constitutes a structural information asymmetry among market participants. Only a few informed traders with professional tools and deep domain knowledge (including researchers at on-chain analysis firms, professional quant funds, and seasoned investors deeply involved in the crypto ecosystem) could turn this raw data into tradable insight. For the vast majority of market participants, interpreting this on-chain data required understanding Ethereum's address-labeling system, the semantics of token transfers, the logic of smart-contract interactions, and the details of FTT's token-economic model, and this knowledge constituted a very high information barrier. Therefore, before the CoinDesk report of November 2, although on-chain alarms were frequent, the market had not formed common knowledge in the economic sense—the state in which all participants know a piece of information and all participants know that other participants also know it. As a result, prices had not reacted fully to these scattered signals.

As described in Section 15.5.1, the CoinDesk report and Zhao's liquidation statement successively translated high-processing-cost on-chain information into low-cost common knowledge, a process essentially of socializing the information processing cost: the media and analysts bore the cost of turning raw data into an understandable signal, so

that ordinary investors could obtain the key information without conducting on-chain analysis themselves. And Zhao's public liquidation as a market giant holding a vast FTT position, from a game-theoretic perspective, turned "private information" about FTX risk into "common knowledge"—every participant knew that others had also seen the message and would act on it.

From this angle, the FTX event strongly corroborates the finding of Azar et al. (2025) about the efficiency loss caused by information processing cost [5]. By analyzing cases of DeFi protocols being hacked, that study found that about 36% of the price decline occurred within the time window between the on-chain recording of the attack and its widespread dissemination on social media, meaning that informed traders who could directly interpret on-chain data acted first before the information became common knowledge. The FTX event presented a similar but more extreme pattern: anomalous on-chain signals existed for months, but only after the media reports turned them into a low-cost signal did the market react violently. Once information crossed the barrier of processing cost, the market's reaction speed was extremely fast: from Zhao's tweet on November 6 to the 80% collapse of the FTT price took less than 48 hours. This shows that along dimensions with high information symmetry, crypto markets can indeed conduct price discovery efficiently. The problem is not that the market lacks the capacity to react but that there is a structural bottleneck in the process of turning raw data into tradable signals.

This finding has important theoretical significance for understanding the semi-strong-form efficiency of crypto markets. The traditional EMH framework assumes that information is incorporated into prices immediately once it is "public," but the FTX event shows that in crypto markets "public" exists at multiple levels: the technical-level disclosure of on-chain data, the narrative-level disclosure of media reporting, and the consensus-level disclosure formed by public statements from core participants. The level of market efficiency depends largely on which "public" level the information is at. This insight not only enriches our understanding of semi-strong-form efficiency but also provides an important analytical perspective for the later discussion of efficiency evolution: as on-chain analysis tools spread and AI technology is applied, information processing cost is falling continuously—does this mean market efficiency will rise accordingly?

15.5.3 The market-cap illusion of FTT

The other side of the FTX event, however, reveals the market's extreme inefficiency in processing complex, systematic fraud information. If the market's reaction to on-chain fund flows was "efficient," then its long-term mispricing of the intrinsic value of the FTT token was a massive efficiency failure lasting years. The core of this inefficiency stems from a carefully constructed market-cap illusion.

The market-cap illusion refers to the vast difference between an asset's reported market capitalization (usually derived by multiplying the last traded price by the total supply or reported circulating supply) and its true, realizable liquid value. Similar problems exist in traditional financial markets (for example, large blocks of restricted shares held by controlling shareholders make the market cap calculated on total shares overstate

the company's true tradable value), but traditional markets have largely mitigated this problem through mandatory position-disclosure requirements, independent audits, and float-certification mechanisms. In crypto markets, the absence of these institutional safeguards amplifies the market-cap illusion to the extreme. In the FTT case, this illusion reached an extreme degree. As shown in Figure 15-16, FTT's total supply was about 328 million tokens, but before the collapse the number of tokens directly or indirectly controlled by FTX and Alameda accounted for nearly 80% of the total, with about another 10% locked or in a vesting period. This meant that the tokens truly available for external investors to trade freely may have accounted for only 10% of the total, or even less.

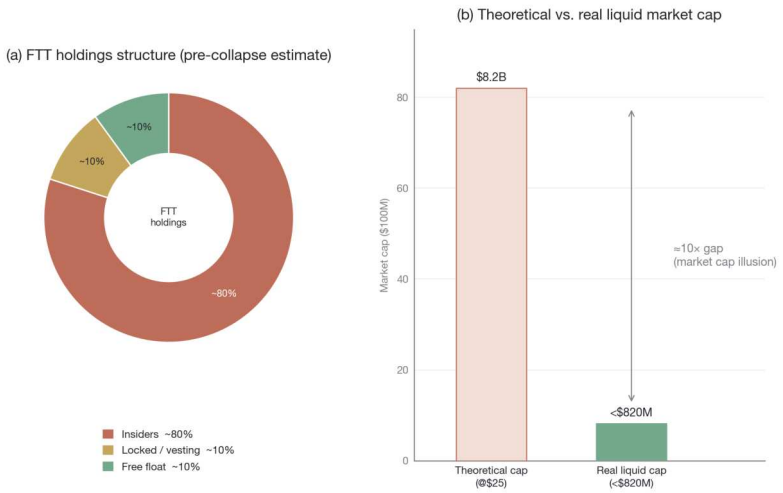


Figure 15-16. The "market-cap illusion" of FTT: the structural imbalance between free float and locked supply

Figure 15-16 uses two panels to reveal FTT's market-cap illusion: the left donut chart shows that before the collapse FTX and Alameda together controlled about 80% of the supply, about 10% was locked or in a vesting period, and the external free float was only about 10%; the right bar chart quantifies the resulting market-cap difference—at an FTT price of $25, the theoretical market cap based on total supply was about $8.2 billion, while the "true liquid market cap" based on the true free float was less than $820 million, a nearly tenfold difference.

Under this highly centralized supply structure, FTT's price discovery mechanism was severely distorted. Because the vast majority of tokens were locked up by insiders, the liquidity of the external trading market was extremely low and the depth of the order book extremely shallow. This meant that FTX and Alameda could maintain or even push up the FTT price with relatively little capital, thereby creating an inflated "market cap." Before the collapse, FTT's reported market cap was as high as several billion dollars, but this was not a fair value determined by a fully competitive market; it was a number easily manipulated by insiders in a low-liquidity market. This manipulation could persist partly because FTX's reputation as a major exchange provided a kind

of implicit credit endorsement for FTT, and investors tended to believe that a token issued by an "industry leader" would not be a hollow shell. FTX's founder was widely regarded before the collapse as a benchmark figure of the crypto industry, appearing frequently at congressional hearings and in mainstream media, and this public image further reinforced investor trust and made them turn a blind eye to potential risk signals.

More important still, this greatly inflated asset was used as the core collateral for circular leverage. FTX and Alameda used the FTT they held as collateral to obtain billions of dollars in loans of real assets—mainly US dollar stablecoins and other major cryptocurrencies—from other lenders (including institutions such as Genesis and BlockFi, and even including FTX's own user funds). This was a classic piece of financial engineering: exchanging a token created at almost zero cost, whose value was highly manipulated, for assets with real value. This circular-collateralization structure can be summarized as a self-referential value loop: the FTT price supported Alameda's balance sheet, Alameda's "solvency" supported the market's confidence in FTX, and FTX's reputation in turn supported the FTT price. As long as any link in this loop was broken, the entire system would trigger a chain reaction and collapse rapidly. Vidal-Tomás et al. (2023) summarized this as a fragile vicious cycle dependent on external shocks [36]. The post-collapse bankruptcy investigation revealed that FTX had misappropriated at least $8 billion in customer funds, most of which was transferred to Alameda for high-risk investments and to repay loans [39].

Why did the market turn a blind eye to such an enormous structural risk for so long? This precisely reflects its inefficiency in processing complex information, whose roots can be analyzed on several levels.

At the level of information processing cost, unlike tracking a single on-chain transfer, understanding the market-cap illusion requires deeper, comprehensive analytical capability. Investors would not only need to analyze position concentration and token distribution but also understand the token's vesting schedule, the depth of the exchange's order book, and the potential coordinated behavior among insiders. More difficult still, they would need to integrate these scattered data points into a coherent judgment that "FTT's true liquid value is far below its reported market cap." The cognitive cost of this comprehensive analysis is extremely high and often involves the details of legal and financial structures, far beyond the reach of ordinary investors.

At the level of cognitive biases, investors were affected by multiple behavioral biases, which were further amplified in the special environment of crypto markets. Authority bias made them place excessive trust in the public image and industry status of FTX's founder, equating his personal reputation with the company's financial health. Confirmation bias made them tend to focus on positive information supporting their bullish expectations (such as FTX's user growth and expanding market share) while ignoring negative signals contradicting their expectations (such as anomalous fund flows shown by on-chain data). The anchoring effect made them use FTT's historical high price (which had once reached about $80 during the 2021 bull market) as a value reference while ignoring the fragility of its fundamentals. Narrative bias made them attracted by the appealing narrative of "a leader of the crypto industry" without calmly examin-

ing the financial reality behind that narrative. The compounded effect of these biases caused the market's pricing of FTT to deviate from its true value for a long time, forming an efficiency failure lasting years.

These cognitive biases did not operate in isolation from one another but formed a self-reinforcing information loop through the social media ecosystem. Through sponsoring sports arenas, making political donations, and appearing frequently in mainstream media, FTX built a "social proof" network far exceeding that of an ordinary enterprise's reputation, so that voices skeptical of FTX were at a structural disadvantage in dissemination. A "spiral of silence" effect existed in the crypto community: when the majority consensus within the community tended to trust a certain narrative, dissenting individuals tended to stay silent, because publicly questioning a widely revered industry leader could invite ostracism from the community. This social-transmission mechanism systematically amplified individual-level cognitive biases into a market-level collective oversight.

At the level of market structure, crypto markets lack the multiple institutional safeguards of traditional financial markets. There was no mandatory independent audit system to verify the exchange's balance sheet; no unified definition and certification standard for float to prevent market-cap manipulation; no strict related-party-transaction disclosure requirement to reveal the conflict of interest between FTX and Alameda; and no effective short-selling mechanism to allow skeptics to express their views through market action and correct the price deviation. In traditional equity markets, short-sellers and sell-side analysts play an important "price-correcting" role, using deep fundamental research to discover overvalued stocks and profiting by shorting them. But in the FTT market, the cost of shorting was extremely high (high coin-borrowing rates, insufficient liquidity), and it faced the risk of being forcibly liquidated as insiders manipulated the price, which made it difficult for potential short-sellers to participate.

At a more fundamental institutional level, the unobservability of counterparty risk in crypto markets constitutes an independent dimension of efficiency assessment. When a trader cannot be sure whether their counterparty (whether an exchange or an affiliated trading party) faces solvency risk, this uncertainty itself structurally damages the quality of price discovery. In the FTX case, the fact that customer funds were misappropriated would, in traditional financial markets, trigger strict fund-segregation and independent-audit mechanisms, but crypto markets had almost no such institutional safeguards before the FTX event. In addition, the legal boundary of "insider trading" in crypto markets is itself blurred, and the legal nature of most tokens remains disputed in the major jurisdictions, which means traditional anti-insider-trading regulations may not apply, and the absence of legal deterrence further exacerbates the persistence of information asymmetry.

From the information-type dimension of the efficiency spectrum, the FTT case clearly displays an efficiency gradient: the market reacted quickly (high efficiency) to simple, observable on-chain facts such as "FTX transferred $X hundred million to Alameda," but reacted extremely slowly (low efficiency) to conclusions such as "FTT's true liquid value is far below its reported market cap," which require comprehensive analysis to reach. The former has a low information processing cost and the latter a high one, and

it is precisely this cost difference that determines the level of market efficiency.

This collective oversight is a deeper market-efficiency failure than any single price crash; it reveals the enormous gaps that still exist in crypto markets in institutional building and information infrastructure. Notably, the market-cap illusion is not unique to FTT. In the broader crypto market, the reported market caps of many tokens differ significantly from their true realizable liquid value, especially for projects in which large token blocks are held by the project team, venture capital firms, or foundations. The extreme case of FTT merely pushed this common problem to its logical extreme. It raises a question of major significance for the entire crypto market: when we talk about a token's "market cap," what are we actually talking about? If the core metric of market cap can itself be easily manipulated, then all value judgments and risk assessments built on market cap lose their reliability. This lesson has profound implications for the building of crypto markets' information infrastructure (including stricter float-definition standards, more transparent position-disclosure requirements, and more reliable price discovery mechanisms).

15.5.4 Contagion and overreaction

When the crisis erupted, market efficiency exhibited its third face: what this chapter calls "over-efficiency"—the market's reaction to negative signals in speed and magnitude exceeding what fundamentals can reasonably explain, which in finance manifests as the contagion effect. The contagion effect refers to the crisis of one institution or asset spreading, through direct financial ties, a common investor base, or purely psychological panic, to other institutions or assets with seemingly no direct connection or limited connection, causing the latter to also suffer price crashes or liquidity crises. In traditional finance, the contagion effect is one of the core topics in systemic-risk research, and the FTX event provides a typical case for observing this phenomenon in crypto markets.

In the FTX event, the experience of Solana is the clearest case of the contagion effect. FTX and Alameda had been early investors in and important supporters of the Solana ecosystem, and Alameda held large amounts of SOL tokens and Solana-based DeFi project tokens on its balance sheet. When FTX collapsed, the market formed a strong expectation: the bankruptcy liquidators would have to dump these SOL tokens on the market to repay debts. This selling expectation alone was enough to push down the price, but the more far-reaching impact came from the panic triggered by tainted association. Market participants began to equate "associated with FTX" with "carrying similar risk," and this simplistic heuristic led to a wholesale sell-off of the Solana ecosystem, not merely of the tokens that might be liquidated. The total value locked (TVL) in DeFi protocols on Solana fell by more than 70% within two weeks of the FTX collapse, as users and developers fled, forming a self-reinforcing negative loop.

As shown in Figure 15-17, from the start of the crisis on November 2 to the end of November, the price reactions of different assets diverged significantly. Bitcoin and Ethereum, as the benchmark assets of the crypto market, were also affected by the FTX collapse, but their declines were relatively contained, at about 22% and 29% respectively from the November 2 baseline to the late-November low. This decline

partly reflected the market's declining confidence in the overall crypto industry and its reduced risk appetite, but as Section 15.5.5 discusses in detail below, it also contained a component of cross-asset mechanical liquidation triggered by the unified margin system; the two sources are hard to decouple precisely in the total declines of BTC and ETH, so characterizing them entirely as "a reasonable repricing of systemic risk" may be an oversimplification. SOL's price, however, plunged from about $32 to near $13, a decline of nearly 60%, far exceeding the adjustment of the market benchmark.

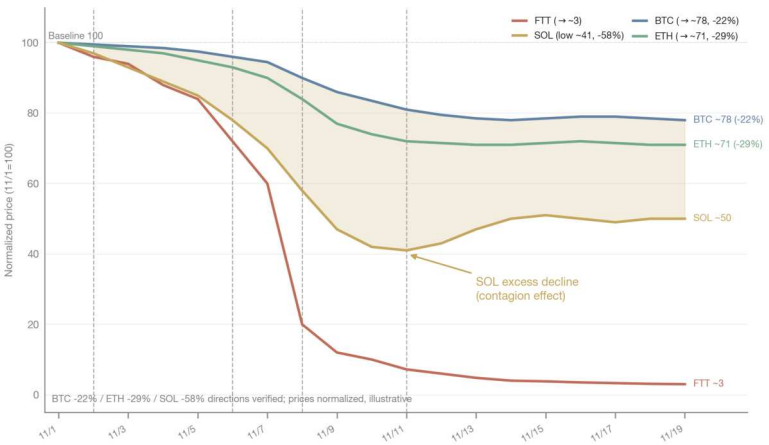


Figure 15-17. The contagion effect: the relative price reactions of SOL, BTC, and ETH during the FTX event

Figure 15-17 normalizes the prices of the four assets FTT, SOL, BTC, and ETH (with November 1, 2022 set to 100) to compare their relative performance during the FTX crisis: compared with the mild declines of BTC and ETH, SOL suffered a sharp decline similar to FTT's, far exceeding the market benchmark, and the "excess decline" region of SOL relative to BTC specially marked in the figure intuitively reflects the market's excessive punishment of the "associated party."

SOL's excess decline relative to BTC and ETH (about 36 percentage points, roughly 2.6 times BTC's decline) clearly illustrates the "over-efficiency" of the contagion effect. This so-called "over-efficiency" refers to the price discovery mechanism, under extreme panic, shifting from rational analysis based on fundamentals to a "one-size-fits-all" sell-off based on simple heuristics. Traders did not carefully assess how much actual impact FTX's bankruptcy would have on the long-term fundamentals of the Solana network (such as the independence of its technical architecture, the activity of its developer community, the growth potential of its decentralized applications, and its consensus mechanism and network security); they simply equated "associated with FTX" with "high risk" and sold off excessively as a result. This behavior seemed "efficient" in the short run, because it rapidly priced in the negative information (association risk), but in the long run it led to severe price overshooting and value destruction.

From the perspective of AMH, this overreaction is not simple "irrationality" but a form of adaptive behavior under extreme uncertainty. When market participants face

an environment of severely incomplete information and sharply rising counterparty risk, prioritizing position reduction to limit the maximum loss is a reasonable survival strategy. As Lo (2004) noted, in times of crisis, fear and the survival instinct overwhelm rational analysis, causing market efficiency to fall sharply [9]. During the FTX collapse, no one could know precisely how wide a chain reaction FTX's bankruptcy would trigger, which institutions would be the next to fall, or whether the Solana ecosystem would fall into permanent decline after losing its largest supporter. Under this deep uncertainty, overreaction is a reasonable risk-avoidance strategy at the individual level. When all market participants adopt this strategy simultaneously, however, it produces a collectively irrational outcome at the market level—prices deviating severely from fundamental value.

More noteworthy is the information revealed by the subsequent market performance. After the heavy blow of the FTX collapse, SOL's price did not correct quickly but lingered in the low range of $13 to $25 for more than a year. Not until the second half of 2023, as the Solana community carried out a series of technical upgrades (including the development of the Firedancer validator client and a marked improvement in network stability) and new ecosystem narratives emerged (including the return of users brought by the rise of meme-coin culture), did SOL's price begin to recover gradually; it briefly broke above $260 in November 2024, modestly exceeding its 2021 high, and rose further to an all-time high of about $290 in early 2025. Using this subsequent recovery to argue that the pricing of November 2022 was an "overreaction" requires caution, however. On the one hand, SOL's recovery was largely driven by new factors entirely unforeseeable at the time of the collapse, and inferring an original mispricing from a price recovery under a different causal chain faces the methodological risk of hindsight bias. On the other hand, under the information set of November 2022, market participants could not know for certain whether the large quantity of SOL held by Alameda would be liquidated in an orderly manner or dumped in a panic, nor whether the Solana network would fall into permanent decline after losing its largest supporter. A more prudent conclusion is therefore that SOL's excess decline may have contained both a reasonable risk discount (pricing the expected liquidation shock) and a panic-driven excess discount (selling beyond the reasonable range of fundamentals), but precisely distinguishing the boundary between the two would require a more rigorous quantitative estimate of Alameda's position size and the expected liquidation shock, which is beyond the scope of this chapter.

The contagion effect was not limited to SOL. FTX's collapse also directly pushed several institutions with lending or investment relationships with it into distress, forming a chain reaction that spread across the entire industry. The crypto lending platform BlockFi filed for bankruptcy in late November because of its large exposure to FTX; the lending arm of the crypto broker Genesis suspended withdrawals and ultimately went into bankruptcy reorganization; and the digital-asset management firm Galois Capital had about half of its assets (about $40 million) trapped in FTX and was forced to shut down its flagship fund in early 2023. These chain reactions further intensified market panic and formed a self-reinforcing negative feedback loop: the failure of one institution triggered worries about its counterparties, leading to more withdrawals and sell-offs, which in turn threatened the solvency of more institutions. This process is

highly similar to the transmission mechanism of systemic risk in traditional financial crises (the global financial crisis triggered by the Lehman Brothers bankruptcy of 2008 followed a similar contagion path), but in crypto markets, because of the absence of a central bank's lender-of-last-resort function, deposit insurance, and an orderly liquidation mechanism, the destructive power of the contagion effect is further amplified and its transmission speed is also faster.

From the market-state dimension of the efficiency spectrum, the appearance of the contagion effect marks the market's rapid leap from the "normal" state to the "decay" state. In the normal state, the market's arbitrage mechanism can effectively correct price deviations, keeping the prices of different assets roughly consistent with their fundamentals. But in the decay state, the arbitrage mechanism itself is paralyzed: market makers withdraw quotes to avoid risk, cross-exchange arbitrage cannot be executed because of withdrawal restrictions, and leveraged traders are forcibly liquidated for insufficient margin. The market loses its ability to self-correct; price deviations are not only not corrected but keep expanding, driven by panic. This nonlinear collapse of efficiency is precisely the extreme case predicted by the efficiency spectrum framework and a strong validation of AMH's core prediction that market efficiency falls sharply in times of crisis.

15.5.5 The liquidation cascade

The preceding analysis examined the macro timeline, on-chain information efficiency, and contagion effect of the FTX collapse, and this section focuses on a market tier that played a key but often overlooked role in the crisis: the perpetual futures market. As the core research object of this book, perpetual futures dominate price discovery in the normal state through their high liquidity and leverage advantage (see Section 15.4); but under the extreme stress test of the FTX collapse, the institutional design of perpetual futures (especially their margin mechanism and forced-liquidation logic) exposed enormous fragility and became an amplifying mechanism accelerating the market's collapse. Through the dedicated panel of perpetual futures, we can observe the nonlinear flip from institutional advantage to institutional risk during the crisis.

Consider first the price discovery leadership and basis inversion of FTT perpetual futures. During the information diffusion period of the crisis (November 2 to November 6), although the spot market reacted slowly, the perpetual futures market had already begun to show its acuity in integrating information. In theory, traders holding inside information or possessing advanced on-chain analysis capabilities are more inclined to use perpetual futures to establish short positions, because this does not require actually borrowing FTT spot and can achieve leveraged gains. Before Zhao announced the liquidation of FTT on November 6, the price of FTT perpetual futures had already begun to deviate subtly from the spot price, and the basis (the perpetual price minus the spot price) turned from slightly positive in the normal state to negative. This basis inversion was in fact an early signal that the perpetual futures market was pricing in the imminent selling pressure.

When the sell-off truly began on November 6, the price discovery leadership of perpetual futures was fully displayed. As liquidity in the spot market dried up rapidly and market

makers widened spreads or even withdrew quotes, a large number of panic sell orders and speculative short orders poured into the perpetual futures market. This caused the FTT perpetual futures price to lead the spot price throughout the crash, with the basis inversion reaching several hundred basis points at the most panicked moment on November 8. This deep basis inversion indicates that under extreme crisis, perpetual futures are not only the leader in price discovery but also the concentrated channel reflecting market panic.

The liquidation cascade produced a powerful amplifying effect that drained liquidity. As the FTT price fell sharply, another core mechanism of the perpetual futures market—forced liquidation—was triggered on a large scale. The liquidation data during the FTX collapse show an extremely severe situation: a large number of leveraged accounts collateralized with FTT or holding FTT long positions were liquidated as their margin ratios fell below the maintenance margin requirement. Within the 48 hours from November 7 to 8, the long liquidation volume of FTT perpetual futures on major exchanges (including Binance, OKX, and Bybit) surged exponentially.

This liquidation was not an isolated event but formed a typical liquidation cascade. When a long position is forcibly liquidated, the liquidation engine sends market sell orders to the market, which push the mark price down further; the lower mark price then triggers the liquidation conditions of the next batch of leveraged accounts, generating new sell orders. In the absence of sufficient long liquidity to absorb them, this mechanical selling continues to push the price down. More seriously, the EMA smoothing mechanism used by the CEX mark price (for the mechanism, see Section 15.4.3) lags behind the latest traded price under extreme conditions, allowing positions that should have been liquidated early to persist and accumulate larger losses, which finally erupt in concentration when the mark price catches up, further amplifying the FTT liquidation cascade. The open interest of FTT perpetual futures on major exchanges fell sharply within days, which is precisely direct evidence of liquidity evaporation and leverage clearing. Unfortunately, the precise liquidation data for FTT perpetual futures during the FTX event (including the ratio of forced liquidations to voluntary closures, the depletion of each exchange's insurance fund, and the frequency with which the auto-deleveraging mechanism was triggered) are incompletely recorded in public data sources, which limits precise quantitative analysis of the severity of the liquidation cascade. That said, the major exchanges did not merely respond passively during this period: platforms such as Binance and OKX typically activate dynamic margin-rate adjustments, reduce the maximum available leverage, and, when the liquidation engine is under excessive pressure, trigger the auto-deleveraging mechanism to replace conventional market liquidation under extreme conditions. These emergency measures mitigated the worst consequences of the liquidation cascade to some extent, but the opacity of their trigger conditions and execution details itself constitutes an institutional risk, as traders cannot accurately anticipate in advance how the rules will be executed under extreme conditions.

The collapse of perpetual futures was not confined to the single asset FTT. Modern derivatives exchanges such as FTX generally adopt a unified margin system, allowing users to use multiple assets (such as BTC, ETH, and SOL) as shared collateral to support all contract positions. This institutional innovation, which improves capital

efficiency in the normal state, became a perfect conductor of cross-market contagion in the crisis.

When the FTT price crash caused the overall equity of a user's unified margin account to shrink, even if their BTC or ETH perpetual futures positions themselves had not lost money, they could be forcibly liquidated because of insufficient overall account margin. This mechanism led to "indiscriminate selling": to save an account on the verge of liquidation, traders were forced to close the most liquid assets (usually BTC and ETH longs); and when forced liquidation actually occurred, the liquidation engine likewise sold all assets in the account indiscriminately. This explains why, during the FTX collapse, although the spot fundamentals of BTC and ETH had not changed fundamentally, their perpetual futures markets still suffered enormous selling pressure and a wave of liquidations. The unified margin system transmitted the crisis of a specific token to the entire crypto derivatives market through account-level linkage.

The funding rate mechanism itself suffered extreme distortion and ultimately failed. Under this series of shocks, the mechanism designed to anchor the perpetual price to the spot price underwent a historic extreme deviation. From November 7 to 9, the funding rate of FTT perpetual futures on major exchanges hit the funding-rate floor set by the exchange (its absolute value set according to the contract's maintenance margin rate, about -2% per 8 hours for standard contracts). This extreme negative funding rate meant that shorts had to pay a high fee to longs, which in theory should have attracted arbitrageurs to buy perpetual futures and sell spot, thereby narrowing the basis.

Under the "decay state" of the FTX crisis, however, this arbitrage mechanism failed entirely. On the one hand, because the FTT spot lending market was frozen, arbitrageurs could not borrow spot to hedge their shorts; on the other hand, facing an extreme one-sided decline and counterparty risk (especially worries about the FTX platform itself), no rational arbitrageur was willing to bear the risk of building a position in a sharply falling price merely to earn the funding rate. The mean-reverting nature of the funding rate failed at this moment; it was no longer a regulator promoting market efficiency but became a mechanism continually draining the remaining long positions. This mechanism-level paralysis marked the perpetual futures market losing its ability to self-correct under extreme stress, its institutional advantage completely flipping into systemic risk. From a risk-management perspective, the duration of this mechanism failure is equally worthy of attention. After the FTX event, the funding rate of major assets took about one to two weeks to recover from its extreme negative value to a normal level, with catalysts for recovery including the gradual entry of new market makers and the rebuilding of market liquidity. The classic "limits of arbitrage" theory of Shleifer and Vishny (1997) [42] applies strongly here: precisely at the moment when arbitrageurs are most needed to correct price deviations, they are least equipped with capital and most exposed to runs. In the FTX event, the fact that a large number of arbitrageurs' funds were trapped in the bankrupt exchange made the paralysis of the arbitrage mechanism far more persistent than pure market panic.

15.5.6 Summary of the FTX event

The collapse of FTX provides a rich, clearly layered, comprehensive case for the efficiency spectrum analytical framework established in this chapter. It strongly demonstrates that any attempt to describe digital asset markets with a single label ("efficient" or "inefficient") is bound to be incomplete. Only within a multidimensional efficiency spectrum can we accurately locate and understand the market's complex behavior. Figure 15-18 systematically deconstructs the FTX event within the efficiency spectrum analytical framework.

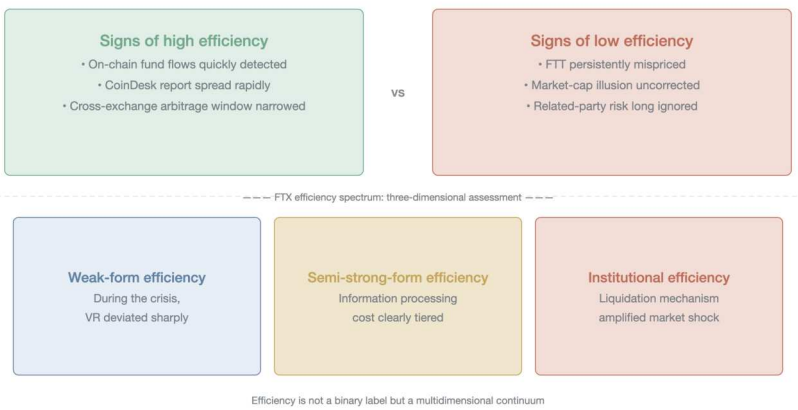


Figure 15-18. The efficiency spectrum analytical framework applied to the FTX event

Figure 15-18 systematically deconstructs the starkly different efficiency features of the FTX event along three layers—weak-form efficiency, semi-strong-form efficiency, and institutional efficiency—supplemented by a contrast of high versus low efficiency: from the high efficiency for simple on-chain data to the extremely low efficiency for complex structural information, from the nonlinear leap from the normal state to the decay state, and the role reversal of perpetual futures from price discovery leader to collapse accelerator. Table 15-5 further summarizes these manifestations systematically along the five dimensions of the efficiency spectrum.

Efficiency spectrum dimension	Specific manifestation	Efficiency assessment	Key evidence
Information type: simple on-chain data	Once the media translated the signal, the market completed 80% of the price adjustment within 48 hours	High efficiency	FTT price plunged 80% within 48 hours of CZ's tweet
Information type: complex structural information	The market-cap illusion and circular collateralization went unpriced for years	Extremely low efficiency	About 80% of FTT supply was controlled by insiders; market cap overstated by nearly tenfold
Market state: from normal to decay	Liquidity evaporated within hours and the price discovery mechanism seized up	Nonlinear collapse	FTX faced $6 billion in withdrawal requests and was forced to halt withdrawals
Asset dimension: contagion	SOL's decline (about 58%) far exceeded BTC's (about 22%), and part of it was later shown to be an overreaction	Over-efficiency	SOL modestly exceeded its 2021 high in November 2024 and rose further to about $290 in early 2025
Market tier: perpetual futures	The institutional advantage that dominates price discovery in normal times flipped into institutional risk—"cascading liquidations plus liquidity evaporation"—during the crisis	Nonlinear reversal	Open interest fell off a cliff, funding rates went to extremes, the basis inverted deeply, and unified margin triggered cross-asset liquidations

Table 15-5. Summary of the multidimensional manifestations of the efficiency spectrum during the FTX event (Data source: compiled by the author from public data on the FTX event)

The table shows clearly that the FTX event exhibited starkly different efficiency features along the different dimensions of the efficiency spectrum, and these features together constitute a complete characterization of the multidimensional nature of market efficiency.

Along the information-type dimension, the market reacted quickly to simple on-chain data (fund flows, token transfers) but lagged for years on complex structural fraud (the market-cap illusion, circular collateralization), pushing the information-processing-cost stratification revealed in Section 15.3 to the extreme—the simplest information was priced at the hourly level, while the most complex information was "discovered" in the form of a collapse rather than a gradual adjustment. Along the market-state dimension, the liquidity evaporation and negative-feedback spiral triggered by the collapse of trust paralyzed the arbitrage and price discovery mechanisms within hours, corroborating the nonlinear collapse predicted by the efficiency spectrum and validating AMH's core prediction that efficiency falls sharply in a crisis; and this collapse was not uniform—

BTC and ETH, with deeper liquidity and more diverse participants, retained relative resilience, while FTT and SOL suffered more severe failures, again highlighting the importance of the asset dimension. Along the time dimension, the market's long-term sluggishness toward FTT's structural risk, its medium-term moderate reaction to the CoinDesk report and Zhao's tweet, and its short-term overreaction during the collapse constitute an efficiency gradient that echoes the weak-form efficiency findings of Section 15.2 but is amplified by extreme conditions.

Along the asset dimension, the divergent reactions of FTT, SOL, and BTC/ETH correspond precisely to the full spectrum from low efficiency (the belated correction of fraud information), through over-efficiency (the contagion punishment of an associated party), to reasonable repricing (the benchmark assets). Distinguishing "fundamental contagion" (a reasonable risk repricing transmitted through quantifiable financial ties and position exposure) from "pure contagion" (panic spillover beyond the range explainable by fundamentals) is the core analytical framework for assessing the nature of the contagion effect, and SOL's excess decline in the FTX event may have contained both components, though the literature has not yet provided a quantitative decomposition. Along the market-tier dimension, perpetual futures reflected the selling pressure earlier than spot during the information diffusion period, displaying price discovery leadership, and then, when the crisis fully erupted, became the core mechanism amplifying the downside and accelerating the price adjustment because of their leverage nature, liquidation mechanism, and unified margin system—this flip from institutional advantage to institutional risk reveals the fragile boundary of derivatives-market efficiency.

Ultimately, the FTX case is not only a warning to market participants but also important research material for students of market efficiency. In an extreme way, it forces us to abandon a simplistic, idealized understanding of the market and embrace a more realistic, dynamic, multidimensional analytical paradigm. Informational efficiency is not a fixed attribute but a continuous spectrum that changes continually with information type, market state, asset characteristics, and time scale. Understanding this not only helps us explain past market behavior more accurately but also provides a more robust analytical foundation for future risk management and investment decisions. Just as AMH predicts, the market will learn and adapt from the FTX crisis: stricter exchange-audit requirements, more transparent proof-of-reserves mechanisms, and more prudent counterparty-risk management practices are all concrete manifestations of this adaptive process. After the FTX event, major exchanges rolled out proof-of-reserves systems based on Merkle trees. Proof-of-reserves systems in their current form have significant limitations, however: they can usually prove only the existence of assets at a specific point in time and cannot verify the completeness of the liabilities side, and an exchange may have off-chain borrowing and liabilities not included in the scope of the proof. Moreover, most proof-of-reserves systems take the form of monthly or quarterly snapshots, and this low-frequency auditing leaves ample windows for manipulation between snapshots. These limitations mean that proof of reserves may to some extent give market participants a false sense of security rather than truly bridging the gulf of information asymmetry. At the same time, regulators have accelerated the legislative process for the crypto industry, and these institutional-level changes are fundamentally reshaping the information environment and efficiency features of crypto markets. From

AMH's population-environment framework, these ex post institutional changes can be understood as a typical "population adaptation after an environmental change": the collapse of FTX, as a strong environmental shock, eliminated the fragile strategies relying on centralized trust assumptions and gave rise to a new generation of market participants and infrastructure characterized by higher transparency and stricter risk control. And the long-term trend of this adaptive evolution is precisely the subject we will discuss systematically below.

Efficiency Spectrum Locator Card · Section 15.5

Dimension	Positioning	Quantitative proxy metric
Time	A nonlinear leap from long-term sluggishness to a short-term instantaneous eruption	FTT completed 80% of the adjustment within 48 hours of the key information release
Asset	The collapse of the direct underlying (FTT) and the contagion overshoot of the associated party (SOL)	SOL's excess decline of about 36 percentage points; the subsequent rebound partly proved the overreaction
Information type	High efficiency for simple on-chain data, extremely low efficiency for complex structural information	Rapid reaction after the media translated the signal, but the "market-cap illusion" lay latent and unpriced for years
Market state	A rapid leap from the normal state to a "decay state" of arbitrage paralysis	FTX faced $6 billion in withdrawals; liquidity evaporation triggered a negative-feedback spiral
Market tier	Perpetual futures flipped from price discovery leader to collapse-accelerating mechanism	Open interest fell sharply, funding rates went to extremes, a cross-asset liquidation cascade erupted

One-sentence takeaway: Under extreme stress, digital asset market efficiency exhibits a nonlinear collapse, and the institutional design of perpetual futures transforms from an efficiency accelerator in the normal state into a risk amplifier in the crisis.

15.6 The evolution of efficiency

After the preceding sections' in-depth dissection of the informational efficiency of digital asset markets along different dimensions, we have clearly seen that efficiency is not a static, unchanging attribute but a dynamic, multilevel complex system. From the time-varying nature of weak-form efficiency, through the differentiated reaction of semi-strong-form efficiency to different information types, to the distinctive insti-

tutional efficiency cycles of perpetual futures, all the evidence points to Lo's (2004) adaptive markets hypothesis as the appropriate lens for understanding this emerging market [9]. This section aims to integrate this fragmented evidence into a unified historical-evolutionary framework, systematically expounding the evolution of crypto market informational efficiency from its birth to the present by constructing a three-stage model. This trajectory is not only a history of the evolution of technology and capital but also a history of the ongoing game, adaptation, and co-evolution among three parties: market participants, infrastructure, and the regulatory environment.

Our core argument is that the informational efficiency of digital asset markets has passed through three distinct but interlocking stages: the frontier era (2009–2017), the institutionalization era (2018–2023), and the regulated era (2024–). Figure 15-19 presents a panorama of this evolution and marks within it the institutionalization process of perpetual futures.

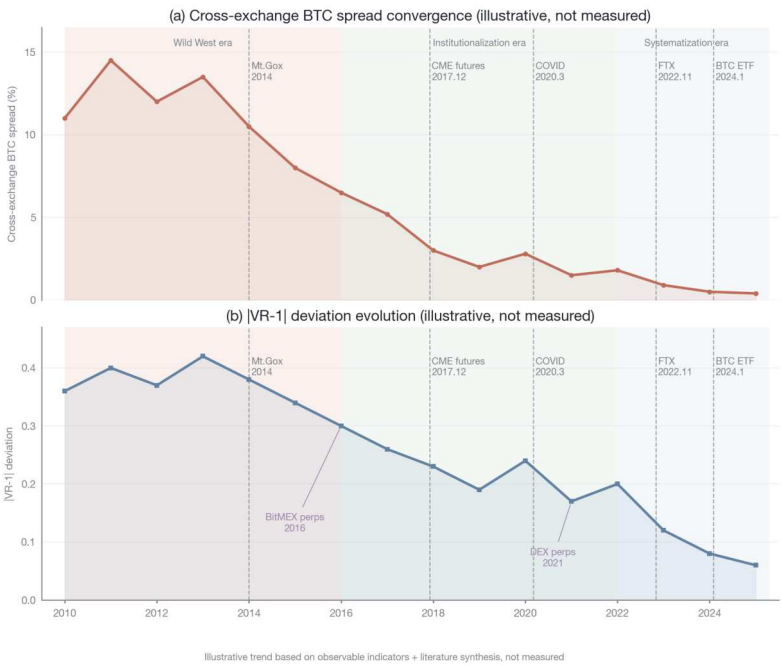


Figure 15-19. The three-stage evolution of crypto market efficiency

Figure 15-19 uses two panels to present the illustrative trend of efficiency evolution (the direction and magnitude are based on a synthesis of the literature rather than the point-by-point measurement of a single continuous data source): panel (a) shows the convergence of cross-exchange BTC spreads from 2010 to 2025, with lower spreads representing higher arbitrage efficiency; panel (b) shows the decline of the variance-ratio deviation $|\mathrm{VR}-1|$, with lower deviation representing prices closer to a random walk. The timeline is colored by the three stages—frontier, institutionalization, and regulated—and marks key events such as the Mt. Gox collapse, the CME futures launch, COVID-19, the FTX collapse, and the BTC ETF approval, along with perpetual futures milestones (such as the first perpetual futures on BitMEX in 2016 and the scaling of

DEX perpetuals in 2021), intuitively displaying the brief rebounds of both metrics during crises and their long-run convergence.

As the figure shows, each stage is defined by its distinctive participant composition, infrastructure level, and regulatory environment, and these factors together shape the efficiency features of that stage. Overall, market efficiency exhibits a long-term trend of spiral improvement, but this process is not linear; it is accompanied by cyclical crises and resets triggered by technical defects, speculative mania, and fraud. These crises are both stress tests of the market structure of the time and often the key catalysts that give rise to the next stage of evolution. Table 15-6 summarizes the core differences among the three stages along key dimensions.

Dimension	Frontier era (2009–2017)	Institutionalization era (2018–2023)	Regulated era (2024–)
Main participants	Cypherpunks, early retail investors, tech enthusiasts	Professional market makers, quant funds, crypto-native institutions	Traditional finance giants, ETF issuers, sovereign wealth funds
Infrastructure	Primitive exchanges, no derivatives	Regulated futures, professional APIs, institutional custody	Spot ETFs, global compliant custody, high-speed Layer 2
Perpetual futures	Not yet born / very early (BitMEX monopoly)	Became the highest-volume derivative; institutional design iterated rapidly	CEX perpetuals highly mature; DEX perpetuals catching up fast; regulatory frameworks intervening
Regulatory environment	Almost a vacuum	Frameworks first established; enforcement actions increasing	Comprehensively regulated
Cross-exchange BTC spread	5%–30% (normal)	1%–5% (normal); spikes during crises	<0.5% (major exchanges)
Weak-form efficiency level	Significant departure from a random walk	Intraday approaches a random walk; patterns persist across days	Close to a random walk across multiple time scales
Signature crisis	Mt. Gox collapse (2014)	Black Thursday (March 12), FTX collapse	To be observed
Catalyst for the phase transition	—	CME futures launch (December 2017)	Spot BTC ETF approval (January 2024)

Table 15-6. Comparison of the core features of the three-stage evolution of crypto market efficiency (Data source: compiled by the author)

This section examines in turn the efficiency features, driving factors, and key turning points of these three stages, uses quantitative evidence to show the concrete trajectory of efficiency evolution, and finally offers a prospect for the emerging forces shaping the next stage's efficiency paradigm. This historical perspective is both a comprehensive summary of all the findings in this chapter and a key clue for understanding the future development path of digital asset markets.

15.6.1 The frontier era (2009–2017)

The first decade of digital asset markets belonged to a highly uncertain early stage. From the birth of the Bitcoin genesis block to the eve of the CME Bitcoin futures launch in late 2017, the market efficiency of this stage was defined by extremely high informational and transactional friction, which offered enormous arbitrage room to the few early participants with specific skills. Market participants at this time were mainly cypherpunks, software developers, early tech enthusiasts, and a few retail speculators attuned to market opportunities. Their motives were diverse, ranging from technological faith in the decentralization ideal to a speculative desire for high returns. This homogeneity of the participant base meant that the market lacked the sustained price discovery process driven by heterogeneous beliefs and diversified strategies found in mature financial markets.

The infrastructure of this period was extremely primitive and fragile. Trading was concentrated mainly on a few centralized exchanges, the largest by volume being Tokyo-based Mt. Gox. At its peak from 2013 to early 2014, Mt. Gox handled roughly 70% of global Bitcoin trading volume, forming a de facto price monopoly. These early exchanges generally lacked sound internal controls, security measures, and customer-asset protection mechanisms, however. Trading engines crashed frequently, withdrawal delays were the norm, and the catastrophic hack that ultimately forced Mt. Gox to file for bankruptcy in February 2014 fully exposed the fragility of this period's infrastructure. In such an environment, even simple cross-exchange arbitrage faced multiple barriers, because traders had to pre-position funds on multiple exchanges, bear extremely high counterparty and operational risk, and endure the execution delay imposed by Bitcoin network confirmation times.

The near-total absence of regulation was another core feature of the frontier era. Governments and financial regulators around the world were still exploring the legal nature, accounting standards, and regulatory framework of Bitcoin and its derivatives, and the market expanded in disorder in a near-anarchic state. This regulatory vacuum provided fertile soil for innovation on the one hand but, on the other, condoned the proliferation of market manipulation, fraud, and money laundering. The large-scale expansion of ICOs in 2017 was a concentrated embodiment of the consequences of this regulatory absence.

Together, these factors produced the extremely low overall market efficiency of the frontier era. As detailed in Section 15.2.5, cross-exchange spreads long stayed above 5%, and regional spreads such as the "Kimchi premium" reached double digits [25]; weak-form efficiency tests also showed significant predictability (see Section 15.2.2), with return series departing significantly from a random walk [2].

The final years of the frontier era saw the emergence of an innovation that would reshape the efficiency of crypto markets. In 2016, BitMEX launched the first Bitcoin perpetual futures (XBTUSD). As a derivative with no expiration date that anchors to the spot price through a funding rate mechanism, it was initially intended only to solve the problems of fragmented liquidity and difficult fiat on- and off-ramps in the early crypto market. In its early days, however, the efficiency of the perpetual futures market was extremely low. Because of the absence of mature market makers

and arbitrageurs, the funding rate fluctuated enormously, frequently reaching extreme positive or negative values, and the magnitude of the basis's deviation from the spot price was also substantial. Funding-rate arbitrage in this period was very limited, conducted mainly by a few technical geeks and early quant teams. This early low-efficiency state offered exactly the rich profit space for the arbitrage capital that would follow and foreshadowed that perpetual futures would become the core force driving the evolution of market efficiency.

15.6.2 The institutionalization era (2018–2023)

In December 2017, the Chicago Mercantile Exchange and the Chicago Board Options Exchange almost simultaneously launched cash-settled Bitcoin futures contracts. This event is widely regarded as a key watershed in the evolution of digital asset markets, marking the opening of the institutionalization era. In this stage, the market's core driving force shifted from the early tech enthusiasts and retail investors to professional financial institutions, quant funds, and market makers. Their entry dramatically changed the market's participant composition, infrastructure level, and efficiency paradigm.

The influx of professional participants brought capital, technology, and experience. The deep participation of top market makers such as Jump Trading, Wintermute, and GSR significantly improved the liquidity depth and order-book quality of major exchanges and greatly compressed cross-exchange spreads; as shown in Section 15.2.5, spreads between major exchanges converged rapidly to below 1% after 2018. The emergence of regulated derivatives provided professional institutions with effective hedging tools and price benchmarks, significantly promoting the price discovery efficiency of the spot market [43].

Infrastructure also underwent a fundamental upgrade in this period. A new generation of exchanges represented by Binance, Coinbase Pro, and FTX offered more powerful, more responsive trading APIs, facilitating programmatic trading. The emergence of compliant custody solutions addressed institutional investors' core concern about asset security. At the same time, the wide adoption of stablecoins (especially USDT and USDC) provided an efficient on-chain channel for cross-exchange fund transfers, greatly reducing the capital friction of arbitrage. Figure 15-20 shows the core driving factors of this stage—professional participants, upgraded infrastructure, and an initially established regulatory framework working together to drive efficiency improvement.



Figure 15-20. The driving factors and efficiency features of the three-stage evolution

Figure 15-20 uses a structured view to show the core factors driving efficiency evolution in the three stages: the bottleneck of the frontier era lay in low participation and primitive infrastructure; the institutionalization era saw efficiency leap forward driven jointly by professional market makers, regulated derivatives, and stablecoin infrastructure; and the regulated era pushed efficiency to new highs through ETF channels, global regulatory frameworks, and the entry of traditional finance giants, with each phase transition catalyzed by a signature event.

The institutionalization era witnessed the explosive growth of perpetual futures and a comprehensive improvement in their efficiency. Between 2019 and 2020, major exchanges such as Binance, OKX, and Bybit fully introduced perpetual futures, whose volume rapidly overtook that of traditional dated futures to become the most central liquidity pool in crypto markets. With the deep participation of professional market makers and quant funds, the efficiency of the perpetual futures market improved significantly. Cross-exchange perpetual-futures basis arbitrage became extremely active, causing price differences across platforms to be flattened rapidly. The fluctuation of the funding rate also gradually converged, the frequency of extreme rates fell sharply, and the speed of mean reversion clearly accelerated [21]. During this period, perpetual futures gradually captured the price discovery leadership of crypto markets and became the tool of choice for informed traders to express their views. This efficiency improvement was not without cost, however. The collapse of FTX in 2022 exposed the systemic fragility of the centralized perpetual futures market: when the collapse of trust caused liquidity to evaporate instantly, high leverage and complex liquidation mechanisms triggered a cascade effect, causing a severe efficiency regression in the perpetual futures market under extreme stress. This event also prompted the market to reflect on the risks of the centralized architecture and laid the groundwork for the rise of decentralized perpetual futures.

The improvement in efficiency was not without cost, however. Although the deep participation of professional market makers significantly improved liquidity and price discovery quality in the normal state, their homogeneous strategies and risk models could also constitute a new source of fragility under extreme conditions. When most market makers use similar volatility models and inventory-management algorithms, market stress can trigger synchronized liquidity withdrawal, causing the seemingly ample order-book depth of the normal state to evaporate within minutes; the "Black Thursday" of March 2020 is a typical case of this "phantom depth" risk. This tension between efficiency improvement and the accumulation of fragility is a key dimension for understanding the evolution of crypto markets.

At the weak-form efficiency level, as shown by the variance ratio analysis in Section 15.2.2, the market efficiency of Bitcoin and Ethereum exhibited a significant and sustained improvement trend after 2018 [7], with the daily variance ratio converging toward 1.0 and the intraday scale already highly close to a random walk [6]. This improvement was stratified, however: the efficiency of the core market improved far faster than that of the peripheral market [44].

15.6.3 The regulated era (2024–)

On January 10, 2024, the U.S. Securities and Exchange Commission formally approved the first spot Bitcoin ETFs, and this milestone event announced that digital asset markets had entered a new regulated era. If the protagonists of the institutionalization era were crypto-native professional funds and market makers, then the protagonists of the regulated era are the traditional finance giants of Wall Street, such as BlackRock, Fidelity, and Invesco, the world's largest asset managers. Their entry brought not only an unprecedented scale of capital but, more importantly, integrated digital assets into the compliance framework and infrastructure of traditional finance, fundamentally changing the operating logic of the market.

The launch of spot ETFs provided a low-barrier, compliant channel for vast amounts of regulated capital to invest in Bitcoin. The ETFs' authorized participants (APs) (usually large trading firms such as Jane Street and Virtu Financial) took on the role of arbitraging between the ETF-share market and the Bitcoin spot market. This mechanism ensured that the ETF's market price closely tracks its net asset value and also coupled Bitcoin's price discovery mechanism with traditional financial markets more tightly than ever before. The ETF's authorized participants can be said to serve as an "efficiency bridge" between the traditional financial world and the crypto-native world. In assessing the attribution of the regulated era's efficiency improvement, however, one must also consider the potential contribution of the contemporaneous global macro-liquidity environment. The improvement in efficiency metrics in 2024 coincided with rising market expectations of Federal Reserve rate cuts and a rebound in global risk appetite, and these macro factors may have had a positive impact on crypto market efficiency independent of the ETF channel; future research needs to control for the macro-liquidity cycle to isolate more precisely the independent contribution of the regulatory factor.

At the same time, regulatory frameworks around the world matured. The European Union's Markets in Crypto-Assets Regulation (MiCA) began phased implementation in 2024, and similarly clear regulatory guidance was successively issued in Hong Kong, Singapore, Dubai, and elsewhere. This regulatory certainty greatly reduced the compliance risk and legal uncertainty faced by institutional investors and laid the foundation for larger-scale adoption.

In the regulated era, the market's efficiency paradigm exhibits features starkly different from the previous two stages. Riskless arbitrage opportunities in core assets have been almost entirely eliminated. Spreads between major compliant exchanges have been compressed to within a few basis points. At the weak-form efficiency level, the market has reached a very high level across all time scales. As shown in Figure 15-21(b), the daily and weekly variance ratios of Bitcoin are already very close to 1.0, indicating that exploitable price patterns have become extremely scarce.

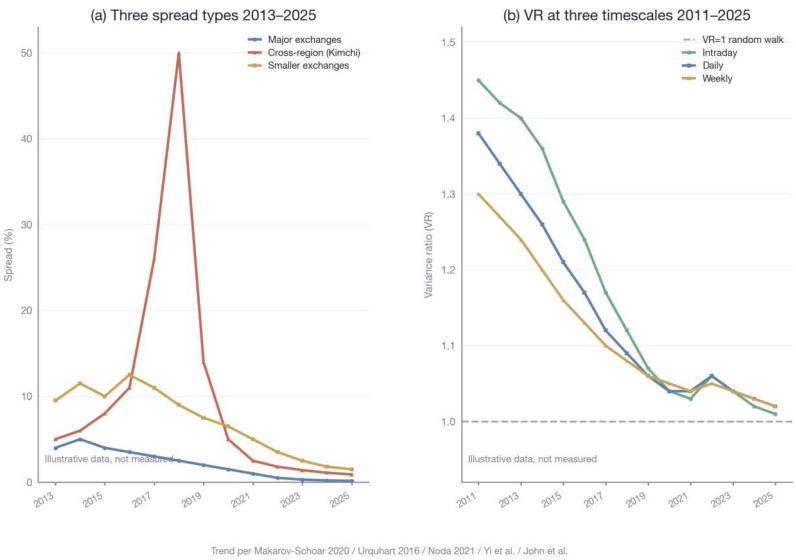


Figure 15-21. Quantitative evidence of efficiency evolution (illustrative)

Figure 15-21 uses illustrative data to show the qualitative trend of efficiency evolution, with the direction and magnitude synthesized from the relevant literature: the spread convergence references Makarov and Schoar (2020) [25] and John et al. (2024) [44], and the variance-ratio evolution references Urquhart (2016) [2], Noda (2021) [7], and Yi et al. (2023) [6]. Panel (a) shows that the three types of cross-exchange Bitcoin spreads (between major exchanges, cross-region such as the Kimchi premium, and between small and mid-sized exchanges) have declined continuously since 2013 and accelerated their convergence after the two nodes of 2018 and 2024; panel (b) shows that the intraday, daily, and weekly variance ratios have all converged toward 1.0 (the random-walk benchmark) since 2011, indicating that price predictability has gradually declined and weak-form efficiency has strengthened continuously.

On entering the regulated era, the perpetual futures market faces new challenges and reshaping. First, with the approval of spot ETFs, the arbitrage activity of traditional financial capital (such as the cross-market arbitrage of ETF authorized participants) has had a profound impact on the funding-rate structure of perpetual futures, making the dynamics of the basis and funding rate more complex. Second, global regulators' restrictions on high-leverage derivatives are increasingly strict, and some jurisdictions even prohibit offering perpetual futures trading to retail investors. This regulatory pressure forces liquidity to be reallocated across platforms. A more striking change is the rapid rise of DEX perpetual futures. High-performance decentralized exchanges represented by dYdX v4 and Hyperliquid have significantly narrowed the gap with centralized exchanges in the latency and execution cost of day-to-day trading, attracting a large amount of institutional capital seeking to avoid counterparty risk. These platforms still face tests in execution quality and manipulation resistance under extreme conditions, however: in the whale manipulation incident on Hyperliquid in March 2025, validators voted to forcibly liquidate an anomalous position and rolled back the associ-

ated price; this intervention protected the system from larger losses but also exposed the centralized nature of its governance. In fact, the sequencers of most on-chain perpetual futures platforms are still operated by a limited validator set controlled by the project team, and this architecture of "technical decentralization but centralized governance" means that risks such as transaction censorship and priority ordering still exist. Nonetheless, on-chain perpetual futures still differ from off-chain ones in oracle update frequency and liquidation efficiency, and this structural difference between CEXs and DEXs creates new arbitrage room and constitutes a frontier topic in the study of perpetual futures market efficiency in the regulated era [22]. The empirical conclusion that CEXs dominate price discovery (such as the study by Zhivkov (2026) [22]) is based mainly on the period when DEX perpetual futures had not yet scaled, and with the explosive growth of on-chain perpetual futures volume since the second half of 2024, the direction of information flow between CEXs and DEXs may be changing.

This does not mean the market is already fully efficient, however. As discussed earlier with respect to semi-strong-form efficiency, information processing cost remains the key to earning excess returns. In the regulated era, the focus of competition has shifted from eliminating simple price differences to the deep analysis of and rapid reaction to complex information. This includes microstructure analysis of on-chain data, interpretation of the information implied in derivatives markets, and quantitative modeling of how macroeconomic and geopolitical factors affect crypto assets. The market's "information advantage" has shifted from discovering simple signals that are public but unpriced to gaining a competitive edge through better models, faster execution, and lower operating costs.

15.6.4 Future drivers

Looking ahead, the evolution of digital asset market efficiency is far from over. New technology, market structures, and regulatory paradigms will continue to reshape its face, and several key drivers deserve especially close attention.

Artificial intelligence, and especially autonomous trading agents driven by large language models, is likely to become an important force in the field of information processing. AI agents have the potential to greatly reduce information processing cost, raising the market's reaction speed to public information from the minute level to the second or even millisecond level. This may also bring new risks, however: "algorithmic resonance" among AI agents may amplify market volatility under specific conditions, forming a new source of systemic risk. A deeper risk is that when a large number of AI agents use similar large-language-model architectures, similar training data, and similar decision frameworks, they may produce an "algorithmic herding" that goes beyond traditional herd behavior, with a transmission speed and strategy consistency far higher than those of human traders, potentially producing a collective directional bias when fundamentals are ambiguous.

The continued iteration of the underlying infrastructure will likewise strongly affect the market's execution efficiency. Faster blockchains, more efficient cross-chain communication protocols, and more advanced exchange matching engines will further reduce trading latency and execution cost.

The global coordination and deepening of regulation will be another key variable shaping the future of market efficiency. As major economies complete the regulatory characterization and framework-building for crypto assets, the room for cross-border regulatory arbitrage will be compressed. The impact of regulation is not a one-way efficiency improvement, however. Strict compliance requirements may raise the friction cost of arbitrage—for example, MiCA's capital requirements for stablecoin issuers may affect the availability of cross-exchange arbitrage media such as USDT; differentiated regulation across jurisdictions may drive liquidity to migrate from compliant platforms to loosely regulated offshore platforms, creating new market fragmentation; and restrictions on perpetual futures leverage may lead to a reallocation of informed traders across platforms. In addition, even for spot ETFs already approved, the SEC retains ongoing regulatory authority and the possibility of suspending or revoking approval upon finding evidence of market manipulation, and this policy uncertainty itself constitutes a risk factor in efficiency evolution.

The perpetual futures market will continue to play a key role in the efficiency evolution to come. As the underlying infrastructure upgrades, we can expect perpetual futures to evolve toward higher-frequency settlement cycles and more transparent liquidation mechanisms. The maturation of on-chain perpetual futures will further narrow their efficiency gap with centralized exchanges and may even, through innovative decentralized oracles and fully on-chain order-book designs, surpass them along certain dimensions. In addition, as the correlations among different crypto assets change, cross-asset funding-rate arbitrage and complex statistical-arbitrage strategies will become more prevalent. Perpetual futures will not only continue to serve as the engine of crypto market price discovery, but the evolution of their own institutional design will also become a core driving force pushing the entire market toward a more efficient form.

In sum, the evolution of crypto market efficiency is an ongoing adaptive process. From the chaos and disorder of the frontier era, through the professional competition of the institutionalization era, to the compliant integration of the regulated era, the rules of the market's game keep changing, and efficiency improvement is always accompanied by the emergence of new forms of fragility. The future market will be an ecosystem shaped jointly by AI agents, high-speed infrastructure, and global regulatory frameworks—more efficient but possibly also more complex. For all market participants, understanding and adapting to this evolutionary trend (just as the adaptive markets hypothesis reveals) will be the key to continued survival and development. This is precisely the core insight that this chapter—from theoretical toolkit to empirical testing, and then to case analysis and the historical-evolutionary framework—has sought to offer the reader.

15.7 Chapter summary

This chapter, centering on the informational efficiency of digital asset markets, constructed the multidimensional "efficiency spectrum" analytical framework and applied it systematically to the empirical testing of weak-form efficiency, semi-strong-form efficiency, institutional efficiency, and extreme-stress scenarios. The core claim of this framework is that any statement about market efficiency divorced from specific dimensional qualifications may be misleading. A market can exhibit random-walk characteris-

tics at the millisecond scale of high-frequency trading yet reveal predictable momentum or reversal patterns at the daily or weekly level; the efficiency of the Bitcoin market differs significantly from that of small-cap tokens; and the performance of efficiency differs enormously between normal markets and crisis states of dried-up liquidity. The efficiency spectrum framework decomposes efficiency into five dimensions—time, asset, information type, market state, and market tier—providing an operational tool for precisely locating the efficiency features of any market phenomenon.

Within this framework, the efficiency features of perpetual futures ran through the analysis at every level of this chapter. In the weak-form efficiency tests, the perpetual futures of major assets are closer to a random walk than spot on the high-frequency intraday scale, but the mean-reverting nature of the funding rate and the basis constitutes an institutional weak-form efficiency gap [21]. In the semi-strong-form efficiency tests, perpetual futures exhibit an information-amplification effect because of their leverage and short-selling convenience, reacting to sudden events faster than spot. At the institutional efficiency level, the funding settlement once every 8 hours endogenously creates a predictable efficiency cycle (the U-shaped pattern), and the mark-price smoothing mechanism and discrete oracle updates introduce a systematic lag in information reflection. In the extreme stress test of the FTX collapse, the normal-state efficiency advantage of perpetual futures rapidly reversed into a source of systemic risk, as high leverage and the liquidation mechanism triggered a cascade effect that caused liquidity to evaporate. Complementing the institutional analysis of perpetual futures, information processing cost was shown to be the key variable determining semi-strong-form efficiency. Blockchain transparency makes vast amounts of data publicly available, but a huge processing gulf exists between raw on-chain transaction records and tradable signals: the market's reaction to standardized macro information is nearly instantaneous, while its reaction to complex on-chain information lags significantly, and this structural advantage created by differences in processing cost is one of the core sources of excess returns in the current market.

From the perspective of dynamic evolution, drawing on Lo's (2004) adaptive markets hypothesis [9], crypto market efficiency adjusts continually with changes in participant composition, infrastructure, and the regulatory environment: from early chaos and disorder, through the professional competition of the institutionalization era, to the compliant integration of the regulated era. The gradual weakening of anomalies in the weak-form efficiency tests and the long-run convergence trend of cross-exchange spreads provide ample empirical support for this evolutionary process. These findings have direct practical implications: the success or failure of a trading strategy depends on whether the inefficiency it relies on truly exists, whether it is sustainable, and whether it is robust across different market states. Table 15-7 turns the efficiency spectrum framework into a decision checklist for traders and risk managers.

Efficiency dimension	Key assessment question	Application for traders	Application for risk managers	Warning signs
Time dimension	What time scale does the strategy rely on? What is the efficiency level at that scale?	High-frequency strategies must verify whether predictable patterns still exist at the minute level; daily strategies must test the current strength of momentum or reversal effects.	Assess whether the time frequency used in the risk model matches the actual trading frequency; avoid inferring high-frequency risk from daily data.	The backtest performs excellently at one time scale but fails at others; strategy returns fluctuate sharply as the time scale changes.
Asset dimension	How liquid is the target asset, what is its market cap, and how much analyst coverage does it have?	Strategies on small-cap coins require higher risk-premium compensation; do not apply the efficiency assumptions of major coins to long-tail assets.	Set stricter position limits and stop-losses for long-tail-asset exposure; liquidity stress tests should account for differences across asset tiers.	The strategy works on BTC/ETH but fails on small-cap coins; or, conversely, the "excess return" on long-tail assets may be a liquidity premium rather than alpha.
Information-type dimension	How high is the processing cost of the information the strategy relies on? How fast does the market react to that type of information?	Strategies exploiting complex on-chain signals require investment in analytical tools and expertise; the arbitrage window for standardized information is extremely short and demands high-frequency infrastructure.	Assess whether the team has the capacity to interpret the key information types; exposure to high-processing-cost information requires a longer risk window.	The information advantage the strategy relies on is being eroded by the spread of tools; or the strategy overestimates its own speed advantage in interpreting complex information.
Market-state dimension	How does the strategy perform across different market states? Which state are we currently in?	A strategy that works in the normal state may fail or even reverse in a crisis; build dynamic position-adjustment mechanisms for state transitions.	Stress tests must simulate efficiency-collapse scenarios, not merely price volatility; liquidity risk amplifies nonlinearly in a crisis.	The strategy is robust in backtests but suffers abnormal, massive losses during March 2020 or the FTX event; risk indicators gave no warning on the eve of the crisis.
Institutional efficiency	Are there predictable institutional efficiency cycles (such as funding settlement)?	Perpetual-futures strategies must account for price behavior around funding settlement; mark-price lag may affect stop-loss execution.	Assess the cost of derivatives exposure under extreme funding rates; a divergence between the mark price and the index price may trigger unexpected liquidations.	Strategy returns are highly concentrated in specific windows around funding settlement; an annualized funding rate above 30% in absolute value is a significant tail-risk signal; a cross-exchange spread persistently above 20 basis points warrants caution about the limits of arbitrage.
Cross-dimensional synthesis	Is the strategy simultaneously exposed to risk across multiple dimensions of the efficiency spectrum?	Avoid using high-frequency strategies on low-efficiency assets (a time × asset mismatch); avoid relying on a complex-information advantage during a crisis (a state × information mismatch).	Identify the possibility of multidimensional risk resonance: small-cap coin + crisis state + complex information = extremely high risk; build a multidimensional risk matrix.	The strategy looks robust along a single dimension but collapses when pressures stack across dimensions; risk exposure overlaps across several low-efficiency quadrants of the efficiency spectrum.

Table 15-7. A practical guide: a decision checklist based on the efficiency spectrum (Data source: compiled by the author)

Note: This checklist is intended to turn the chapter's efficiency spectrum framework into

an actionable decision tool. In use, evaluate and record answers item by item according to the specific strategy and market environment, in order to identify potential risks in efficiency assumptions.

Efficiency Spectrum Locator Card · Chapter 15 Summary

Dimension	Core finding	Key evidence
Time	Major assets approach weak-form efficiency intraday, with predictability across days; funding settlement creates periodic efficiency fluctuations	High-frequency variance ratio approaches 1.0; U-shaped volatility pattern around the 8-hour settlement
Asset	Core assets (BTC/ETH) are far more efficient than long-tail assets; the stratification of liquidity determines the stratification of efficiency	Spreads between major exchanges <0.5%; significant momentum/reversal effects in small-cap tokens
Information type	The market reacts instantly to standardized information but slowly to complex on-chain information; information processing cost is the core friction	ETF approval priced within minutes; about 36% of the price move after a DeFi hack precedes the announcement
Market state	Efficiency improves steadily in the normal state but collapses nonlinearly under extreme stress; efficiency is highly state-dependent	Long-run convergence of cross-exchange spreads; liquidity evaporation and basis dislocation during the FTX collapse
Market tier	Perpetual futures dominate price discovery in the normal state but may amplify systemic risk in a crisis; the funding rate has endogenous predictability	Perpetual futures hold the majority of the information share; liquidation cascades cause price overshooting during crises

One-sentence takeaway for the chapter: The informational efficiency of digital asset markets is a multidimensional, dynamic, and state-dependent complex system; perpetual futures, as its core hub, improve market efficiency in the normal state by aggregating informed traders, but leave predictable periodic gaps in their institutional design and face the risk of nonlinear collapse under extreme stress.

This chapter's exploration of the time-varying nature of informational efficiency also naturally draws our attention to the core themes of the chapters to come. The dynamic evolution of market efficiency, and especially its sharp collapse under extreme stress, is the key to understanding the microstructure and price-benchmark governance of digital asset markets. When information processing cost is high and market participants struggle to reach consensus on an asset's fundamental value, price volatility intensifies; when liquidity dries up and the price discovery mechanism fails, even a small order can trigger enormous price impact. In the chapters that follow, we further explore themes such as leverage, liquidation mechanisms, and price-benchmark governance, and the efficiency analytical framework established in this chapter will provide a solid theoretical foundation for understanding these mechanisms.

References

[1] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417. https://doi.org/10.2307/2325486

[2] Urquhart, A. (2016). The inefficiency of Bitcoin. *Economics Letters*, *148*, 80–82. https://doi.org/10.1016/j.econlet.2016.09.019

[3] Tran, V. L., & Leirvik, T. (2020). Efficiency in the markets of crypto-currencies. *Finance Research Letters*, *35*, Article 101382. https://doi.org/10.1016/j.frl.2019.101382

[4] Kang, H. J., Lee, S. G., & Park, S. Y. (2022). Information efficiency in the cryptocurrency market: The efficient-market hypothesis. *Journal of Computer Information Systems*, *62*(3), 622–631. https://doi.org/10.1080/08874417.2021.1872046

[5] Azar, P. D., Olivas, S., & Sinha, N. D. (2025). The price of processing: Information frictions and market efficiency in DeFi (Staff Report No. 1153). Federal Reserve Bank of New York. https://doi.org/10.59576/sr.1153

[6] Yi, E., Yang, B., Jeong, M., Sohn, S., & Ahn, K. (2023). Market efficiency of cryptocurrency: Evidence from the bitcoin market. *Scientific Reports*, *13*, Article 4789. https://doi.org/10.1038/s41598-023-31618-4

[7] Noda, A. (2021). On the evolution of cryptocurrency market efficiency. *Applied Economics Letters*, *28*(6), 433–439. https://doi.org/10.1080/13504851.2020.1758617

[8] Vidal-Tomás, D., Ibáñez, A. M., & Farinós, J. E. (2019). Weak efficiency of the cryptocurrency market: A market portfolio approach. *Applied Economics Letters*, *26*(19), 1627–1633. https://doi.org/10.1080/13504851.2019.1591583

[9] Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, *30*(5), 15–29. https://doi.org/10.3905/jpm.2004.442611

[10] Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, *1*(1), 41–66. https://doi.org/10.1093/rfs/1.1.41

[11] Chow, K. V., & Denning, K. C. (1993). A simple multiple variance ratio test. *Journal of Econometrics*, *58*(3), 385–401. https://doi.org/10.1016/0304-4076(93)90051-6

[12] Wright, J. H. (2000). Alternative variance-ratio tests using ranks and signs. *Journal of Business & Economic Statistics*, *18*(1), 1–9. https://doi.org/10.1080/07350015.2000.10524842

[13] Box, G. E. P., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332), 1509–1526. https://doi.org/10.1080/01621459.1970.10481180

[14] Brock, W. A., Dechert, W. D., Scheinkman, J. A., & LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, 15(3), 197–235. https://doi.org/10.1080/07474939608800353

[15] Apopo, N., & Phiri, A. (2021). On the (in)efficiency of cryptocurrencies: Have they taken daily or weekly random walks? *Heliyon*, *7*(4), Article e06685. https://doi.org/10.1016/j.heliyon.2021.e06685

[16] Aslan, A., & Sensoy, A. (2020). Intraday efficiency-frequency nexus in the cryptocurrency markets. *Finance Research Letters*, *35*, Article 101298. https://doi.org/10.1016/j.frl.2019.09.013

[17] Zhang, Y., Chan, S., Chu, J., & Sulieman, H. (2020). On the market efficiency and liquidity of high-frequency cryptocurrencies in a bull and bear market. *Journal of Risk and Financial Management*, *13*(1), Article 8. https://doi.org/10.3390/jrfm13010008

[18] Zaremba, A., Bilgin, M. H., Long, H., Mercik, J., & Szczygielski, J. J. (2021). Up or down? Short-term reversal, momentum, and liquidity effects in cryptocurrency markets. *International Review of Financial Analysis*, *78*, Article 101908. https://doi.org/10.1016/j.irfa.2021.101908

[19] Caporale, G. M., & Plastun, A. (2020). Momentum effects in the cryptocurrency market after one-day abnormal returns. *Financial Markets and Portfolio Management*, *34*, 251–266. https://doi.org/10.1007/s11408-020-00357-1

[20] Bianchi, D., Guidolin, M., & Pedio, M. (2023). The dynamics of returns predictability in cryptocurrency markets. *The European Journal of Finance*, *29*(6), 583–611. https://doi.org/10.1080/1351847X.2022.2084343

[21] He, S., Manela, A., Ross, O., & von Wachter, V. (2024). Fundamentals of perpetual futures (Working Paper). https://doi.org/10.48550/arXiv.2212.06888

[22] Zhivkov, P. (2026). The two-tiered structure of cryptocurrency funding rate markets. *Mathematics*, *14*(2), Article 346. https://doi.org/10.3390/math14020346

[23] Caporale, G. M., & Plastun, A. (2019). The day of the week effect in the cryptocurrency market. *Finance Research Letters*, *31*, 258–269. https://doi.org/10.1016/j.frl.2018.11.012

[24] Aharon, D. Y., & Qadan, M. (2019). Bitcoin and the day-of-the-week effect. *Finance Research Letters*, *31*, 415–424. https://doi.org/10.1016/j.frl.2018.12.004

[25] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, *135*(2), 293–319. https://doi.org/10.1016/j.jfineco.2019.07.001

[26] Giudici, P., & Pagnottoni, P. (2019). High frequency price change spillovers in bitcoin markets. *Risks*, *7*(4), Article 111. https://doi.org/10.3390/risks7040111

[27] Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International Economic Review*, *10*(1), 1–21. https://doi.org/10.2307/2525569

[28] Saggu, A., Ante, L., & Kopiec, K. (2025). Uncertain regulations, definite impacts: The U.S. SEC's regulatory interventions and their effects on crypto assets. *Finance Research Letters*, *72*, Article 106413. https://doi.org/10.1016/j.frl.2024.106413

[29] LMAX Group & Macro Hive. (2025, April). *Millisecond reactions to market shocks: From order flow to cross-asset opportunity.* https://www.lmax.com/documents/LMA

XGroup-Macro-Hive-report-2nd-Edition.pdf

[30] Ruan, Q., & Streltsov, A. (2022). Perpetual futures contracts and cryptocurrency market quality (Working Paper). https://doi.org/10.2139/ssrn.4218907

[31] Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *The Journal of Finance*, *50*(4), 1175–1199. https://doi.org/10.1111/j.1540-6261.1995.tb04054.x

[32] Kapar, B., & Olmo, J. (2019). An analysis of price discovery between bitcoin futures and spot markets. *Economics Letters*, *174*, 62–64. https://doi.org/10.1016/j.econlet.2018.10.031

[33] Alexander, C., Choi, J., Park, H., & Sohn, S. (2020). BitMEX bitcoin derivatives: Price discovery, informational efficiency, and hedging effectiveness. *Journal of Futures Markets*, *40*(1), 23–43. https://doi.org/10.1002/fut.22050

[34] Robertson, K., & Zhang, R. (2025). Price discovery in bitcoin spot and futures markets. *Journal of International Money and Finance*, *159*, Article 103415. https://doi.org/10.1016/j.jimonfin.2025.103415

[35] De Blasis, R., & Webb, A. (2022). Arbitrage, contract design, and market structure in bitcoin futures markets. *Journal of Futures Markets*, *42*(3), 492–524. https://doi.org/10.1002/fut.22305

[36] Vidal-Tomás, D., Briola, A., & Aste, T. (2023). FTX's downfall and Binance's consolidation: The fragility of centralized digital finance. *Physica A: Statistical Mechanics and Its Applications*, *625*, Article 129044. https://doi.org/10.1016/j.physa.2023.129044

[37] Allison, I. (2022, November 2). Divisions in Sam Bankman-Fried's crypto empire blur on his trading titan Alameda's balance sheet. *CoinDesk*. https://www.coindesk.com/business/2022/11/02/divisions-in-sam-bankman-frieds-crypto-empire-blur-on-his-trading-titan-alamedas-balance-sheet/

[38] Bloomberg. (2022, November 6). *Binance to liquidate its FTX token (FTT) holdings amid Alameda balance-sheet revelations*. Bloomberg News.

[39] Ray, J. J., III. (2022, November 17). *Declaration of John J. Ray III in support of Chapter 11 petitions and first day pleadings*. United States Bankruptcy Court for the District of Delaware.

[40] Yang, Y., & Ghosh, S. (2022, November 9). Binance backs out of FTX rescue, citing finances, investigations. *Bloomberg*. https://www.bloomberg.com/news/articles/2022-11-09/binance-is-said-to-be-very-unlikely-to-go-ahead-with-ftx-deal

[41] Khoo, Y. L., Leow, S., Polk, N., & Chia, D. (2022, November 17). Blockchain analysis: The collapse of Alameda and FTX. *Nansen*. https://www.nansen.ai/research/blockchain-analysis-the-collapse-of-alameda-and-ftx

[42] Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, 52(1), 35–55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

[43] Augustin, P., Rubtsov, A., & Shin, D. (2023). The impact of derivatives on spot markets: Evidence from the introduction of bitcoin futures contracts. *Management Science*, *69*(11), 6752–6776. https://doi.org/10.1287/mnsc.2023.4900

[44] John, K., Li, J., & Liu, R. (2024). Pricing and arbitrage across 80 cryptocurrency exchanges (Working Paper). New York University. https://doi.org/10.2139/ssrn.4816710

Part VI
Arbitrage and Anomalies

Core question: If markets are not fully efficient, where do systematic profit opportunities lie? How do arbitrageurs eliminate them, and why do they sometimes fail?

Arbitrage and the Transmission Mechanisms of Market Efficiency

On December 18, 2024, the Federal Open Market Committee (FOMC) of the Federal Reserve released its final interest rate decision of the year. Within seconds of the release, the microstructure of the digital asset market displayed its distinctive complexity. The price of Binance's Bitcoin perpetual futures fell 1.2%, Coinbase's spot price fell only 0.7%, and on the decentralized derivatives protocol Hyperliquid an oracle-update delay meant that the price had not yet fully caught up. This rapid divergence caused the basis of the perpetual futures relative to spot to widen abruptly from +15 basis points (bps) to +65 bps. Within a short window after the divergence emerged, three types of arbitrage mechanism engaged simultaneously. High-frequency traders' automated systems detected the price gap between Binance and Coinbase at microsecond resolution and executed cross-market hedges. Systematic quantitative funds detected the abnormally widened basis between spot and perpetual futures, which triggered signals to open basis-arbitrage positions. And on-chain-native MEV (maximal extractable value) searchers captured the 42-bps gap between Hyperliquid's oracle price and the real-time price on centralized exchanges. Starting from their respective areas of specialization, these three forces converged on a single objective: eliminating the spread. Over the following tens of seconds, the basis was gradually compressed back to +22 bps; by the end of the first minute, prices across the three markets had realigned tightly. The market dynamics of these 30 seconds crystallize the central questions of this chapter: Who drives the rapid convergence of prices, and what does their profit-seeking behavior mean for the efficiency of the market as a whole? (This passage is a stylized scenario reconstructed from real market mechanisms; the specific figures are illustrative and do not constitute a precise observational record of any particular event.)

These questions carry us from the descriptive analysis of Chapter 15 to the causal analysis of this chapter. Using an "efficiency spectrum" framework, Chapter 15 systematically assessed the level of efficiency in the digital asset market and revealed its pronounced heterogeneity across assets, platforms, and time. The central task of this chapter is to dissect the mechanisms by which market efficiency is generated, maintained, and self-repaired. In a crypto market that lacks a central clearinghouse, unified regulation, and standardized settlement infrastructure, the traditional financial definition of arbitrage as the "capture of riskless profit" proves inadequate: it cannot explain why spreads persist, nor can it fully capture the systemic function that arbitrage performs in this ecosystem. This chapter therefore proposes the arbitrage infrastructure

hypothesis: arbitrageurs are not merely passive correctors of price deviations but active providers of market infrastructure. Their profit-seeking behavior, in effect, performs three key infrastructure functions: along the price-linkage dimension it partly serves as a cross-market price-transmission bus (but does not assume the concentrated management of counterparty credit risk or the default-fund functions of a traditional central counterparty); through continuous basis arbitrage it provides the stabilizing mechanism that anchors perpetual-futures prices, substituting for the delivery at maturity of traditional futures; and it integrates the local liquidity dispersed across hundreds of isolated exchanges and protocols into a virtual global liquidity pool.

Building on this theoretical perspective, the chapter analyzes the arbitrage ecosystem of the digital asset market at three levels. Theoretically, it re-examines the shift of arbitrage from the capture of "riskless profit" to the provision of "infrastructure services" and dissects the multi-tiered ecological structure formed by high-frequency traders, quantitative funds, protocol-level arbitrageurs, MEV searchers, and individual traders. Mechanistically, it decomposes the sources of return and the efficiency contributions of six core mechanisms: spot-perpetual basis arbitrage, cross-exchange arbitrage, spot-futures-perpetual triangular arbitrage, funding rate arbitrage, cross-asset statistical arbitrage, and CEX-DEX cross-tier arbitrage. At the level of transmission, it constructs an "arbitrage-efficiency transmission model" that explains how micro-level arbitrage behavior propagates to macro-level market efficiency through price calibration, information aggregation, and liquidity integration, thereby forming a causal bridge to the efficiency diagnosis of Chapter 15.

16.1 The theoretical foundations of arbitrage

In financial theory, arbitrage is regarded as the core mechanism driving market efficiency—a persistent price-correcting force that pushes deviating asset prices back toward their intrinsic value. Yet when this classical theoretical framework is applied to the emerging domain of digital assets, a series of striking *anomalies* comes into view. The crypto market—a distinctive ecosystem of globalized, fragmented, around-the-clock (24/7) trading and complex derivatives—not only challenges the conventional understanding of riskless arbitrage but forces a reconsideration of the fundamental role that arbitrageurs play in the market. The systematic violation of the classical law of one price in the crypto market is the starting point for rethinking the nature of arbitrage. By replacing delivery at maturity with a funding rate mechanism, perpetual futures fundamentally alter the logic of arbitrage, transforming it from a one-off convergence-at-maturity trade into an ongoing position-management activity. This paradigm shift provides the micro-mechanistic foundation for the chapter's central theoretical contribution: the arbitrage infrastructure hypothesis.

16.1.1 The persistence paradox of the law of one price

One of the most basic and intuitive principles in economics is the law of one price. It holds that in an efficient market free of trading frictions—such as transaction costs, taxes, or capital controls—any homogeneous asset should have only one price, wherever

it is traded. If a price gap appears, rational arbitrageurs profit by buying in the low-price market and selling in the high-price market, and their actions quickly erase the gap and restore equilibrium. In mature traditional financial markets, such as equity or foreign exchange markets, the law of one price holds to a high degree almost all the time. Brief deviations occasionally occur, but they are typically eliminated within milliseconds or seconds by high-frequency trading algorithms.

In the crypto asset market, however, the situation is entirely different. There, the law of one price is not a stable empirical regularity but a theoretical assumption that is systematically and persistently violated. Early academic research on crypto markets quickly identified this striking phenomenon. Makarov and Schoar (2020) [1] were the first to systematically document the vast arbitrage opportunities pervasive in crypto markets. Analyzing Bitcoin trading data from dozens of exchanges worldwide, they found that cross-exchange spreads were not only large but persistent. They reported that cross-border spreads were far larger than spreads within a single country: about 3% between the United States and Europe, about 10% between Japan and the United States, and about 15%—peaking above 40%—between Korea and the United States. For example, the "kimchi premium" of the Korean market relative to the U.S. market once exceeded 40% at the peak of the bull market in late 2017 and early 2018, meaning that buying Bitcoin in Korea cost nearly half again as much as buying it in the United States.

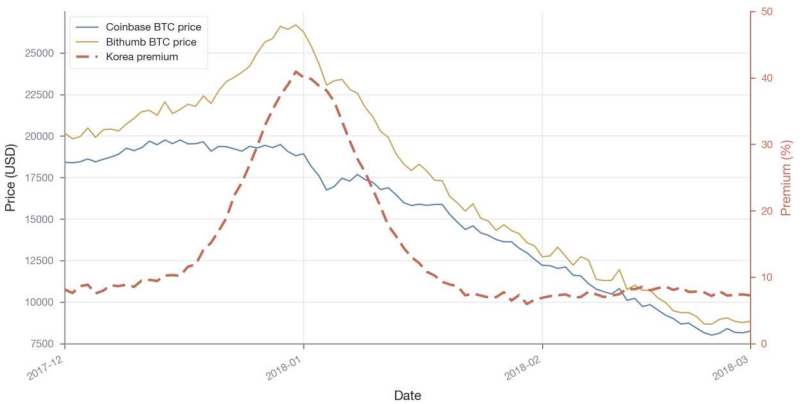


Figure 16-1. Cross-exchange Bitcoin price spreads, December 2017 to February 2018 (this figure focuses on the Korea-versus-U.S. *kimchi premium*, which peaked above 40%; the cross-border spreads of about 3% for U.S.-Europe and about 10% for Japan-U.S. are discussed in the text; data source: Makarov & Schoar 2020 [1])

These spreads were not fleeting price deviations but structural phenomena that could persist for hours, days, or even weeks. Makarov and Schoar estimated that, over their sample period, the theoretically capturable arbitrage profit exceeded $75 million per day on average. This raises a central paradox: if such a large and substantial space for riskless profit existed, why did rational arbitrageurs fail to eliminate it quickly, as classical theory would predict?

This "persistence paradox" is the starting point for understanding arbitrage mecha-

nisms in crypto markets. It strongly suggests that treating arbitrage in crypto markets simply as the capture of riskless profit is an oversimplification. The answer appears to lie not in arbitrageurs being insufficiently clever or fast, but in the fact that arbitrage in crypto markets faces frictions and costs of a magnitude that traditional markets never encountered. These frictions may arise from strict capital controls across countries, which make it difficult for arbitrageurs to move fiat-currency profits freely from one market to another; from the weak clearing and settlement links between exchanges, which mean that transferring assets across platforms entails time delays and counterparty risk; or from the execution risk and liquidity evaporation that arbitrage exchanges face during sharp price swings. The persistent spreads we observe may therefore be not a sign of complete market inefficiency but rather a monetized pricing of these real "friction costs" that impede the free flow of arbitrage capital. This requires us to shift the analytical framework from an idealized frictionless assumption toward an arbitrage-engineering perspective that takes real-world constraints into account. This perspective aligns closely with the classic analysis of the limits of arbitrage by Shleifer and Vishny (1997) [2]: real-world arbitrageurs are not ideal agents with unlimited capital and zero risk but economic actors subject to multiple constraints—funding constraints, principal-agent problems, and noise-trader risk—so that even in the face of a clear price deviation, their corrective capacity is systematically limited by these constraints. The empirical study of equity-market arbitrage by Mitchell, Pulvino, and Stafford (2002) [3] further corroborates this point: even in mature, highly liquid markets, arbitrageurs are often unable to eliminate known pricing deviations in time because of capital constraints and execution risk, allowing spreads to persist.

16.1.2 From *point convergence* to *flow anchoring*

As noted above, the core innovation of perpetual futures is that they replace the delivery-at-maturity regime of traditional futures with a funding rate mechanism, fundamentally changing the logic of arbitrage. The no-arbitrage pricing of traditional futures relies on the cost-of-carry model and on the forced price convergence at maturity, which we term *point convergence*. Through the funding rate—a continuous cash-flow payment mechanism—perpetual futures extend price anchoring from a discrete point in time into a continuous, dynamic process, forming a new paradigm we call *flow anchoring*. Under this paradigm, arbitrage is no longer a one-off convergence-at-maturity trade but an ongoing position-management activity: by constructing a delta-neutral position, the arbitrageur earns the profit from basis convergence while continuously collecting the funding rate as compensation for risk.

He et al. (2022) [4] established a rigorous no-arbitrage pricing framework for perpetual futures and revealed a key fact: the deviation between the perpetual-futures price and the theoretical no-arbitrage price is not an anomaly signaling market failure but an equilibrium outcome that reflects the true cost of arbitrage. They found an annualized mean absolute deviation from no-arbitrage prices of as much as 60% to 90%; He et al. emphasize that this magnitude of deviation is significantly larger than the deviations recorded for comparable products in traditional money markets. This measure captures the cumulative deviation of the price from the no-arbitrage benchmark (annualized); its economic meaning can be further understood as the systematic difference between

actual financing costs and the theoretical no-arbitrage-implied financing cost, reflecting market frictions and risk premia that the model does not fully capture. Such a large magnitude of deviation is rare in traditional financial markets but is the norm in the crypto world. It reflects the various risks and costs that arbitrageurs must bear when executing basis arbitrage, including transaction fees, slippage, funding costs, and the risk of forced liquidation in extreme conditions. The expected return that arbitrageurs demand must be sufficient to compensate for these costs and risks, so a significant and persistent basis and funding rate is precisely the natural state in which the market reaches equilibrium.

From a macro-financial perspective, this deviation can be decomposed into at least three components. The first is the macro financing cost that reflects the global dollar interest rate environment; during a Federal Reserve tightening cycle, the opportunity cost of arbitrage capital rises, and the equilibrium basis and rate levels increase accordingly. The second is the leverage premium endogenous to the crypto market; during bull markets, participants' excess demand for leverage pushes up the perpetual-futures premium and the funding rate, and this component is strongly procyclical. The third is the purely structural friction cost, including transaction fees, cross-platform fund-transfer delays, and margin lock-up—inherent frictions that do not vary with the macro cycle. Attributing such a large deviation entirely to structural frictions while ignoring the transmission effect of macro-financial conditions could lead to a misjudgment of market efficiency.



Figure 16-2. Annualized deviation of perpetual futures from no-arbitrage prices (Data source: He et al. 2022 [4])

Figure 16-2 shows that the annualized mean absolute deviation of perpetual futures from no-arbitrage prices remains persistently in the 60%–90% range, far higher than for comparable products in traditional markets—a monetized expression of the costs and risks that arbitrageurs bear.

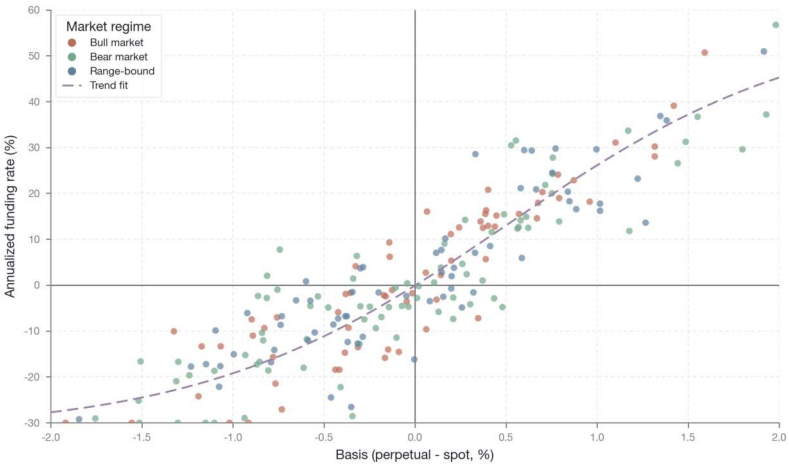


Figure 16-3. The relationship between the funding rate and the basis (Data source: He et al. 2022 [4])

Figure 16-3 depicts the pronounced positive correlation between the funding rate and the perpetual-futures basis: when the basis widens (the perpetual premium rises), the funding rate rises in step, reflecting that the financing cost of maintaining a leveraged long position grows with the premium—the empirical basis of the *flow anchoring* mechanism.

The emergence of perpetual futures thus marks a paradigm shift in arbitrage theory. It forces us to move our focus from the convergence of prices at some future "point" to the anchoring relationship among prices in the present, driven by a cash "flow." Under this new paradigm, arbitrageurs are no longer mere discoverers of price deviations; they effectively perform the dual functions of liquidity provision and risk management. By continuously managing arbitrage positions, they provide the perpetual-futures market with a critical price-stabilization mechanism, earning a corresponding risk premium in return. This lays a solid micro-mechanistic foundation for the arbitrage infrastructure hypothesis that we propose next.

16.1.3 The arbitrage infrastructure hypothesis

Building on the foregoing analysis of persistent spreads in crypto markets and of the *flow anchoring* mechanism of perpetual futures, we can now propose an overarching theoretical framework in this chapter: the arbitrage infrastructure hypothesis (AIH). The hypothesis aims to redefine the core role of arbitrageurs in the digital asset market ecosystem and to provide a unified theoretical lens for understanding the various arbitrage activities examined later in the chapter.

The central claim of the arbitrage infrastructure hypothesis is this: in a crypto market that lacks a central clearinghouse and whose regulatory frameworks are still evolving rapidly and have not yet converged on a globally unified standard (although the EU's Markets in Crypto-Assets (MiCA) regulation has been implemented in phases—the stablecoin rules applying from June 2024 and the rules for other issuers and service

providers from December 30, 2024, with a transition period running to 2026—so that regulatory frameworks across major jurisdictions are taking shape at an accelerating pace), the role of the arbitrageur has evolved from the "price corrector" of classical financial theory, which reacts passively to price deviations, into an "infrastructure provider" that actively supplies core functions to the market. Their profit-seeking behavior, in effect, builds and maintains three pillar functions on which the market's efficient operation depends:

The first function is the cross-market clearing bus. In the traditional financial system, institutions such as the Depository Trust & Clearing Corporation or central counterparty clearinghouses ensure that trades among different exchanges and market participants can be cleared and settled smoothly and efficiently. The crypto market inherently lacks such centralized institutions, and the individual exchanges are independent of one another. Arbitrageurs—especially high-frequency trading firms—by simultaneously buying and selling across different exchanges, effectively act as the price-transmission hub connecting these independent markets; their capital and technology constitute a cross-market value-transfer network that, along the price-linkage dimension, partly assumes a function that should be provided by a central institution (but does not assume that institution's concentrated management of counterparty credit risk or its default-fund functions; see Proposition 1 below).

The second function is the price-anchoring mechanism. Because of their no-maturity design, perpetual futures lack the convergence-at-maturity mechanism of traditional futures. By continuously hedging between the spot and perpetual-futures markets, basis arbitrageurs anchor the perpetual-futures price around the spot price; their collective action constitutes a dynamic, funding-rate-based, continuous anchoring mechanism that substitutes for delivery at maturity.

The third function is the liquidity-integration network. Liquidity in the crypto market is highly fragmented, dispersed across hundreds of exchanges worldwide, thousands of trading pairs, and two starkly different tiers—centralized and decentralized. Arbitrage activity—particularly cross-exchange arbitrage, cross-asset statistical arbitrage, and CEX-DEX arbitrage—serves to integrate this dispersed liquidity. When a large buy order appears on one exchange, arbitrageurs buy on other exchanges with deeper liquidity or lower prices and sell on the first exchange, thereby transmitting global liquidity to the market where demand is concentrated and forming what this chapter calls the virtual global liquidity pool (the market-wide depth, aggregated from the local liquidity of individual exchanges through arbitrageurs' cross-market transmission, that a single trader can effectively access; at the order-book level it also manifests as a virtual consolidated order book).

Although these three functions derive from similar cross-market trading behavior, they operate on different dimensions and are not equivalent to one another: the price-transmission bus operates on the net transfer of value or positions across participants (the value-flow dimension), the price-anchoring mechanism operates on the continuous-in-time convergence between perpetuals and spot (the time dimension), and the liquidity-integration network operates on the cross-exchange aggregation of available depth (the depth dimension).

The central claim above can be further formalized as the following testable propositions:

Proposition 1 (infrastructure-function substitutability along the price dimension): In a market structure that lacks a central clearinghouse, the collective behavior of arbitrageurs partly substitutes for the functions of a traditional central clearinghouse along price dimensions such as price calibration, liquidity integration, and cross-market value transfer. It must be stressed that this substitution is confined to the price dimension: the concentrated management of counterparty credit risk and the default-fund guarantee that a traditional central counterparty clearinghouse also provides cannot be substituted by dispersed arbitrage behavior. When an exchange defaults (as in the FTX collapse of 2022), arbitrageurs cannot serve as a credit-risk buffer; on the contrary, their cross-platform position exposure may make them transmitters and amplifiers of risk.

Proposition 2 (the cost-function nature of the equilibrium spread): Given the macro interest rate environment r and the competitive density of arbitrageurs N, the equilibrium level of the cross-market spread s^* is a monotonically increasing function of the cost c of providing arbitrage-infrastructure services (including transaction costs, capital costs, technology investment, and risk premia), that is, $s^* = f(c \mid r, N)$, $\partial f/\partial c > 0$. We retain r and N explicitly as conditioning variables because the macro financing cost and the competitive density also affect s^* (see Sections 16.6.3 and 16.11); s^* should therefore not be reduced to a single-variable function of c.

Corollary 2.1: Any technological or institutional innovation that reduces the arbitrage friction cost c will reduce the equilibrium spread s^*—that is, reduce the cost the market pays for the infrastructure service. Two kinds of efficiency must be distinguished. This corollary bears directly on *spread efficiency* (the law-of-one-price dimension—how wide or narrow the spread is), whereas *informational efficiency* (the speed and completeness with which prices reflect information) must be measured independently using metrics orthogonal to the spread, such as information share and the speed of price discovery. Where the remainder of this chapter speaks of "improving market efficiency," it means the former when spreads are at issue and the latter when price discovery is at issue; the two should not be conflated.

Proposition 3 (the compensatory nature of returns): The persistent positive returns that arbitrageurs earn are not "abnormal returns" signaling market inefficiency but the reasonable, risk-adjusted economic compensation they receive as infrastructure providers.

The AIH framework allows us to understand many features of the crypto market from a fresh perspective. It stands in sharp contrast to classical arbitrage theory, as shown in Table 16-1:

Dimension	Arbitrage infrastructure hypothesis (AIH)	Classical arbitrage theory
Role of the arbitrageur	Infrastructure provider	Price corrector
Meaning of the spread	Equilibrium price of an infrastructure service	Sign of market inefficiency
Core assumption	Frictions and costs present	Frictionless / perfectly competitive
Market structure	Fragmented / no central clearing	Unified, centrally cleared market
View of efficiency	Dynamic-equilibrium efficiency	Perfectly efficient market attainable
View of risk	Arbitrage has costs and risks	Arbitrage is riskless

Table 16-1. Classical arbitrage theory versus the arbitrage infrastructure hypothesis (Data source: constructed by the author)

This framework matters, first, because it explains the "persistence paradox" raised in Section 16.1.1. The persistently positive arbitrage returns in crypto markets are not "costless returns" signaling market inefficiency but the reasonable economic reward that arbitrageurs earn as infrastructure providers for bearing risk and investing capital and technology. The spreads we observe are, in essence, the fees the market pays for these "infrastructure services." Second, the AIH provides a coherent logic for the discussions in the chapters that follow. This framework resonates with the theoretical analysis of Gromb and Vayanos (2010) [5] on the relationship between arbitrageurs' capital constraints and financial stability: when the capital base of the arbitrage infrastructure is shocked, its market-stabilizing function can reverse. The arbitrage risks that Chapter 17 will examine can, under the AIH framework, be reinterpreted as "fragilities of the infrastructure," while the arbitrage failures and market anomalies that Chapter 18 will analyze can be seen as "local or systemic ruptures of the infrastructure." Finally, the AIH directly links micro-level arbitrage behavior to the macro-level market-efficiency-spectrum framework of Chapter 15: the efficiency level of a market is, to a large extent, a function of the completeness, robustness, and cost of the arbitrage infrastructure behind it. The more complete the arbitrage infrastructure, the lower its service cost (the equilibrium spread) and the higher the market's efficiency.

16.1.4 The Grossman-Stiglitz paradox

In discussions at the frontier of arbitrage theory, it is worth introducing a classic proposition in financial economics—the Grossman-Stiglitz paradox—and examining its distinctive expression in crypto markets. This paradox provides an important theoretical precedent for the AIH sub-claim that "the equilibrium spread is persistently positive" (namely Propositions 2 and 3 below) and reveals the intrinsic tension in market efficiency and the nature of dynamic equilibrium. The paradox does not, by itself, directly support the stronger core metaphor that "arbitrageurs are infrastructure providers."

Grossman and Stiglitz (1980) [6] posed a central question: if a market is fully informationally efficient—that is, if all information is already fully reflected in prices—then it would be unprofitable for any individual or institution to spend resources gathering and analyzing information. But if no one has an incentive to gather information, how

could prices come to reflect all information and thereby become efficient? The paradox points to a conclusion: a fully informationally efficient market is logically impossible. A market must contain some degree of "inefficiency"—prices must deviate from fundamental value—so as to provide a sufficient return to those who bear the cost of acquiring information, keeping them engaged in information discovery.

We can now extend this classic paradox, by analogy, to the arbitrage context of crypto markets. A caveat must be drawn first—the two are structurally isomorphic but concern different objects: the original Grossman-Stiglitz proposition concerns traders' acquisition of "costly private fundamental information" and whether prices can fully reveal that information, and its equilibrium spread arises from an information rent; the arbitrage in this chapter, by contrast, exploits "publicly observable cross-market spreads," and its equilibrium spread arises from the cost of infrastructure services. Under this analogy, "information gathering" corresponds to "arbitrage activity" and "information cost" corresponds to the "infrastructure-service cost" of the AIH framework. The Grossman-Stiglitz-style paradox for crypto markets can be stated as follows.

If arbitrage activity were perfectly efficient and could rapidly eliminate all cross-market and cross-asset spreads to zero, then arbitrageurs would earn nothing—because their return (the spread) would be zero while the cost they incur to execute arbitrage (transaction fees, funding costs, technology investment, risk-bearing, and so on) would be positive. In that case, rational arbitrageurs would cease activity and exit the market. Yet once arbitrageurs—the "infrastructure providers"—collectively exit, the market's clearing bus, price anchoring, and liquidity-integration functions would fail, and spreads would inevitably reappear and widen. When spreads widen enough to cover arbitrage costs and offer an attractive risk-adjusted return, new or returning arbitrageurs would again be drawn into the market. This process forms a never-ending dynamic cycle.

The equilibrium state of this cycle is not a "perfectly efficient market" with zero spreads but a "dynamic-equilibrium market" in which spreads are maintained at some persistently positive level. At this equilibrium point, the return the marginal arbitrageur can earn just covers all the costs of executing arbitrage activity. The level of this equilibrium spread defines the practical ceiling on that market's efficiency. It is at once the minimum compensation the market must pay to keep arbitrageurs providing infrastructure services and the implicit cost that market participants must bear to obtain an integrated, relatively efficient trading environment.

This analytical conclusion provides a theoretical foundation for connecting to the chapters that follow. When arbitrage costs (the infrastructure-service cost of the AIH) rise for some reason—for example, tighter regulation that strengthens capital controls, network congestion that increases transaction delays, or violent market swings that enlarge risk exposure (all of which Chapter 17 will discuss in detail)—the equilibrium spread must rise correspondingly to maintain the incentive for arbitrageurs. A systematic widening of spreads, that is, a decline in market efficiency, is the condition under which various market anomalies (to be examined in Chapter 18) arise and persist. The crypto version of the Grossman-Stiglitz paradox thus tells us that the spread is not an inefficiency to be eliminated entirely but a key indicator of the operating state of the market ecosystem; it measures the operating cost of arbitrage—the market's core

infrastructure—and reflects the efficiency frontier and the potential risk level of the entire market.

16.2 The structure of the arbitrage ecosystem

If the arbitrage infrastructure hypothesis proposed in Section 16.1 provides a theoretical lens for understanding the micro-foundations of crypto-market efficiency, the aim of this section is to analyze further the entity structure that constitutes this infrastructure—a complex ecosystem composed of diverse participants, clearly stratified and dynamically evolving. In the perpetual-futures market, arbitrage has evolved from the individual behavior of a handful of speculators into a highly specialized pattern of collective cooperation and competition with a clear division of labor. Participants of different backgrounds, drawing on their divergent endowments in capital scale, technological latency, risk appetite, and strategy complexity, occupy different tiers of the ecological pyramid and form a complex network that is at once mutually dependent and fiercely competitive. The collective action of this network jointly constitutes the key infrastructure functions—cross-market clearing, price anchoring, and liquidity integration—described by the AIH.

The analysis of this ecosystem begins with the tiered structure of participants. A map of participants comprising five main tiers (Tier 0 through Tier 4) identifies the distinct roles and niches ranging from protocol-level participants to individual traders. As a representative case of "protocol-level arbitrage," the Ethena protocol reveals how this crypto-native phenomenon has reshaped the capital structure and dynamics of the entire ecosystem. The interaction network among participants combines tiered competition with functional complementarity, and the systemic risks latent within it warrant attention. Placing this ecosystem in a framework of historical evolution allows us to trace its path from the early stage to the institutionalized era and to consider the transformations that an AI-driven new paradigm may bring.

16.2.1 A five-tier map of participants

The arbitrage ecosystem of perpetual futures is not a flat competitive structure but exhibits clearly differentiated tiers. According to capital scale, technological capability (especially execution speed), strategy complexity, and market influence, the main participants can be grouped into a five-tier pyramid model from Tier 0 to Tier 4. This model reveals not only the core competitiveness of the different participants but also the different divisions of labor they assume in playing the role of "arbitrage infrastructure." Note that Tier 0 (protocol-level arbitrageurs) is a special tier that emerged only in recent years; following an evolutionary logic from traditional to emerging players, this section discusses it after Tiers 1 through 4. The "0" in its label signals that its capital scale and systemic influence are the most prominent, not its order of appearance.

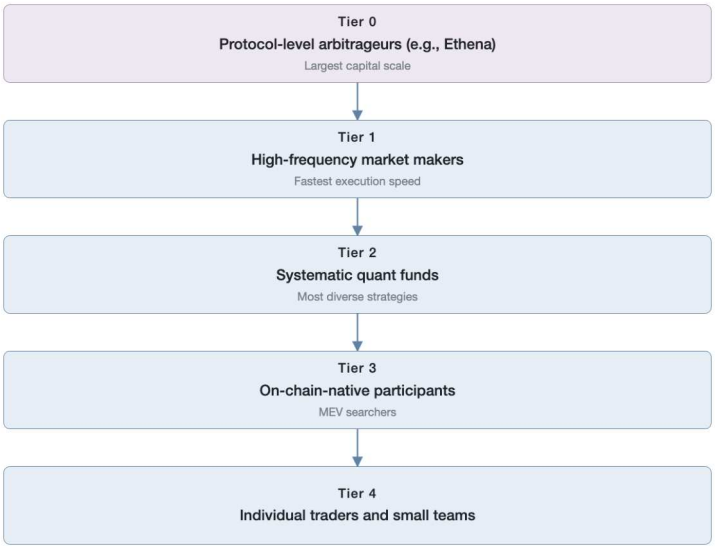


Figure 16-4. The five-tier participant structure of the arbitrage ecosystem (Data source: a conceptual schematic constructed by the author from public market data)

Figure 16-4 shows the vertical stratification of the arbitrage ecosystem from Tier 0 to Tier 4: the higher the tier, the more concentrated the capital and the fewer the participants. The top tiers, drawing on their capital and technology advantages, capture the most stable and largest arbitrage opportunities, while the bottom tiers compete in more crowded, thin-margin domains.

Tier 1: High-frequency market makers

Tier 1 sits at the top of the traditional arbitrage hierarchy and consists mainly of high-frequency trading firms with deep traditional-finance backgrounds, such as Jump Trading, Wintermute, and GSR. Their capital typically runs into the hundreds of millions of dollars or more, and their core advantage is extremely low trading latency. Through server colocation at exchanges and optimized network connectivity, they can execute trades with latencies on the order of microseconds to milliseconds. Their core strategy is to capture fleeting cross-exchange spreads. For example, when the price of BTC perpetual futures on Binance is a few basis points higher than the spot price on Coinbase, Tier 1 algorithms quickly short the perpetual on Binance while buying spot on Coinbase, locking in a nearly riskless, direction-neutral spread. The profit per operation is thin, but through enormous volume and high frequency it accumulates into a considerable return. Under the AIH framework, Tier 1 participants act as the core cross-market price-clearing hub; their near-real-time arbitrage helps bring prices across the world's major exchanges into close alignment, forming the micro transmission network of the virtual global liquidity pool.

Tier 2: Systematic quantitative funds

Next come the Tier 2 systematic quantitative funds. These participants typically command capital ranging from tens of millions to hundreds of millions of dollars. Although they cannot match Tier 1 on execution speed (they operate at second-to-minute latencies), their core advantage lies in the depth and complexity of their strategies. Once Tier 1 has fully consumed the simple microsecond-scale spread opportunities, Tier 2 focuses on more durable and complex arbitrage opportunities. Their strategy libraries include, but are not limited to, spot-perpetual basis arbitrage, funding rate arbitrage, calendar-spread arbitrage (such as the spread between perpetual futures and quarterly futures), and pairs trading based on statistical models. These strategies place relatively low demands on speed but higher demands on the accuracy of market models, the granularity of risk management, and capital efficiency. For example, a Tier 2 fund might construct a complex portfolio, simultaneously going long a set of perpetual futures with expected positive funding rates and shorting another set with expected negative rates, so as to earn a stable rate differential. Within the AIH framework, Tier 2 acts as a cross-maturity liquidity provider and a maintainer of relative pricing; it connects contract markets across different maturities and instruments, ensuring pricing consistency within the entire derivatives ecosystem.

Tier 3: On-chain-native participants

Tier 3 represents an innovative force unique to the crypto market, comprising mainly MEV searchers and liquidation bots. Their capital may range from millions to tens of millions of dollars, but their core arena is the blockchain itself, and their execution speed is measured in "block time." MEV searchers use complex algorithms to scan the mempool for and exploit arbitrage opportunities in transaction ordering—for example, when a spread exists between an automated market maker (AMM) pool on a decentralized exchange (DEX) and a centralized exchange (CEX), capturing profit through "sandwich attacks" or front-running. Liquidation bots monitor lending protocols (such as Aave and Compound) and DEX margin trading for under-collateralized positions; once one is found, they trigger the liquidation process, buying the collateral at a discount and repaying the debt to earn the liquidation penalty. Under the AIH framework, Tier 3 participants are the key transmission nodes connecting the CEX and DEX tiers of the market; they rapidly propagate and correct pricing deviations from the on-chain world and are also the actual executors of the risk-control mechanisms (such as liquidation) of the on-chain financial system.

Tier 4: Individual traders and small teams

At the base of the pyramid is the large population of Tier 4 participants, including sophisticated individual traders and small proprietary trading teams. Their capital is typically below a million dollars, their execution speed depends on commercial trading terminals or manual operation, and their reaction time runs from minutes to hours. Because of their disadvantages in capital and technology, they cannot compete with Tiers 1 and 3, and their strategies typically center on simplified versions of Tier 2 strategies—for example, manually executed spot-perpetual basis arbitrage, or capturing large funding rate differentials across exchanges. Although the market influence of any single Tier 4 participant is negligible, their collective behavior has a significant aggregate effect on the market, especially in driving the funding rate back toward its

mean. When the funding rate is too high, large numbers of individual traders pile into basis arbitrage, and their collective selling pressure helps dampen the rate. In the AIH framework, therefore, Tier 4 serves as an "auxiliary rate stabilizer" and forms the broad participatory base of the market's self-regulating mechanism.

Tier 0: Protocol-level arbitrageurs

Finally, and most distinctively, at the apex of the pyramid is Tier 0: protocol-level arbitrageurs. This is an entirely new type of participant that has emerged only in recent years, whose outstanding representative is Ethena. Such a participant is itself a decentralized protocol that pools massive amounts of user capital (which can reach billions or even tens of billions of dollars) through smart contracts and executes, programmatically and at large scale, the arbitrage strategies that formerly belonged to Tier 2. The emergence of Tier 0 marks a paradigm leap in arbitrage from "individual or institutional strategy" to "protocolized infrastructure." It "productizes" and "democratizes" complex arbitrage strategies, allowing any user to share indirectly in arbitrage returns by holding a protocol token (such as USDe). This model substantially extends the capital frontier of arbitrage, from the balance sheets of a few professional institutions to the liquidity of the entire DeFi world. Under the AIH framework, Tier 0 is an "infrastructure provider" that has itself become a huge, systemic tool for price anchoring and liquidity integration. This scale, however, also brings new problems, which we examine in detail in the Ethena case in the next section.

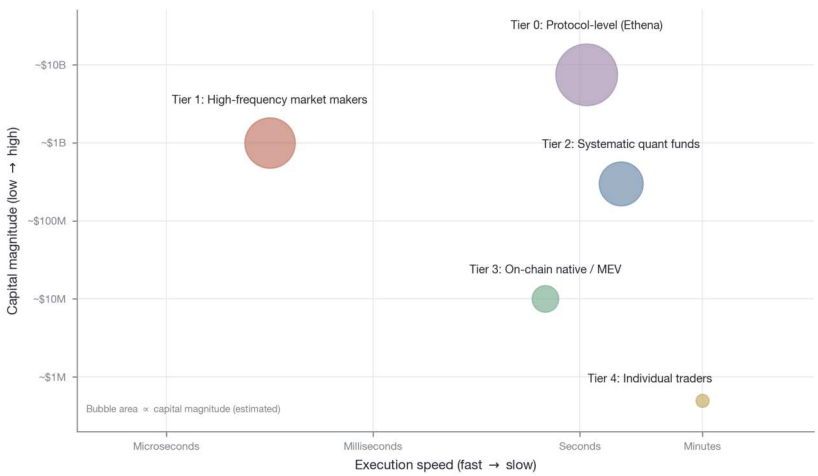


Figure 16-5. The distribution of capital scale and execution speed across the five tiers of arbitrage participants (Data source: estimated by the author from public disclosures of firms such as Jump Trading and Wintermute and from industry reports; data as of the fourth quarter of 2025)

Figure 16-5 uses a bubble chart (bubble size represents capital scale, the vertical axis represents execution speed) to depict each tier's niche: Tier 1 is fastest but not the largest in capital, Tier 0 far exceeds the other tiers in capital but executes more slowly, and Tiers 2, 3, and 4 each occupy their own position in the trade-off between capital and speed.

16.2.2 Ethena and arbitrage-as-a-service

In the evolution of the crypto arbitrage ecosystem, the rise of the Ethena protocol is a representative case. It not only created the synthetic dollar USDe at an unprecedented scale but, more importantly, pioneered an entirely new business model: "arbitrage-as-a-service." By packaging an institutional-grade spot-perpetual basis-arbitrage strategy into a permissionless, composable on-chain protocol, Ethena transformed an arbitrage strategy once confined to a few professional institutions into yield infrastructure open to all DeFi users. This paradigm breakthrough provides a representative empirical case for understanding the arbitrage infrastructure hypothesis.

Ethena's core mechanism is conceptually simple. Users deposit assets such as ETH, BTC, or liquid staking tokens into the protocol; the protocol automatically uses the deposited assets as collateral and opens an equivalent short position in the perpetual-futures market on a CEX (such as Binance or Bybit), constructing a delta-neutral portfolio. The captured funding rate income, net of operating costs, is distributed to sUSDe holders. In essence, Ethena has built a large-scale automated basis-arbitrage system, and USDe is the tokenized claim on the stable cash flow that this system generates.

Although Ethena manages user deposits and withdrawals through on-chain smart contracts, its core operation—opening and managing perpetual-futures short positions on CEXs—relies on off-chain custodians for execution. This hybrid architecture of "on-chain interface plus off-chain execution" means that Ethena is not a fully decentralized protocol, and its custody chain contains at least three critical risk nodes: the operational security and key-management risk of the custodian, the counterparty credit risk of the CEX (the lesson of the FTX collapse remains fresh), and the operational risk of the settlement channel between the custodian and the CEX. In addition, the CEX where Ethena's positions reside could unilaterally freeze accounts, change margin rules, or suspend trading in extreme conditions, and any such event would instantly invalidate the protocol's delta-neutral assumption.

Ethena's structural innovation as a Tier 0 participant manifests in three respects. First, it "productizes" and "democratizes" the arbitrage strategy, allowing any user to share indirectly in arbitrage returns by holding a protocol token, substantially extending the capital frontier of arbitrage from the balance sheets of a few professional institutions to the liquidity of the entire DeFi world. Second, it operates within DeFi's composability framework: USDe and sUSDe can be embedded in other DeFi protocols as collateral, liquidity, or yield instruments, forming a secondary distribution network for arbitrage returns across the DeFi ecosystem. Third, its protocolized mode of operation makes strategy execution transparent and auditable, reducing the principal-agent risk of the traditional fund model.

Viewed through a behavioral-finance lens, the growth curve of Ethena's total value locked (TVL) is not driven entirely by rational capital allocation. The early sUSDe yield of over 20% triggered a pronounced fear-of-missing-out (FOMO) effect in the crypto community, and the spread of the yield narrative on social media accelerated the pace of capital inflows, quite possibly far faster than rational expectations would support. Investors tended to anchor the early high yield as the "normal" level and

displayed lagged exit behavior when the yield declined as the strategy became crowded. The 2022 collapse of Terra/UST, which followed a similar high-yield narrative, offers a cautionary precedent for these behavioral dynamics.

This model, however, also introduces a new dimension of risk. When a single protocol grows to command a significant share of the market's total open interest, it shifts from a price "taker" to a price "maker," and its own behavior can trigger positive or negative feedback loops in the market. This "Ethena effect"—how the protocol's own expansion systematically affects the very target of its arbitrage (in essence, the reflexive suppression that large-scale arbitrage exerts on its own source of returns, a protocol-scale manifestation of arbitrage's self-limiting nature)—is analyzed quantitatively in Section 16.4.3.

Moreover, the regulatory characterization of the "arbitrage-as-a-service" model remains deeply uncertain. An entity that raises funds from the general public, executes an investment strategy, and distributes returns may constitute a regulated investment product under the securities laws of major jurisdictions. "Protocolization" does not equal "deregulation," and this legal risk poses a potential challenge to the long-term sustainability of Ethena and all similar models.

The systemic risks that protocol-level arbitrageurs bring to the ecosystem as a whole are examined further in Chapter 17.

16.2.3 The participant interaction network

The participants in the arbitrage ecosystem do not exist in isolation; through complex trading relationships and information flows, they form a tightly coupled interaction network. This network contains both a tiered competitive structure and complementary collaborative relationships, and it also harbors systemic risks that could trigger chain reactions. Understanding this interaction network is key to grasping the stability and fragility of the arbitrage ecosystem.

The logic of tiered competition is the most salient feature of this network. It follows a simple principle: faster, better-capitalized participants preferentially capture the simplest and most obvious arbitrage opportunities, forcing others toward more complex, thinner-margin domains. Tier 1 high-frequency traders sit at the top of the competitive hierarchy; at microsecond speed they rapidly eliminate spreads among the major exchanges, so that any simple cross-market arbitrage opportunity lasting more than a second quickly disappears. This forces Tier 2 quantitative funds to abandon direct competition on speed and instead develop strategies requiring longer holding periods and more complex models, such as basis arbitrage and statistical arbitrage. And when Tier 0's Ethena enters with tens of billions of dollars to execute basis arbitrage at scale, it significantly compresses the profit space of Tiers 2 and 4 in that strategy. The funding rate is systematically depressed, rendering the manual carry trade—once quite attractive to individual traders—unprofitable. This effect of progressively compressing profit margins tier by tier is the inevitable result of the arms race in arbitrage technology; it pushes the entire ecosystem to evolve toward ever greater efficiency and refinement, but it also raises the barriers to entry for new participants.

This ecosystem, however, is not defined solely by competition; complementary collaborative relationships are equally important to keeping it running. Participants at different tiers are, in effect, one another's counterparties and create the conditions for one another's operations. For example, Tier 1 market makers, in providing liquidity, also lower the transaction costs (slippage) of strategy execution for Tiers 2 and 4. A market with deeper order books and narrower bid-ask spreads benefits all participants. Conversely, the basis-arbitrage strategies that Tier 2 funds execute (such as buying BTC spot and shorting BTC perpetual futures) provide important hedging positions for Tier 1 market makers. When large numbers of speculators pile into long perpetual-futures positions, it is Tier 2 arbitrageurs who supply the short-side counterparty, thereby maintaining market balance. Likewise, when Tier 3 MEV searchers execute CEX-DEX arbitrage, although they extract value from the DEX's liquidity providers, they also help keep the DEX's price aligned with the external market, lowering transaction costs for ordinary users. This complex symbiotic network gives the arbitrage ecosystem a dynamic, self-regulating balance.

Competition within a single tier also plays the role of an "automatic stabilizer." When a particular arbitrage strategy (such as funding rate arbitrage on some altcoin) shows excess profit, it attracts many same-tier participants. For example, several Tier 2 funds may simultaneously discover and execute the strategy. Their collective action rapidly compresses the arbitrage space (the funding rate falls, the basis converges) until the strategy's expected return declines to a level commensurate with its risk and cost. At that point, some less efficient funds or those with higher capital costs choose to exit, so the profit space rebounds slightly and attracts a new round of competition. This process constitutes a negative feedback mechanism that ensures no riskless arbitrage opportunity can persist for long and keeps market efficiency at a dynamic equilibrium determined by the average arbitrage cost of all participants. This negative feedback mechanism may, however, turn nonlinear during the de-crowding phase. Behavioral-finance research shows that the upswing phase of strategy crowding is often accompanied by participant overconfidence and an underestimation of tail risk, and that when market conditions deteriorate, panicked exit can cause arbitrage intensity to fall far more than equilibrium requires—that is, spreads to widen excessively—similar to the "quant quake" that quantitative strategies experienced in 2007. In a crypto market where information spreads extremely fast and participant behavior is highly homogeneous, the nonlinear overshoot risk of de-crowding is especially pronounced.

But this tight coupling and interdependence also accumulate systemic risk. When a key node in the ecosystem fails, risk can propagate rapidly across the whole network. The Ethena case again serves as a warning. If Ethena had to urgently unwind its tens of billions of dollars in short positions for some reason (such as a security vulnerability in its custody scheme, or a default by the CEX platform it hedges on), it would generate enormous buying demand in the market. This would not only push perpetual-futures prices up rapidly and turn the funding rate sharply positive but also inflict large mark-to-market losses on all the long-side speculators and market makers whose short counterparty was Ethena. More seriously, such a shock could trigger chain liquidations at other institutions and set off a full-blown market crash. This "centralization risk," arising from a single entity's excessive scale, is a new challenge that protocol-level

arbitrageurs bring to the whole ecosystem. It shows that when arbitrageurs evolve from dispersed individuals into concentrated infrastructure, efficiency improves but the system's fragility may increase as well.

A deeper systemic risk is that an Ethena risk event could infect the broader market through three channels simultaneously. The first channel is a direct shock to the CEX perpetual-futures market: large-scale short covering pushes up the funding rate and perpetual-futures prices, affecting all traders holding positions on the relevant CEX. The second channel is a chain reaction in the DeFi collateral chain: USDe and sUSDe are already widely used as collateral in DeFi lending protocols and liquidity pools, and an Ethena risk event could cause the price of USDe to fall, triggering chain liquidations of positions collateralized by USDe. The third channel is a confidence crisis leading to a stablecoin depeg: market panic could drive the secondary-market price of USDe below its $1 peg, further intensifying the positive feedback of the first two channels. Under stress, these three channels could activate in synchrony, constituting a textbook "correlation jump" scenario in which risk factors that appear independent in normal times suddenly become highly correlated in a crisis, amplifying a local event into a systemic shock.

16.2.4 The protocolization of arbitrage

The perpetual-futures arbitrage ecosystem is not static; it has undergone dramatic evolution over the past several years. This process is closely synchronized with the three-stage evolutionary framework of crypto-market efficiency described in Chapter 15 (the early stage, the institutionalized era, and the regulated era) and clearly displays the path along which arbitrage behavior has migrated from scattered individual strategies to a highly organized infrastructure service.

In the early market stage, during the infancy of perpetual futures (roughly 2016–2019), the arbitrage ecosystem consisted mainly of Tier 4 individual traders and a few small teams with market insight. At the time, the market was severely fragmented, spreads across exchanges were large and persistent, and funding rates were extremely volatile. The pronounced efficiency gaps offered rich profits to early participants but also reflected the market's extreme lack of infrastructure. Arbitrage at this stage was opportunistic and unsystematic, far from constituting an "infrastructure" capable of stably anchoring prices.

Entering the "institutionalized era" (roughly 2020–2023), as crypto assets gradually drew the attention of mainstream financial institutions, Tier 1 high-frequency traders and Tier 2 systematic quantitative funds entered at scale. These institutions brought mature trading algorithms, large-scale capital, and professional risk-management systems. In this phase, the efficiency of core markets improved markedly. Cross-exchange spreads were compressed to the millisecond scale, and the funding rates of mainstream assets became smoother and more predictable. Arbitrage shifted from experience-driven individual behavior to systematized quantitative strategy and became a routine allocation for institutional capital. It was in this phase that arbitrage's infrastructure functions as a "cross-market clearing bus" and a "price anchor" took initial shape.

We are now in the "regulated era" (roughly 2024 to the present). The defining feature of

this phase is the rise of Tier 0 (protocol-level arbitrageurs) and Tier 3 (on-chain-native participants). The emergence of Ethena symbolizes the "institutionalization" and "protocolization" of the arbitrage strategy itself, which has become a foundational building block of the DeFi world. At the same time, as DEXs and on-chain derivatives protocols have flourished, on-chain-native forces such as MEV searchers and liquidation bots have grown steadily in importance. They have established new efficiency links between CEXs and DEXs and among different on-chain protocols, extending the infrastructure functions of arbitrage from the centralized world into the decentralized substrate. In this phase, the complexity and integration of the arbitrage ecosystem have reached levels not seen in any earlier phase, but they are accompanied by the centralization risk noted above and by new ethical challenges (such as the fairness of MEV).

Looking ahead, the evolution of the arbitrage ecosystem is far from over. One frontier direction worth watching is the application of artificial intelligence in arbitrage. As large language models and reinforcement-learning techniques advance, fully autonomous AI arbitrage agents may become a reality. Such potential AI-native arbitrage agents, surpassing today's Tier 1, might process and respond to unstructured information (such as news and social-media sentiment) faster than Tier 1 high-frequency traders, or autonomously discover statistical-arbitrage patterns more complex than those of Tier 2 quantitative funds. If such AI agents emerge, they could reshape the existing ecosystem, pushing the technological competition of arbitrage into a new dimension—from a contest of execution speed based on code and hardware toward a contest of cognitive capability based on algorithms and intelligence. At that point, the form, efficiency, and risk characteristics of arbitrage infrastructure would undergo another round of structural transformation.

16.3 A taxonomy of arbitrage

Before dissecting the various concrete arbitrage mechanisms in depth, establishing a clear classification framework is a necessary prerequisite for the analysis that follows. The institutional distinctiveness of the perpetual-futures market—its funding rate mechanism, its lack of a fixed maturity, its high leverage, its round-the-clock trading, and its complex ecosystem in which centralized and decentralized exchanges coexist—jointly forms an arbitrage strategy space far richer and more multidimensional than that of traditional financial derivatives markets. These strategies do not exist in disarray but follow a specific structural logic. This section aims to establish a unified analytical coordinate system that systematically classifies the core arbitrage activities of the perpetual-futures ecosystem by identifying the type of boundary each arbitrage "crosses," and reveals their intrinsic progression from simple to complex.

16.3.1 The classification framework

The traditional way to classify market arbitrage often relies on asset class or the mathematical model of the strategy. In the highly heterogeneous and rapidly evolving environment of crypto markets, however, a more robust taxonomy should focus on the essence of arbitrage behavior: identifying and exploiting the "spread" or "inconsistency"

between different markets, assets, or instruments. The core of the classification framework we propose therefore lies in defining precisely which boundary the "deviation" that arbitrageurs seek to eliminate occurs across. We decompose this core idea into two principal dimensions and one implicit dimension.

The first principal dimension is the "type of boundary" that the arbitrage crosses. This defines the fundamental cause of the spread—the dimension along which the market has become segmented or inconsistent. We identify six key boundary types. A cross-market-tier boundary refers to the price deviation of the same asset across different market tiers, the most typical being the spread between the spot market and the derivatives market. A cross-space/platform boundary refers to the price deviation of the same asset or contract across different geographic locations or trading platforms. A cross-institutional-parameter boundary refers to arbitrage opportunities that arise for the same asset across different trading venues because of their distinctive institutional designs (such as the funding rate, settlement timing, or margin rules). A cross-asset boundary refers to a temporary deviation in the price relationship between two or more strongly correlated or cointegrated assets within the same market. A cross-term-structure boundary refers to a distortion in the pricing relationship implied among derivatives of different maturities on the same underlying asset. A cross-infrastructure-architecture boundary refers specifically to spreads that arise between platforms operating on the two starkly different technical and trust assumptions of CEXs and DEXs, owing to differences in oracle latency, transaction-packing mechanisms, or liquidity models.

The second principal dimension is the "combination of market tiers" involved in the arbitrage operation. This defines specifically which market segments the two or more legs of the arbitrage trade fall into—for example, whether it connects spot and perpetual futures, hedges between different perpetual futures, or constructs a complex portfolio involving spot, futures, and perpetual futures.

Beyond these two explicit dimensions, there is an implicit dimension: the "holding period." The time scales required to discover, execute, and close different arbitrage strategies vary enormously. A high-frequency trader may complete a cross-exchange arbitrage within microseconds, whereas funding rate arbitrage may require holding for days or even weeks. The length of the holding period directly reflects the risk exposure, capital cost, and technological threshold of an arbitrage strategy; together with the two preceding dimensions, it determines the complexity and profit model of a given arbitrage strategy.

Through this two-dimensional "boundary type $\times$ market tier" matrix, supplemented by the holding-period consideration, we can construct a clear map of the seemingly tangled arbitrage activities in the perpetual-futures ecosystem, providing a unified logical starting point for the deconstruction in the sections that follow.

16.3.2 An overview of the six arbitrage types

Based on the classification framework above, we can group the mainstream arbitrage activities of the perpetual-futures ecosystem into six core types. The table below systematically organizes the characteristics of these six types, including the boundary each crosses, the market tiers involved, the core instrument, and the "efficiency function"

each plays for the market ecosystem as a whole—that is, which dimension of market inefficiency each primarily repairs.

No.	Arbitrage type	Boundary crossed	Market tiers	Core instrument	Efficiency function
①	Spot-perpetual basis arbitrage	Cross-market-tier	Spot ↔ perpetual	Basis / funding rate	Anchors the perpetual-futures price to spot
②	Cross-exchange perpetual arbitrage	Cross-space/platform	Perpetual ↔ perpetual	Cross-exchange spread	Unifies global perpetual-futures prices
③	Funding rate arbitrage	Cross-institutional-parameter	Perpetual ↔ perpetual	Rate differential / rate surface	Smooths the rate structure
④	Cross-asset statistical arbitrage	Cross-asset	Perpetual ↔ perpetual	Spread / correlation	Maintains relative-pricing consistency
⑤	Spot-futures-perpetual triangular arbitrage	Cross-term-structure	Spot ↔ futures ↔ perpetual	Term structure	Integrates term-structure information
⑥	CEX-DEX cross-tier arbitrage	Cross-infrastructure-architecture	CEX perpetual ↔ DEX perpetual	Spread / rate differential	Bridges the CEX-DEX efficiency gap

Table 16-2. The six core arbitrage types of the perpetual-futures ecosystem and their efficiency functions (Data source: constructed by the author)

These six arbitrage mechanisms do not exist in isolation but form a functionally complementary network that jointly maintains market efficiency. To understand their "division of labor" more intuitively, we can map each arbitrage type to the market boundary it crosses—this is precisely the concrete unfolding, at the micro-mechanistic level, of the market-efficiency spectrum of Chapter 15. As the figure below shows (corresponding to the six boundary types summarized in Table 16-2), the six arbitrage types cover six distinct boundary types, each assuming a corresponding price-correction function, and together they drive the improvement of market efficiency.

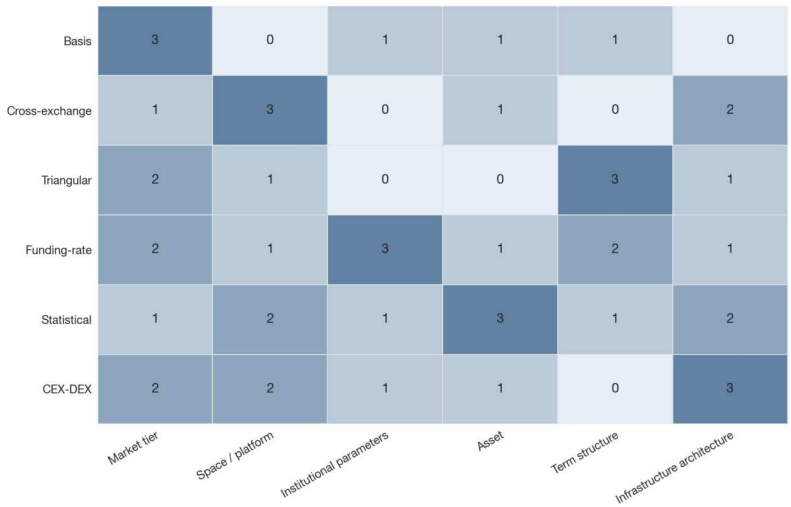


Figure 16-6. The functional coverage of the six arbitrage types across the six boundary types they cross (corresponding to Table 16-2; source: a conceptual schematic constructed by the author from the analysis in each section)

The six arbitrage types form a functional division of labor across the six boundary types listed in Table 16-2: spot-perpetual basis arbitrage anchors the "market-tier" boundary, CEX-DEX cross-tier arbitrage bridges the "infrastructure-architecture" boundary, cross-exchange perpetual arbitrage governs the "space/platform" boundary, funding rate arbitrage smooths the "institutional-parameter" boundary, cross-asset statistical arbitrage maintains relative pricing across the "asset" boundary, and triangular arbitrage integrates the "term-structure" boundary.

Through this mapping, we can see clearly that a seemingly unified market price is, behind the scenes, the result of different types of arbitrageur continuously performing price-correction operations within their respective functional domains. Together they constitute a dynamic, multi-tiered error-correction system that is the micro-foundation on which market efficiency is maintained.

16.3.3 The complexity gradient

Dividing arbitrage activity into six types is not merely a matter of classificatory tidiness; more importantly, it reveals the intrinsic progression among them. From type ① to type ⑥, we observe a clear evolutionary trend from simple to complex. This increase in complexity manifests in several respects: the number of markets involved, the sources of institutional friction to be overcome, the demands on technological infrastructure, and the risk dimensions embedded in the strategy itself.

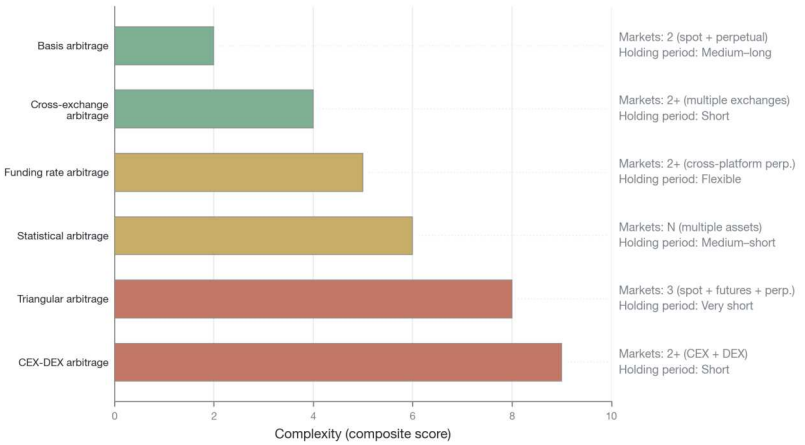


Figure 16-7. The complexity gradient of the six arbitrage types (Data source: a qualitative assessment constructed by the author from three indicators—the number of markets involved, the variety of friction sources, and the technological-infrastructure requirements)

As the figure shows, complexity escalates step by step. Spot-perpetual basis arbitrage (①) is the most basic form; it typically involves only two markets (one spot, one perpetual), and its core frictions arise mainly from transaction costs and short-term fluctuations in the funding rate. This makes it the arbitrage type with the lowest threshold and the widest range of participants, from large quantitative funds to individual traders.

By contrast, although cross-exchange perpetual arbitrage (②) also involves only two markets, it demands extremely high speed; the technological threshold rises significantly, and it is dominated mainly by high-frequency trading firms. Funding rate arbitrage (③), meanwhile, introduces the need to understand differences in institutional parameters across platforms, and its complexity lies in the fine-grained grasp of rate-calculation mechanisms and settlement cycles.

The complexity of cross-asset statistical arbitrage (④) stems from the word "statistical." It is no longer based on a deterministic riskless pricing relationship but relies on modeling the historical correlation between assets, which introduces model risk and a higher technological threshold.

The complexity of spot-futures-perpetual triangular arbitrage (⑤) manifests in the increase to three markets involved and in the need to handle simultaneously the pricing logic of different term structures, with a corresponding increase in risk dimensions.

Finally, CEX-DEX cross-tier arbitrage (⑥) stands at the top of the complexity ladder. It must contend not only with the two starkly different trading environments of CEXs and DEXs but also with a variety of distinctive friction sources and risk dimensions, including on-chain congestion, gas-fee fluctuations, oracle latency, and even MEV. This confines the effective participants in this domain to a small number of on-chain searchers and high-frequency trading firms with top-tier technical capabilities.

This progression from simple to complex also broadly matches the chronological order in which these arbitrage strategies historically emerged and scaled. Basis arbi-

trage existed from the very birth of perpetual futures; as the number of exchanges grew, cross-exchange arbitrage followed; after competition among exchanges intensified and institutional designs diverged, funding rate arbitrage became popular; and CEX-DEX arbitrage became possible only after the rise of DeFi and decentralized perpetual-futures protocols. This evolutionary path clearly shows that the arbitrage ecosystem is itself continuously adapting and developing: as the market deepens and expands, new and more complex arbitrage opportunities keep emerging, drawing in more specialized participants and together constituting a dynamically evolving efficiency landscape.

16.4 Spot-perpetual basis arbitrage

In the arbitrage ecosystem of perpetual futures, spot-perpetual basis arbitrage—often called "cash-and-carry" arbitrage—constitutes the most basic price-stabilization mechanism. Through continuous price-correction operations, basis arbitrage anchors the perpetual-futures price near the spot price and is the micro-foundation that allows the funding rate mechanism to function effectively.

16.4.1 The bidirectional operating mechanism

The core of basis arbitrage lies in capturing temporary deviations between the perpetual-futures price and the spot price. When the basis deviates significantly from zero because of market sentiment, leverage demand, or liquidity shocks, an arbitrage opportunity arises. Depending on the direction of the basis, the operation can be divided into two modes: positive-basis arbitrage and negative-basis arbitrage.

Positive-basis arbitrage is executed in an environment of positive basis and positive funding rate: the arbitrageur buys spot and shorts an equivalent amount of perpetual futures at nearly the same time, constructing a delta-neutral position. The phrase "at nearly the same time" matters here, because in actual execution there is always a time gap of milliseconds to seconds between the two legs—the spot leg and the perpetual-futures leg—whether across exchanges or even within the same exchange. This "execution lag" implies a window of directional risk exposure between the two legs, known as "leg risk." In a highly volatile market, even a lag of a few hundred milliseconds can cause significant slippage or a single-leg fill, so managing the execution lag constitutes a core component of arbitrage technology. Its sources of return include the funding rate collected periodically and the spread profit when the basis converges. From the standpoint of market efficiency, positive-basis arbitrage increases the short-side supply in the perpetual-futures market, exerts downward pressure on the overheated perpetual-futures price, and prompts the basis to narrow, achieving an automatic calibration of prices.

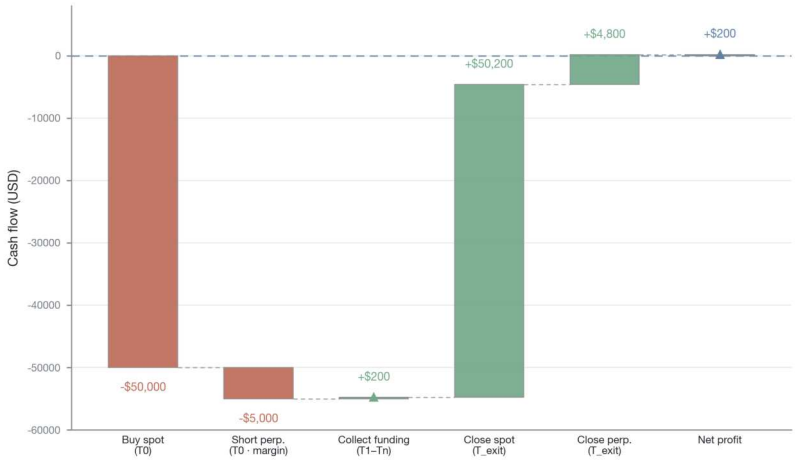


Figure 16-8. The cash-flow timeline of basis arbitrage (Data source: a conceptual schematic constructed by the author)

Figure 16-8 depicts the cash-flow timeline of positive-basis arbitrage over the holding period: at entry, buying spot and shorting the perpetual; during the holding period, collecting the funding rate at each settlement cycle; and at exit, realizing the profit from basis convergence—so that the net return equals the sum of the accumulated funding rate and the basis-convergence gain, less all transaction costs. The diversification of settlement frequency (4-hour and 1-hour intervals now coexist with the traditional 8-hour interval) directly affects returns: the more frequent the settlement, the smaller the amount collected each time, the shorter the lock-up period, and the lower the funding-rate volatility risk the arbitrageur faces.

Negative-basis arbitrage performs the reverse operation: shorting spot and going long the perpetual futures. In the crypto market, however, spot shorting faces pronounced asymmetric friction, so the execution efficiency and scale of negative-basis arbitrage fall far short of those of positive-basis arbitrage. Specifically, the currently available channels for spot shorting include borrowing through centralized lending platforms (the available platforms have been greatly reduced after the bankruptcies of several leading institutions, and borrowing rates are typically between 5% and 30% annualized); borrowing through DeFi lending protocols such as Aave (which requires over-collateralization and is extremely capital-inefficient); and the coin-borrowing (margin) features of some exchanges (with limited quotas and opaque rates). The availability, cost, and counterparty risk of each channel differ starkly. This asymmetry of shorting friction directly explains why the negative-basis state tends to persist longer, and deviate further, than the positive-basis state—one of the market anomalies that Chapter 18 will examine.

16.4.2 Decomposing the returns

The returns from basis arbitrage are not riskless profit but compensation for bearing specific risks and friction costs, fully consistent with the arbitrage infrastructure hy-

pothesis proposed in this chapter. He et al. (2022) [4] note that the perpetual-futures price shows an annualized mean absolute deviation of 60% to 90% from its theoretical no-arbitrage price, while a simulated basis-arbitrage strategy can generate a very high Sharpe ratio (for Bitcoin perpetual futures, about 1.8 for retail traders facing high transaction costs and as much as 3.5 for zero-fee high-frequency market makers). This statistic should be viewed with caution, however: the Sharpe ratio assumes normally distributed returns, whereas the return distribution of basis arbitrage exhibits pronounced negative skewness and fat tails—stable positive returns in normal conditions and enormous losses under extreme stress. This return structure means that the Sharpe ratio may severely understate the strategy's risk exposure in tail events. A more robust risk assessment should be supplemented with metrics sensitive to tail risk, such as maximum drawdown and conditional value-at-risk.

The net return of basis arbitrage can be decomposed into three main parts: the funding rate income, the basis-convergence income, and the various transaction costs. The formula can be expressed as:

$$\text{Net return} = \sum_{k} \text{funding rate}_k + (\text{basis}_{\text{open}} - \text{basis}_{\text{close}}) - \text{transaction costs} - \text{margin opportunity cost} - \mathbb{E}[\text{liquidation/ADL loss}]$$

where k indexes the settlement cycles within the holding period.

Here, transaction costs are a composite concept encompassing trading fees, the bid-ask spread (slippage), spot carrying costs or borrowing costs (in negative-basis arbitrage), and the cost and time delay of cross-platform fund transfers. These costs directly erode the gross return and determine that arbitrage is economically viable only when the gross basis is large enough to cover them. In addition, two cost items often overlooked but critical in actual execution are: the margin opportunity cost—the initial and maintenance margin that must be locked up on the perpetual-futures leg cannot be used for other investments, and in a high-rate environment this implicit cost is not negligible; and the probability-weighted loss from forced-liquidation and auto-deleveraging (ADL) risk—during violent market swings, unrealized losses on the perpetual-futures leg may trigger margin calls or even forced deleveraging, while at that moment the unrealized gains on the spot leg cannot be used to top up margin because they sit on a different platform. Some exchanges have introduced unified margin modes (such as OKX's portfolio margin and Binance's unified account) that allow spot assets to serve directly as margin for perpetual futures, partly alleviating the problem of cross-platform margin separation and improving the capital efficiency of basis arbitrage.

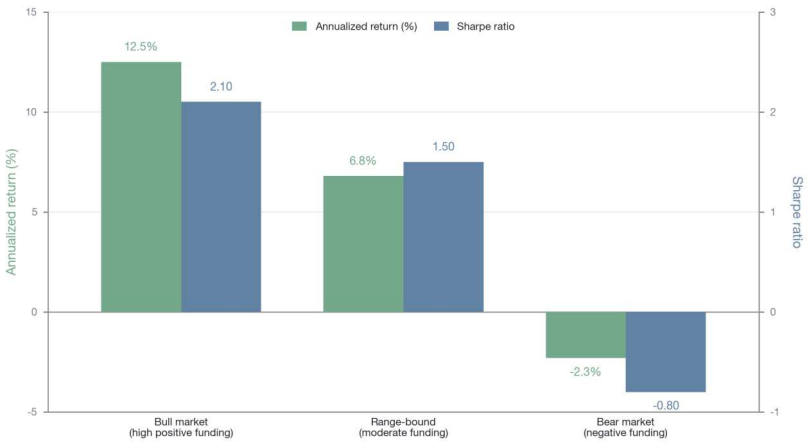


Figure 16-9. The returns and Sharpe ratios of basis arbitrage across different market states (the overall Sharpe ratio is based on He et al. 2022 [4]; the by-state decomposition into bull, bear, and range-bound markets is a representative illustration, not the original by-state series of [4])

Figure 16-9 compares the returns and Sharpe ratios of basis arbitrage across three market states—bull, bear, and range-bound: returns are highest and volatility lowest in a bull market (the Sharpe ratio can exceed 2.0), returns fall markedly in a bear market as the funding rate turns negative, and the range-bound market lies in between.

Importantly, the returns of basis arbitrage exhibit pronounced dependence on the market state. During bull markets, sentiment runs high, leverage demand keeps the perpetual-futures price persistently above the spot price, and the funding rate stays at a relatively high positive level for a long time. In such conditions, the returns of the positive-basis-arbitrage strategy are the highest and least volatile of the three market states. Conversely, during bear markets, panic drives the perpetual-futures price below spot, and the funding rate turns negative. In such conditions, positive-basis arbitrage faces losses, and arbitrageurs must turn to the more difficult negative-basis arbitrage, whose returns are also fraught with uncertainty owing to the cost of spot shorting. In a range-bound market, the basis and funding rate fluctuate little, and although arbitrage returns are lower than in a bull market, they can still provide a relatively stable cash flow. This time-variation in returns is closely tied to the crowding effect. When an arbitrage opportunity becomes widely known and profitable, large amounts of capital pile into the strategy, rapidly compressing the basis and driving down the funding rate, thereby squeezing the excess-profit space. This is precisely the crypto-market version of the Grossman-Stiglitz paradox: arbitrage, in the very act of creating market efficiency, also eliminates the profit space that drives it.

16.4.3 The Ethena effect

The rise of the Ethena protocol marks the shift of basis arbitrage from dispersed execution to a protocolized, scaled paradigm. Its "arbitrage-as-a-service" mechanism (users deposit collateral, the protocol automatically constructs a delta-neutral position, and the funding rate and basis income are distributed in the form of sUSDe) was detailed in Section 16.2.2; this section focuses on the quantitative effects that its scaling produces.

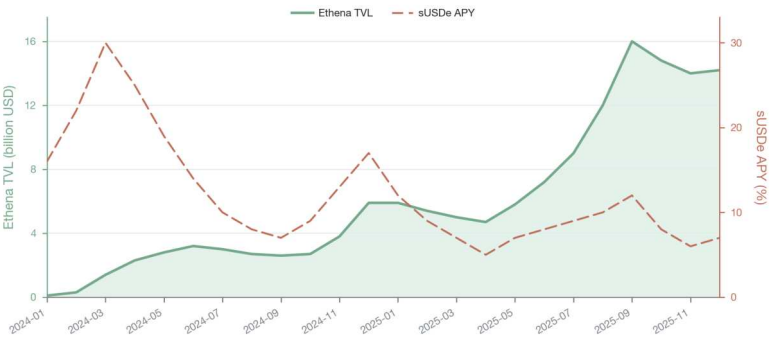


Figure 16-10. The historical trajectory of Ethena's TVL and sUSDe APY (data source: DefiLlama [7], accessed March 10, 2026; the TVL peak of about $16 billion is on the DefiLlama basis, or about $14.5 billion on the USDe-supply/market-cap basis—the two bases are not interchangeable—as of September 2025)

Ethena's success produced an extremely rare scale effect. Its TVL climbed rapidly from under $1 billion within a few months and peaked in September 2025—about $16 billion on the DefiLlama TVL basis, or about $14.5 billion if measured by USDe supply/market cap (the two bases are not interchangeable; the DefiLlama figure is used here [7], accessed March 10, 2026). According to derivatives-industry data, the perpetual short positions Ethena held accounted for about 5% of ether perpetual-futures open interest across the market in early 2024 [8]; at the peak of its scale, on specific BTC and ETH perpetual-futures contracts at some centralized exchanges, industry estimates put its share of that contract's open interest at as much as roughly 10-15% (the basis depends on the specific contract; see Section 17.10 in Chapter 17). This scale turned it from a mere "arbitrageur" into a "market maker" capable of influencing the market itself, producing what is called the "Ethena effect." When the Ethena protocol absorbs vast amounts of capital and builds large short positions, it itself becomes one of the main forces depressing the funding rate. The market thereby enters a textbook reflexive loop: Ethena's high yield attracts large TVL inflows → the growth in TVL forces Ethena to open more perpetual-futures shorts → the enormous short position exerts downward pressure on the funding rate → the decline in the funding rate in turn erodes Ethena's own source of returns. This is a protocolized version of the classic "arbitrage paradox": when an arbitrage strategy is scaled up excessively, its own behavior compresses the space that generates its profit.

Ethena's quarterly revenue data clearly reflect this dynamic. As the figure below shows, Ethena's total revenue peaked at $151 million in the third quarter of 2025, precisely during the high-yield phase when its TVL was growing rapidly, market sentiment was optimistic, and the funding rate was elevated. However, as its TVL topped out at the end of the third quarter and began to decline, and as changing market conditions compressed the funding rate, its quarterly revenue fell accordingly (to about $96 million in the fourth quarter). The staking yield of sUSDe also retreated from an early high of over 20% to single digits. This shows that even protocolized arbitrage infrastructure cannot escape the economic law of diminishing returns.

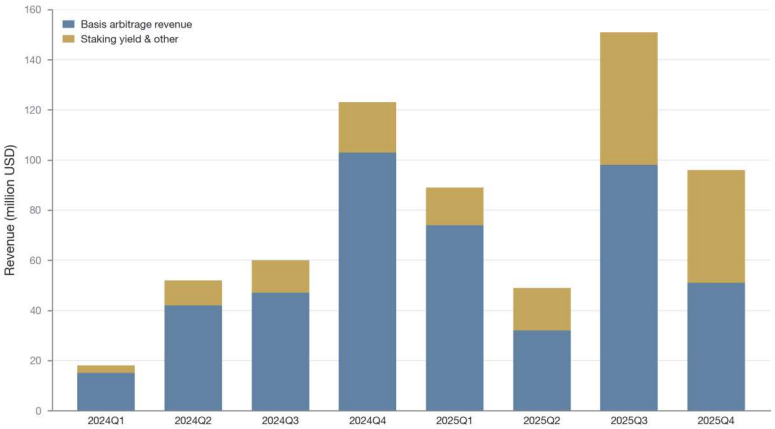


Figure 16-11. The composition of Ethena's quarterly total revenue (data source: Token Terminal [9] "total revenue" basis, accessed March 10, 2026; not interchangeable with the DefiLlama "fees" basis)

Historical precedents warn of the concentrated-unwinding risk of large-scale basis trades. During the global panic triggered by the COVID-19 pandemic in March 2020, the basis trade in the U.S. Treasury market (hedge funds going long Treasury cash securities and short Treasury futures at large scale to earn the basis) suffered chain forced liquidations as margin calls hit; positions that had been nearly riskless caused billions of dollars in losses within days, and stability was ultimately restored only through the Federal Reserve's emergency intervention as lender of last resort. In the crypto space, during the Terra/LUNA collapse of May 2022, many delta-neutral strategies based on perpetual futures saw the "market-neutral" assumption fail under extreme stress as the liquidity of the underlying evaporated instantly and the basis widened sharply. The fundamental institutional difference between the two is that the crypto market has no equivalent "lender of last resort" mechanism to interrupt the positive feedback of a liquidity spiral, and cross-platform margin separation (margin cannot be shared between different CEXs or between a CEX and a DEX) makes the transmission path of a liquidity crisis more fragmented. These cases show that the systemic risk of basis arbitrage is highly nonlinear in strategy scale: when the executors of a single strategy command too large a share, their unwinding demand can exceed the market's absorptive capacity, turning a "riskless" strategy into a source of systemic risk. Ethena's current scale places it under similar risk constraints.

This process clearly displays the feedback loop between arbitrage intensity and market efficiency and provides strong case support for the arbitrage infrastructure hypothesis: when infrastructure providers evolve from dispersed individuals into a concentrated protocol, efficiency improves, but concentration risk follows—a point that Chapter 17 will examine further.

16.4.4 Empirical evidence for price anchoring

The core efficiency function of basis arbitrage is that it constitutes the micro-economic force maintaining the anchoring of the perpetual-futures price to the spot price. When-

ever the perpetual-futures price deviates from the spot price because of speculation, leverage, or liquidity imbalance, the profit-seeking behavior of basis arbitrageurs constitutes an endogenous price-convergence force that drives the two back into alignment. This price-anchoring function can be verified through empirical data along several dimensions.

First, there is a significant negative correlation between the intensity of basis-arbitrage activity and the level of the basis. We can proxy the intensity of arbitrage activity by observing the estimated share of delta-neutral positions in perpetual-futures open interest, or changes in exchange margin-lending balances. Mechanistically, one would expect that when the arbitrage activity reflected in these proxy indicators strengthens, the level and volatility of the basis typically decline in turn, which supports the corrective effect of arbitrage on price deviations (note that arbitrage intensity here is measured by proxy indicators; one should avoid inferring arbitrage intensity from spread convergence itself, which would be circular). The mean-reversion property of the funding rate observed in Chapter 15 is driven, at bottom, by basis arbitrage. When the funding rate is too high, arbitrageurs enter the market to build short positions and collect the rate, thereby depressing it; when the rate is too low, the reverse operation applies—ultimately pushing the rate back toward its long-run mean.



Figure 16-12. The long-run compression trend of the basis (Data source: Binance and OKX exchange API data [10])

Taking BTC as an example, Figure 16-12 tracks the long-run evolution of the spot-perpetual basis from 2020 to 2026: both the average level and the amplitude of fluctuation have fallen markedly, and the basis of tens of basis points in the early years has been compressed under normal market conditions to a low of a dozen or so basis points—the most direct evidence that the arbitrage infrastructure is steadily maturing.

Second, from a more macro perspective, as the crypto market has matured, the number and capital scale of basis arbitrageurs have grown continuously, producing a long-run "compression" trend in the basis. A review of the data from recent years shows that both the average level and the amplitude of the spot-perpetual basis for mainstream assets (such as BTC and ETH) exhibit a clear downward trend. In the early market, a basis of tens or even over a hundred basis points was not uncommon, whereas today, under normal market conditions, the basis is typically confined to a very narrow range. This

declining curve is the most direct evidence of improving market efficiency; it shows that the arbitrage infrastructure is becoming ever more complete and efficient, able to identify and repair price deviations more quickly. The strength of this "efficiency anchor," however, is not unlimited. Under extreme market stress—for example, during a violent price crash or a liquidity crisis—arbitrageurs may be forced to unwind because of margin risk, forced liquidation, or counterparty risk, causing the arbitrage force to weaken or even vanish. At such times the anchoring mechanism may temporarily "fail," producing a persistent deviation of the basis; these "arbitrage failure" scenarios will be analyzed in depth as core anomalies in Chapter 18.

16.5 Cross-exchange arbitrage

Cross-exchange arbitrage is the core mechanism that unifies globally fragmented trading markets. In a crypto market that lacks central clearing and unified regulation, price differences for the same asset across different trading venues are almost inevitable. Cross-exchange arbitrageurs drive these spreads to converge and integrate global liquidity, constituting the second pillar of this chapter's arbitrage infrastructure hypothesis.

The generation, propagation, and elimination of a spread constitute the micro-foundation for understanding cross-exchange arbitrage. From early manual cross-platform arbitrage to today's sub-millisecond automated competition, the evolutionary path of arbitrage execution clearly reflects how technological progress has advanced market efficiency. The classic study by Makarov and Schoar (2020) [1] revealed the deep institutional frictions behind the persistence of cross-border spreads, while the long-run compression trend of cross-exchange spreads provides the empirical basis for a "core-periphery" efficiency model.

16.5.1 The life cycle of a price spread

The law of one price (see Section 16.1.1) is far from perfectly satisfied in the digital asset market, and the market's distinctive structural features provide the structural conditions for spreads to arise. The life cycle of a spread—from its generation and propagation to its eventual elimination by arbitrage activity—fully reveals the dynamic process by which market efficiency is generated.

The generation of a spread is rooted in the market's inherent segmentation. First is the delay in information propagation. Although information travels globally at extremely high speed, on the scale of milliseconds or even microseconds, delays still exist. The price impact of a large buy order on Coinbase takes some time to propagate fully to the order books of Binance or Kraken. This tiny delay constitutes the "time window" that high-frequency arbitrageurs exploit. Second is the segmentation of liquidity. Hundreds of exchanges worldwide each have their own order books and liquidity pools. When one exchange faces a large one-directional order flow (for example, heavy buying or selling), its limited liquidity depth causes a significant price deviation, while at that moment prices on other exchanges may not yet have been affected, and a spread thereby arises. Finally, differences in institutions and infrastructure also play an important role. Differences in exchanges' fee structures, API performance, physical server locations, and

even deposit and withdrawal speeds all affect trader behavior and the price-formation process, generating or entrenching spreads.

Once a spread forms, it "propagates" through the market via arbitrageurs' observation and action. At first, perhaps only a few Tier 1 high-frequency traders at the trading frontier can capture the signal rapidly on their global server clusters. Their algorithms quickly assess whether the magnitude of the spread is enough to cover their highly optimized execution costs. If the answer is yes, the arbitrage trade is triggered: buying on the lower-priced exchange and selling on the higher-priced exchange. This action is itself the beginning of the spread-elimination mechanism. Buying pushes up the price on the low-price exchange, and selling depresses the price on the high-price exchange, thereby narrowing the spread.

As information spreads further, Tier 2 quantitative funds and a broader set of traders begin to observe the opportunity. Their reaction speed may be on the order of seconds or even minutes, but their collective action further drives prices toward convergence. This process forms a tiered transmission of efficiency: the fastest arbitrageurs capture the largest and most perishable profits and complete the initial price calibration; subsequent participants clean up the remaining, smaller spreads until the magnitude narrows to a level at which even the lowest-cost participants can no longer profit. The spread-convergence process above is not always monotonically decreasing, however. In a highly volatile market, arbitrage itself may temporarily widen the spread by consuming order-book depth: when arbitrageurs execute large buy orders on the low-price exchange, their market impact may push that exchange's price up by more than the spread converges, producing an oscillating "widen-converge" dynamic. In addition, when market makers sense the presence of active arbitrageurs, they may cancel orders or widen their quotes to reduce their own adverse-selection risk, and this strategic behavior may likewise delay or even temporarily reverse the spread's convergence. At such times the spread does not vanish entirely but stabilizes at a "residual" level reflecting the market's average arbitrage cost. The existence of this residual validates the crypto-market version of the Grossman-Stiglitz paradox: in a market with no spread at all, no arbitrageur would have any incentive to keep its prices aligned. A tiny but persistent spread is therefore the cost that must be paid to keep the arbitrage infrastructure functioning effectively.

16.5.2 The evolution of the latency race

The history of cross-exchange arbitrage is a history of continuous iteration in technology and strategy, at the core of which is the relentless pursuit of ever-lower latency. From the manual cross-platform arbitrage of Bitcoin's infancy to today's sub-millisecond automated competition dominated by high-frequency trading firms, this path clearly shows how arbitrage evolved from simple opportunity discovery into a highly precise and capital-intensive science.

In the early market stage before 2017, cross-exchange arbitrage was typically conducted in a rudimentary way, often called manual cross-platform spread arbitrage. Traders manually observed price differences on the web pages of different exchanges, bought Bitcoin manually on one platform, then withdrew it and deposited it on another, higher-priced platform to sell, earning the difference. The latency of this process was measured

in hours or even days and involved cumbersome manual operations and enormous price-fluctuation risk. Market efficiency at the time was extremely low; spreads could reach more than 10% and persist for days, leaving ample profit space for this slow form of arbitrage. The limitations of this model, however, were obvious: the network-confirmation delays of withdrawal and deposit, exchange processing speed, and counterparty risk all severely limited the scale and efficiency of arbitrage.

The 2017 bull market and the spread of API trading ushered arbitrage into the "automation era." Traders began writing scripts to fetch real-time quotes from multiple exchanges via API and to execute trades automatically when a spread was detected. This slashed arbitrage latency from hours to seconds or even minutes. In this phase, arbitrageurs no longer needed to transfer assets frequently between exchanges. They pre-deposited fiat and crypto as inventory at several major exchanges, and once a spread was detected, they could simultaneously execute a buy order on the low-price exchange and a sell order on the high-price exchange, completing the hedge. After the market stabilized, they rebalanced asset positions across exchanges through slow on-chain transfers. This "inventory management" model significantly improved arbitrage efficiency, made high-frequency intraday trading possible, and compressed cross-exchange spreads to below 1%.

After 2020, as traditional high-frequency trading firms such as Jump Trading and Wintermute entered in force, the latency race in the crypto market was pushed to an extremely high level, entering the "sub-millisecond era." These top HFT firms deploy their own servers near the data centers of the world's major exchanges and connect via dedicated lines to obtain the lowest quote and trade latencies. Their competitive edge lies no longer merely in strategy but in a contest of infrastructure. From hardware and networks (microwave towers) to software (highly optimized C++ code), every link involves heavy investment to shave off a few microseconds of latency. In their world, the discovery, decision, and execution of an arbitrage opportunity are all completed within a few hundred microseconds. This extreme speed allows them to capture the fleeting, basis-point-scale spreads. Their high-frequency activity has effectively eliminated the vast majority of simple arbitrage opportunities, forcing other participants toward more complex, longer-horizon strategies.

The ultimate result of this latency race is a great improvement in market efficiency. Spreads that once took hours to erase now converge within seconds. It has also clearly stratified the arbitrage ecosystem: Tier 1 HFT firms dominate the frontier latency game, Tier 2 quantitative funds seek more complex statistical-arbitrage opportunities at the second and minute scales, and Tier 4 individual traders have been almost entirely squeezed out of pure cross-exchange spread arbitrage. This deep coupling with exchange infrastructure, however, has also created a new dimension of operational risk. When a major exchange suffers an outage, an API failure, or a suspension of withdrawals, all Tier 1 participants relying on that exchange as one leg of their arbitrage face single-leg exposure risk—the perilous situation in which one leg of a position has been established while the other cannot be hedged. During the FTX collapse, the funds of many market makers holding arbitrage positions on the platform were frozen, starkly revealing exchange-concentration risk as a systemic vulnerability of the arbitrage infrastructure. The end point of the latency race is a highly automated, global

price-integration mechanism dominated by algorithms and infrastructure.

16.5.3 The Makarov-Schoar effect

Although technological progress has significantly compressed cross-exchange spreads, a central question remains: why can certain spreads—especially cross-border spreads involving different national fiat currencies—persist, sometimes reaching significant magnitudes? Makarov and Schoar (2020) [1] offered a classic explanation. Their findings reveal that beyond technological frictions, institutional frictions (especially capital controls) are the key to understanding the limits of crypto-market efficiency.

By analyzing tick-level data from 34 exchanges worldwide during 2017–2018, Makarov and Schoar documented the extent to which the crypto market violated the law of one price: spreads between exchanges within the same country typically did not exceed 1%, but cross-border spreads were large and could persist for weeks or even months (as noted earlier in Section 16.1.1, the Korean "kimchi premium" peaked above 40% and the Japan-U.S. spread was about 10%). They estimated that during the period of violent market fluctuation from December 2017 to February 2018, the total potential profit from global cross-exchange arbitrage reached at least $2 billion.

If such enormous riskless profit existed, why did arbitrageurs not eliminate it entirely? Makarov and Schoar's analysis points out that the answer lies not in arbitrageurs being insufficiently fast or clever but in the fact that capital could not flow freely. A typical cross-border arbitrage operation is: buy Bitcoin with U.S. dollars on a U.S. exchange, then move the Bitcoin to a Korean exchange and sell it for Korean won. At that point, although the arbitrageur has earned a won profit, the initial capital (U.S. dollars) has become won. Because of Korea's strict capital controls, converting that won profit back into dollars at scale and low cost and remitting it abroad is extremely difficult. This means arbitrage capital flows "one way" and easily gets "stuck" in a particular country. Once large amounts of arbitrage capital flowed into Korea, it was trapped within the local financial system and could not flow back to execute a new round of arbitrage. Arbitrage activity therefore soon halted for lack of capital, and the large spread persisted.

To test this hypothesis further, Makarov and Schoar conducted two ingenious tests. First, they compared the spreads of "crypto-fiat" pairs with those of "crypto-crypto" pairs. They found that at the same Korean exchange, when the price of Bitcoin against the won carried a premium of as much as 20%, the price of Bitcoin against Ethereum was essentially in line with the global market, with a spread of only about 3%. This is because the exchange and transfer between cryptocurrencies is completed entirely on-chain and can bypass the traditional banking system, escaping the constraint of capital controls. This comparison strongly confirms that institutional frictions are the main source of the spread. Second, they found that countries with stricter capital controls showed stronger correlation in the fluctuations and premium levels of their Bitcoin prices with those of other likewise-controlled countries, forming a "regulated club." This indicates that capital controls are a common systemic factor causing market segmentation and efficiency loss.

The Makarov-Schoar effect provides key empirical support for this chapter's arbitrage

infrastructure hypothesis. It strongly shows that arbitrage returns are not always a simple manifestation of market inefficiency but, more often, the risk compensation that arbitrageurs demand for overcoming enormous institutional frictions. When arbitrageurs must confront risks such as capital being frozen and legal-policy uncertainty, the "profit" they earn is in fact the reward for providing the entire market with the infrastructure services of "cross-border liquidity" and "price integration." The size of the spread is a direct measure of the cost of this service.

16.5.4 The long-run compression of spreads

Although the Makarov-Schoar effect reveals persistent spreads caused by institutional frictions, over a longer horizon, as the market has matured and the arbitrage infrastructure has become more complete, the overall level of cross-exchange spreads has undergone a marked long-run compression. This downward trend is the most direct evidence that arbitrage activity continuously compresses spreads. It shows that despite structural obstacles, arbitrageurs' sustained activity is indeed gradually drawing global market prices together and improving overall efficiency.

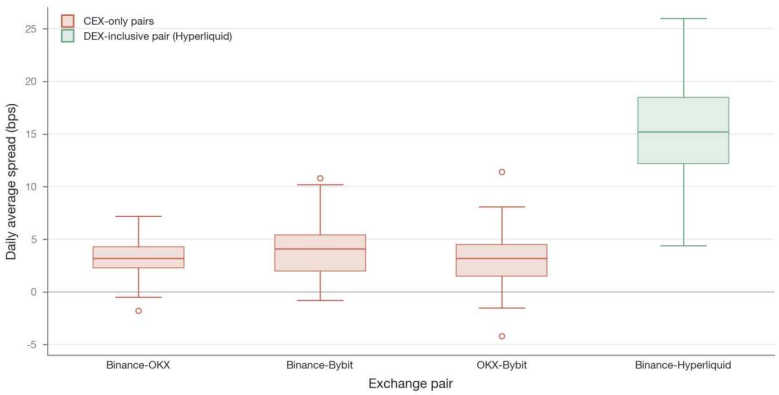


Figure 16-13. The characteristics of cross-exchange perpetual-futures spreads (Data source: Binance, OKX, and Bybit exchange API data [10])

Figure 16-13 shows the spreads of major exchanges' BTC perpetual futures relative to the industry benchmark Binance since early 2023, exhibiting three key features. First, spreads persist but their absolute level is now far below that of the Makarov-Schoar period—the median daily spread among purely centralized exchanges has narrowed to about 3-4 basis points, while pairs involving decentralized exchanges (such as between Binance and Hyperliquid) are higher, at about 15 basis points, and this stratification itself reflects the enormous overall improvement in market efficiency. Second, spreads widen briefly during periods of violent market fluctuation and fall back rapidly as arbitrage intervenes. Third, the systematic stratification of spreads across exchanges introduces the "core-periphery" structural model discussed below.

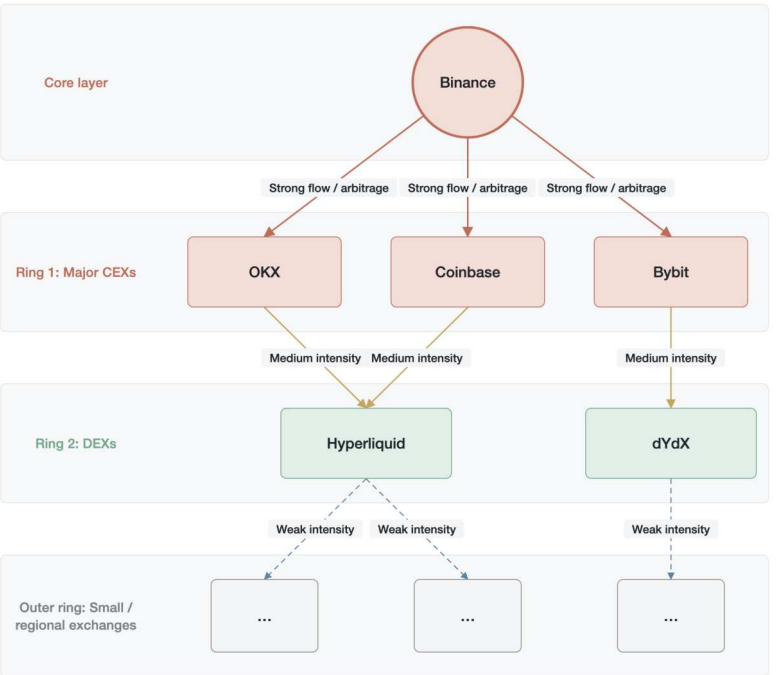


Figure 16-14. The core-periphery efficiency structure of cryptocurrency exchanges (Data source: a conceptual schematic constructed by the author)

The current efficiency structure of the global cryptocurrency market can be depicted as a "core-periphery" model. As the figure above shows, this structure consists of a small number of core "central markets" and a large number of "peripheral markets." The central markets—with Binance at the core and leading exchanges such as Coinbase next—have the highest liquidity, the strongest price-discovery capability, and the most complete infrastructure. Most of the world's price information and trading volume originates here; they are the "emission source" of global prices. The peripheral markets include other second- and third-tier exchanges, whose liquidity is thinner and whose price-discovery capability is weaker, and whose prices largely follow the movements of the central markets. Cross-exchange arbitrage assumes the price-transmission function in this structure: it rapidly propagates price movements from the central markets to the peripheral markets, ensuring that peripheral prices do not deviate too far from the center. The "efficiency distance" $d(e_i, e_c)$ of an exchange e_i from a central market e_c (such as Binance) can be measured along three dimensions: (1) the average absolute spread $\overline{|p_i - p_c|}$, which measures the magnitude of pricing deviation; (2) the spread half-life $\tau_{1/2}$, which measures the time required for the spread to converge from its peak to the equilibrium level and reflects the speed of information transmission; and (3) the Hasbrouck information share IS_i, which measures the exchange's contribution to global price discovery. The greater the efficiency distance (that is, the wider the spread, the slower the convergence, and the lower the information share), the more peripheral the exchange's position in the global efficiency hierarchy.

This core-periphery structure is a natural result of arbitrage competition. Because arbitrageurs always operate first between the markets with the best liquidity and lowest costs, the spreads among central markets are eliminated first and their efficiency is the highest. Arbitrage that connects the center with the periphery, or one peripheral market with another, involves higher slippage and risk costs, so its equilibrium spread is naturally higher. The efficiency of the entire market therefore exhibits a gradient distribution that declines from the center toward the periphery. The significance of this model is that it recasts cross-exchange arbitrage from simple "error correction" into a strategic force that shapes the entire market structure and efficiency hierarchy. Through the optimal allocation of their capital, arbitrageurs in effect define for the whole ecosystem what is "central" and what is "peripheral," and continuously maintain the dynamic stability of this hierarchy.

16.6 Triangular arbitrage

In the arbitrage ecosystem of perpetual futures, spot-futures-perpetual triangular arbitrage (hereafter "triangular arbitrage") is a composite strategy that simultaneously checks the pricing of three interrelated market tiers for the same asset: spot, traditional delivery futures, and perpetual futures. By constructing a three-point closed loop, the strategy captures pricing inconsistencies produced by the joint action of different term structures and financing costs.

When the same asset—Bitcoin, for example—simultaneously has an active spot market (such as Coinbase), a strictly regulated delivery-futures market (such as the CME's BTC futures), and a globally liquid perpetual-futures market (such as Binance's BTC/USDT perpetual futures), a multi-tiered pricing structure forms. In theory, the prices of these three markets should be tightly locked together by no-arbitrage relationships. Because they each have different participant structures, institutional designs, liquidity characteristics, and information-transmission paths, however, temporary inconsistencies among the prices inevitably arise. The work of the triangular arbitrageur is to identify and correct these inconsistencies within the three-market system, thereby integrating information dispersed across different "time layers" and "financial worlds" into a unified, internally consistent asset term structure.

After the United States approved spot Bitcoin ETFs in 2024, this mechanism gained unprecedented importance. Traditional financial institutions enter through compliant spot ETFs and CME futures, while crypto-native forces dominate the perpetual-futures market. Triangular arbitrage has therefore become the key information conduit and efficiency anchor connecting these two parallel but increasingly interwoven financial systems. This section dissects the theoretical foundation and operating logic of triangular arbitrage and, through cases, reveals how it has reshaped the term structure of crypto assets in the post-ETF era and become a key integrator of market efficiency.

16.6.1 Theoretical foundations

Triangular arbitrage can exist because two classic no-arbitrage pricing relationships coexist within an expanded market system and may come into temporary conflict. These

two relationships are the "flow anchoring" relationship connecting spot and perpetual futures, and the "point convergence" relationship connecting spot and delivery futures.

The first is the perpetual-futures "flow anchoring" relationship set out in Section 16.1.2, in which the funding rate mechanism anchors the perpetual-futures price near spot through continuous cash-flow payments:

> Relationship A: $P_{\text{perp}} \approx P_{\text{spot}} + \text{carry cost implied by the funding rate}$

The second is the traditional cost-of-carry model, which defines the no-arbitrage price of delivery futures through forced price convergence at maturity:

> Relationship B: $F_{\text{futures}} \approx S \cdot (1 + r \cdot T)$, where S is the spot price, r is the cost-of-carry rate, and T is the time to maturity

When the spot, perpetual-futures, and delivery-futures markets all exist simultaneously, a "triangular closed loop" forms. If the market were perfectly frictionless and efficient, Relationships A and B should be able to coexist in harmony. In reality, however, the financing cost implied by the perpetual futures under Relationship A (the funding rate) and the financing cost implied by the delivery futures under Relationship B (the futures basis) often diverge. We call this divergence the "triangular residual."

This residual has important economic meaning. It essentially reflects the market's different pricing of "short-term financing cost" (embodied in the high-frequency-settled funding rate) and "forward financing cost" (embodied in the one-off-at-maturity futures basis). For example, when market sentiment is extremely optimistic and large numbers of retail traders enter the perpetual-futures market to build long positions, the funding rate may rise to a high level (say, 30% annualized), while during the same period the CME futures market, dominated by institutional participants, may remain relatively calm, with its basis implying an annualized yield of only 15%. This 15% difference is the direction-neutral profit opportunity that triangular arbitrageurs can capture (as described in Section 16.4.2, such returns are not truly riskless but compensation for frictions such as execution lag, margin lock-up, and counterparty risk).

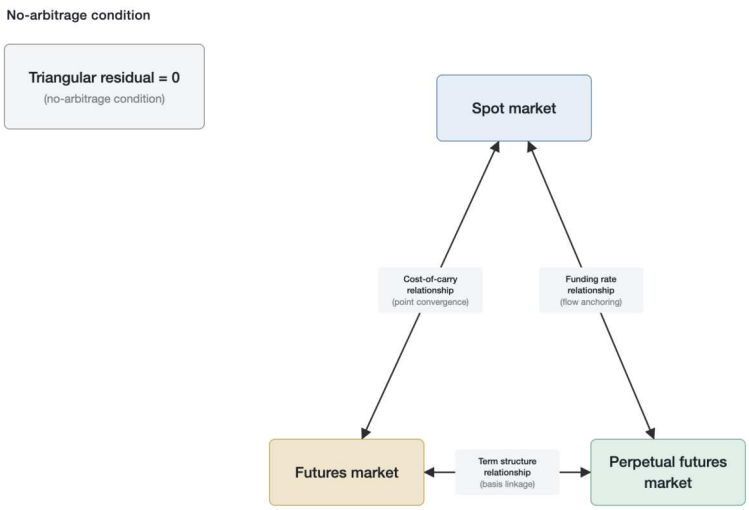


Figure 16-15. The three no-arbitrage relationships of triangular arbitrage (Data source: a conceptual schematic constructed by the author)

Figure 16-15 presents the three no-arbitrage relationships on which triangular arbitrage depends: the "flow anchoring" relationship connecting spot and perpetual futures (Relationship A), the "point convergence" relationship connecting spot and delivery futures (Relationship B), and the implied financing-cost consistency relationship between perpetual and delivery futures derived from the first two (Relationship C). When all three relationships hold simultaneously, no arbitrage opportunity exists; when any one of them deviates—that is, a "triangular residual" appears—the arbitrageur can capture riskless profit by constructing positions across the three legs.

16.6.2 Construction and execution logic

The construction logic of triangular arbitrage is essentially to build a portfolio with zero net risk exposure across three markets so as to lock in and extract the "triangular residual." Its core is to short simultaneously in the market that is overpriced (high implied financing cost) and go long in the market that is underpriced (low implied financing cost), hedging through the spot market.

Take the scenario mentioned earlier: suppose that at a given moment the funding rate of BTC perpetual futures, converted to annualized terms, is 30%, while the annualized basis of the same asset's CME delivery futures maturing in three months is only 15%. The arbitrageur identifies this persistent 15% annualized differential and constructs the arbitrage position along the following logic.

The arbitrageur first shorts a certain quantity (say, 1 BTC) of perpetual futures on a liquid exchange such as Binance, with the aim of collecting the annualized funding rate of as much as 30%. Because the funding rate is typically settled every 8 hours, the arbitrageur will continuously receive cash-flow income. At the same time, the arbitrageur goes long an equivalent amount of delivery futures maturing in three months

on the CME, paying the annualized basis cost of 15%; by buying the futures, the arbitrageur locks in a lower future purchase price and hedges the price risk of the perpetual-futures short. In theory, one perpetual-futures short and one futures long already constitute a delta-neutral portfolio, but in practice, to hedge risk more precisely and manage margin, the arbitrageur may hold a small amount of assets or cash in the spot market to cope with the unrealized profit and loss from price fluctuations.

Through this portfolio, the arbitrageur constructs a market-neutral position whose profit and loss no longer depend on the absolute rise or fall of the Bitcoin price and whose net return approximately equals the difference between the implied financing costs of the two markets: $30\% - 15\% = 15\%$ (annualized).

The theoretical return above is eroded by multiple constraints in actual execution. First, the standard contract size of CME Bitcoin futures is 5 BTC (with a micro contract of 0.1 BTC), whereas the minimum trading unit of Binance perpetual futures is 0.001 BTC, and this difference in contract specifications increases the difficulty of precise hedging. Second, CME trades from Sunday to Friday U.S. Eastern time, whereas Binance perpetual futures trade around the clock, which means the arbitrageur faces the risk of being unable to adjust one leg while the CME is closed. Furthermore, the CME's SPAN margin system is entirely incompatible with Binance's USDT margin system, so the arbitrageur must lock up margin separately at both ends, and the actual capital efficiency is far below the theoretical calculation. Finally, CME futures settle in U.S. dollar cash, whereas Binance perpetual futures settle in USDT, so there is a risk that USDT depegs from the dollar. Although the deviation is usually minimal, USDT has briefly deviated from its \$1 peg during periods of market panic, adding an extra risk dimension to cross-system arbitrage.

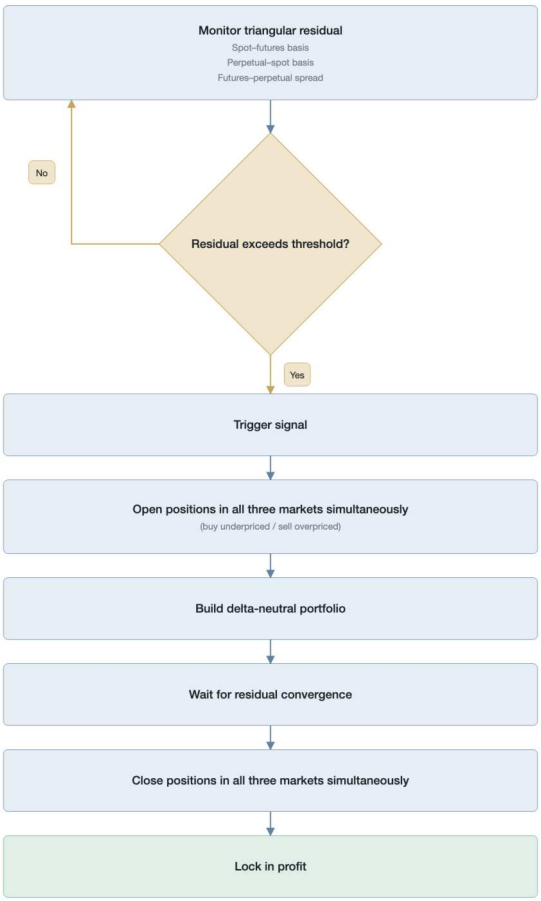


Figure 16-16. The execution decision flow of triangular arbitrage (Data source: constructed by the author)

The figure above presents, in the form of a decision flow, the execution logic of a complete triangular arbitrage. The flow begins with continuous monitoring of the triangular residual—that is, tracking the basis relationships among spot-futures, perpetual-spot, and futures-perpetual. When the residual exceeds a preset threshold, an arbitrage signal is triggered, and the arbitrageur simultaneously opens positions in the three markets, buying the relatively underpriced leg and selling the relatively overpriced leg, and constructing a delta-neutral portfolio to strip out directional price risk. Before the residual converges, the arbitrageur continues to hold the portfolio, collecting the cash flow corresponding to the difference between the implied financing costs of the two markets; once the residual converges, the arbitrageur simultaneously closes positions in the three markets, and the locked-in net return approximately equals the difference between the two markets' implied financing costs, less all execution costs. If the residual does not exceed the threshold, the flow returns to continuous monitoring.

This behavior is central to calibrating market efficiency. When large numbers of arbitrageurs execute this strategy, the short-side force in the perpetual-futures market

increases and exerts downward pressure on its price and funding rate, while the long-side force in the futures market increases and lifts the futures price, compressing its basis. Ultimately the implied financing costs of the two markets converge, the market's term structure is "flattened" again, and the triangular residual vanishes—until the next imbalance appears.

Entering 2024, the approval of spot Bitcoin ETFs brought triangular arbitrage an entirely new dimension and set of participants. A larger-scale, more compliant triangular structure formed: the secondary-market price of the spot ETF $\leftrightarrow$ CME Bitcoin futures $\leftrightarrow$ global perpetual futures. The authorized participants of the ETF are themselves professional arbitrageurs who, through the creation and redemption mechanism, ensure that the ETF price tracks its net asset value closely. When the ETF's secondary-market price, the CME futures price, and the global perpetual-futures price diverge, traditional financial institutions can construct more complex cross-market arbitrage strategies by trading futures on the CME and the ETF on the U.S. stock market while hedging with crypto-native funds in the perpetual-futures market. This established, for the first time at the institutional level, a direct, high-frequency channel of price discovery and efficiency transmission between traditional finance and crypto finance.

The actual smoothness of this transmission channel, however, is constrained by compliance limits across regulatory jurisdictions. Institutional investors regulated by the SEC or the CFTC in the United States are, in most cases, prohibited or strictly limited from trading perpetual futures on unregulated offshore crypto exchanges. This means that the two ends of the "CME futures $\leftrightarrow$ global perpetual futures" arbitrage chain fall under different regulatory jurisdictions, presenting a compliance obstacle for U.S. institutional participants. In practice, this role is more often assumed by crypto-native funds registered in jurisdictions with lighter regulation. This regulatory segmentation itself constitutes an institutional friction of triangular arbitrage and is one reason the cross-market efficiency gap persists.

16.6.3 Triangular arbitrage dynamics in the post-ETF era

In January 2024, the U.S. Securities and Exchange Commission approved the first batch of spot Bitcoin ETFs, marking the formal incorporation of crypto assets into the mainstream financial regulatory system. This event not only introduced hundreds of billions of dollars in incremental capital but also fundamentally reshaped the market's microstructure, creating in particular the previously nonexistent market conditions and complex dynamics for triangular arbitrage.

Before the ETFs' approval, triangular arbitrage was conducted mainly within crypto-native exchanges (such as Binance's spot, futures, and perpetual futures) or with a few offshore regulated exchanges (such as the CME), but both scale and participants were limited. The emergence of the ETFs established a high-capacity capital channel between the spot market and the traditional financial system, allowing institutional-grade capital to participate in a compliant, large-scale manner. According to data released by CME Group, after the ETFs' approval, leveraged funds' net short positions in CME Bitcoin futures increased significantly. This does not mean institutions are bearish on Bitcoin; on the contrary, it is evidence of large-scale basis trading and

triangular arbitrage: obtaining long exposure by buying the spot ETF while shorting futures on the CME to hedge risk and earn the basis.

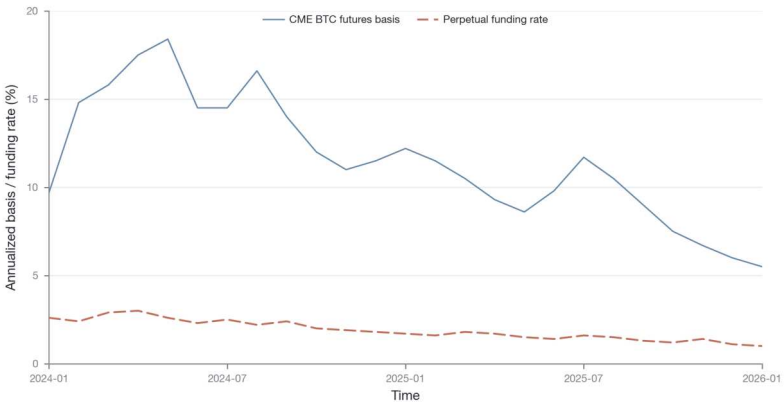


Figure 16-17. The evolution of the CME futures basis and the perpetual-futures rate (illustrative simulation: constructed from public basis/rate parameters of CME Group [11] and Binance [10], drawn by the author)

Figure 16-17 simulates the evolutionary path of the CME futures basis and the mainstream perpetual-futures funding rate from early 2024 to early 2026, from which the following key phenomena can be observed.

In the first half of 2024, strong market expectations of the ETFs and heavy capital inflows pushed the annualized basis of CME futures above 10%, forming a significant triangular-arbitrage space relative to the then relatively calm perpetual-futures funding rate (about 1-3% annualized), and large amounts of arbitrage capital accordingly entered the market to execute the "buy the spot ETF + short CME futures" strategy. As arbitrage activity intensified, market sentiment normalized, and the Federal Reserve's rate policy shifted from hiking to cutting, the CME futures basis was significantly compressed from the second half of 2024 through 2025: by early 2026, the annualized basis had fallen back to 4-5%, while the average perpetual-futures rate remained around 1%, and the potential profit space of triangular arbitrage narrowed from 5-9% in the early bull market to an equilibrium level of 3-4%. Meanwhile, a distinctive linkage formed between the ETFs' trading hours (U.S. stock-market hours) and perpetual futures' around-the-clock trading: during U.S. stock-market hours, large ETF creations and redemptions dominate price discovery and are rapidly transmitted through triangular arbitrage to the CME futures and global perpetual-futures markets; while the U.S. stock market is closed, price movements in the perpetual-futures market become a leading indicator of the next day's ETF opening price. Triangular arbitrageurs continuously perform the function of transmitting price information between these two asynchronous markets. The compression of the CME basis, however, is not driven entirely by arbitrage activity. Under the cost-of-carry model, the risk-free-rate component of the futures basis is directly affected by the Federal Reserve's policy rate. Between 2024 and 2026, the Federal Reserve's policy shift (from a hiking cycle into a cutting cycle) itself compressed the interest-rate component of the basis. The observed basis compression is therefore the joint result of arbitrage activity and the macro inter-

est rate environment, and attributing it entirely to improved arbitrage efficiency may overstate the independent contribution of arbitrage activity.

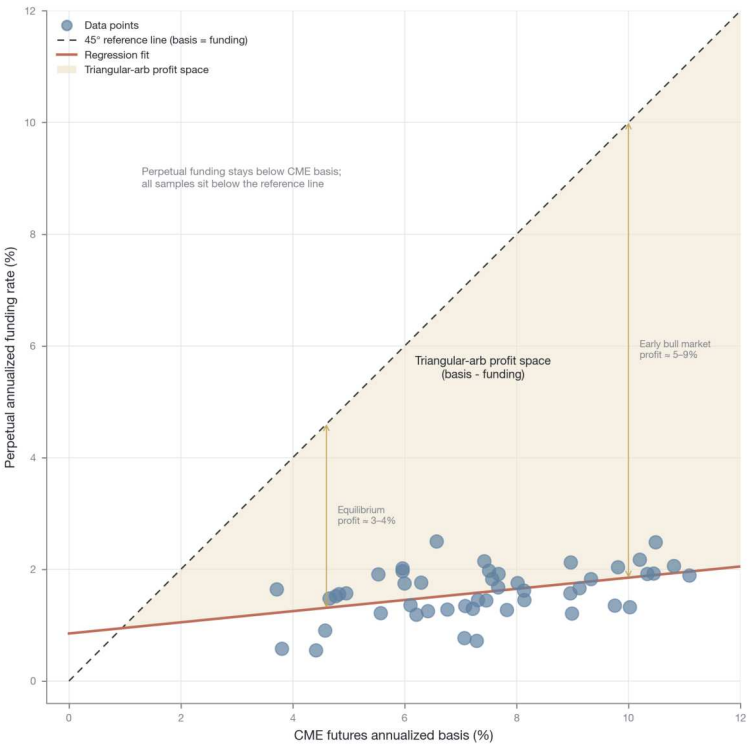


Figure 16-18. A comparison of the CME futures basis and the perpetual-futures rate (illustrative simulation: constructed from public basis/rate parameters of CME Group [11] and Binance [10], drawn by the author)

Figure 16-18 illustrates the efficiency improvement brought by arbitrage: the once-significant cross-market spread has been greatly compressed by arbitrageurs' profit-seeking behavior (again, this figure is an illustrative simulation rather than original data, and, as noted above, the basis compression is also driven by the macro interest rate environment). This confirms that arbitrageurs are not merely profit capturers but builders of market infrastructure that advance global pricing unification and liquidity integration.

16.6.4 Term-structure integration

The core efficiency function of triangular arbitrage is to serve as a systematic calibration mechanism that continuously integrates and unifies the term-structure information of crypto assets. The term structure—the pricing relationship among assets or contracts of different maturities—is a key gauge of a financial market's maturity. In an efficient market, short-, medium-, and long-term financing costs should form a smooth and logically self-consistent yield curve. Triangular arbitrage constitutes the micro force that shapes this curve.

First, by connecting perpetual futures (which can be viewed as a rolling, ultra-short-term contract) and delivery futures of various maturities (such as monthly and quarterly), triangular arbitrage anchors together the market's expectations of the short-term funding cost (the funding rate) and the forward funding cost (the futures basis). Without triangular arbitrage, the perpetual-futures market might see its funding rate deviate substantially from fundamentals for a long time because of short-term speculative sentiment, while the futures market might produce a distorted basis because of shifts in institutional risk appetite. The presence of triangular arbitrageurs forces these two price signals to calibrate each other, forming a more consistent implied yield curve from "now" (perpetual futures) to the "future" (futures).

Second, in the post-ETF era, triangular arbitrage performs the information-transmission function of connecting the two market systems of traditional finance and crypto finance. The CME futures market is dominated by institutional investors bound by U.S. regulation, and its pricing reflects the risk appetite, capital cost, and macroeconomic expectations of the traditional financial system. The offshore perpetual-futures market, represented by Binance and Bybit, is driven more by crypto-native funds and retail users worldwide, and its pricing more sensitively reflects sentiment, narratives, and capital flows within the crypto community. The participants, information sources, and pricing logic of these two markets differ significantly.

By operating simultaneously across the CME, spot ETF, and perpetual-futures markets, triangular arbitrageurs become the key medium for transmitting information. When a change in the Federal Reserve's macro policy affects the risk appetite of CME investors, this effect is transmitted through triangular arbitrage to the funding rate of perpetual futures. Conversely, when a new DeFi protocol or technological breakthrough sparks enthusiasm in the crypto community and drives up perpetual-futures prices, arbitrageurs also transmit this signal in reverse to the CME futures and spot ETF markets. This bidirectional information flow significantly improves the price-discovery efficiency of the entire crypto asset market, so that any local price shock can be absorbed and digested by the whole market more quickly.

In sum, triangular arbitrage is not merely a complex arbitrage strategy but a core driver of the crypto financial market's maturation and integration. By connecting the three key markets of spot, futures, and perpetual futures, it integrates the pricing logic of different participants, different institutional frameworks, and different information sources into a unified system, ultimately forming a more unified, more efficient global crypto-asset pricing system with smoother information transmission. This integration function, which crosses institutional and geographic boundaries, is a concentrated embodiment of this chapter's arbitrage infrastructure hypothesis.

16.7 Funding rate arbitrage

In the complex mechanism design of perpetual futures, the funding rate is not only the core tool for maintaining price anchoring but has itself evolved into a distinctive, tradable financial parameter. When the same asset displays persistent and significant funding rate differences across different trading platforms and market tiers, a crypto-

native, highly institution-driven form of arbitrage arises. Unlike arbitrage strategies that capture short-lived price deviations, the core of funding rate arbitrage is to capture and exploit the "institutional premium" caused by differences in exchange microstructure, participant composition, and leverage demand. This arbitrage activity not only provides traders with a relatively independent source of return but, at the macro level, plays the key role of smoothing the global rate structure and unifying market expectations of financing costs. This section dissects the micro-mechanics of the funding rate as a tradable asset, proposes the analytical framework of the "funding rate surface," and, through representative case studies of centralized and decentralized exchanges, reveals how funding rate arbitrage drives the efficiency calibration of the perpetual-futures ecosystem.

16.7.1 The tradability of the funding rate

When the same underlying asset shows different funding rates across different exchanges, the rate difference itself constitutes a tradable arbitrage instrument. Funding rate arbitrage captures persistent differences in institutional parameters by constructing a delta-neutral portfolio—shorting on the high-rate exchange and going long on the low-rate exchange. Unlike price arbitrage, this strategy typically has a longer holding period, which can last from hours to days.

Compared with traditional cross-exchange spread arbitrage, funding rate arbitrage is deeper along the strategy dimension. The former trades the deviation of "price," a zeroth-order variable, pursuing rapid convergence on the millisecond scale; the latter trades the "funding rate," a first-order institutional parameter derived from price, capturing the persistence of a state. Moreover, differences in the settlement frequency of the funding rate across exchanges (for example, every 8 hours, every 4 hours, or every 1 hour) add a "term-structure" dimension to the strategy, giving rise to a more complex "settlement-window arbitrage" strategy—that is, exploiting the mismatch of different settlement times to optimize returns and risk exposure. This trading logic, which turns the institutional parameter itself into an asset, is an important marker of the perpetual-derivatives market's maturation and growing complexity.

Funding rate arbitrage simultaneously faces multidimensional risk constraints. Because the arbitrage position involves cross-exchange holdings, execution risk, liquidity risk, and operational risk all constitute substantive constraints. In particular, when the arbitrageur must build positions simultaneously in two markets, rapidly changing market conditions can produce "leg risk" (the mechanism is described in Section 16.4.1). In addition, cross-exchange fund-transfer delays, exchange technical failures, or policy changes can all affect the execution of the arbitrage strategy. The successful execution of funding rate arbitrage therefore often depends on the arbitrageur's technological infrastructure, risk-management capability, and thorough understanding of market microstructure.

16.7.2 The funding rate surface framework

To understand and uncover cross-market and cross-asset funding rate arbitrage opportunities systematically, we can draw on the concept of an interest rate surface or a

volatility surface to construct an analytical framework of a "funding rate surface." This is a multidimensional information matrix that can intuitively display the "financing-cost topography" of the global perpetual-futures market.

We can formally define the funding rate surface as a three-dimensional mapping function $F : \mathcal{E} \times \mathcal{A} \times \mathcal{T} \to \mathbb{R}$, where $\mathcal{E} = \{e_1, e_2, \dots, e_m\}$ is the set of exchanges, $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ is the set of assets, and $\mathcal{T}$ is the time domain. $F(e_i, a_j, t)$ denotes the annualized funding rate that exchange e_i quotes for asset a_j at time t. There is an important difference from a traditional volatility surface, however: a volatility surface constructs a smooth surface by interpolation over the two continuous dimensions of strike and maturity, whereas both the exchange dimension and the asset dimension of the funding rate surface are discrete, with no natural basis for interpolation. Calling it a "funding rate matrix" might therefore be more precise mathematically, but the term "surface" helps emphasize the holistic perspective of analyzing arbitrage opportunities across multiple dimensions, so this chapter retains it.

Within this framework, we can define at least two core dimensions. The exchange dimension arranges the world's major perpetual-futures exchanges horizontally—such as Binance, OKX, and Bybit, as well as emerging decentralized platforms such as Hyperliquid and dYdX. The asset dimension arranges different crypto assets vertically, from the mainstream BTC and ETH, to popular public-chain tokens such as SOL and AVAX, to various long-tail assets. The value at each coordinate point on the surface represents the real-time or forecast funding rate of a specific asset on a specific exchange (typically expressed as an annualized percentage).

The "shape" of this surface contains extremely rich market-microstructure information. Along the exchange dimension, the "slope"—the rate differences across exchanges for the same asset—reflects differences in their user composition, leverage preferences, market depth, and risk-management mechanisms. For example, a retail-dominated platform may exhibit a higher funding rate than an institution-dominated platform during a bull market. Along the asset dimension, the "slope"—the rate differences across assets within the same exchange—reveals the degree of the market's short-term chase after specific narratives and hot spots. High rates typically cluster on the "star assets" with large recent gains and high market attention. The local "peaks" and "valleys" of the surface—its extreme points—are the direct embodiment of arbitrage opportunities. The difference between a pronounced "peak" (an exchange with a rate far above the market average for some asset) and a "valley" (a rate significantly negative or far below average) constitutes the potential profit space of an arbitrage trade.



Figure 16-19. A schematic of the funding rate surface (Data source: constructed by the author)

Figure 16-19 presents the funding rate surface at a given moment as a three-dimensional heatmap (the horizontal axis is the major exchanges, the vertical axis is different crypto assets, and the color depth represents the annualized rate): the surface exhibits a pronounced "peak-and-valley" structure, with some exchanges' rates for certain assets far above the market average (a "peak") or significantly below the average or even negative (a "valley"), and the difference between peak and valley is the potential profit space that arbitrageurs can capture.

The core work of arbitrageurs is to monitor this dynamically changing funding rate surface in real time and capture significant "peak-valley differences," smoothing the rate differences by constructing cross-platform, cross-asset hedged positions. Their collective action, in effect, smooths the entire surface, prompting the rate structure of the global perpetual-futures market toward greater balance and efficiency. This process can be understood as a market self-correction mechanism: when an exchange's rate is too high, arbitrageurs' shorting increases the short-side supply in that market and thereby depresses the rate; and vice versa. Through this spontaneous behavior of market participants, the funding rate surface gradually moves from a highly heterogeneous state toward equilibrium, ultimately achieving a balance of global financing costs.

16.7.3 The Hyperliquid case

To understand concretely how funding rate arbitrage operates in practice, and how different market microstructures give rise to arbitrage opportunities, we take Binance, the world's largest centralized exchange, and Hyperliquid, the fastest-growing decentralized perpetual-futures platform, for a comparative analysis. Their significant differences in user base, technical architecture, and mechanism design make them a representative comparative case for analyzing the logic of CEX-DEX rate arbitrage (the mechanism differences between the two are shown in Table 16-3).

The core differences between the two in the funding rate mechanism constitute the institutional basis of the arbitrage opportunity. In the settlement cycle, Binance is typically 8 hours and Hyperliquid is 1 hour. In rate calculation, Binance uses a premium index plus a fixed interest rate, whereas Hyperliquid includes only the premium index. In the execution environment, Binance uses a centralized matching engine, whereas Hyperliquid is based on a self-developed layer-1 blockchain that implements a fully on-chain order book. In addition, the difference in cross-platform fund-transfer delay constitutes an important operational constraint.

These mechanism differences directly cause the two platforms' funding rates to behave differently. The most salient is the mismatch of settlement cycles. Binance's 8-hour settlement cycle means its funding rate is a "blunted" reflection of market sentiment over the past 8 hours, whereas Hyperliquid's 1-hour high-frequency settlement makes its rate more sensitive to short-term market fluctuations. This temporal mismatch creates a "settlement-window arbitrage" opportunity for arbitrageurs. For example, in a rapidly rising market, an arbitrageur can anticipate that Binance will show a high positive rate at its next settlement point while Hyperliquid's rate may already have responded to a short-term price correction. The arbitrageur can short on Binance and go long on Hyperliquid in advance, capturing the pulse-like payoff at Binance's rate settlement.

Dimension	Binance	Hyperliquid
Settlement cycle	8 hours	1 hour
Rate calculation	Premium index + fixed interest rate (set to 0 for some pairs)	Premium index + damping coefficient
Execution environment	Centralized matching engine	Layer-1 on-chain order book
Fund transfers	Instant, internal	Cross-chain bridging
Maximum leverage	125x	50x (after March 2025, reduced to 40x for BTC and 25x for ETH)

Table 16-3. A comparison of the settlement mechanisms of Binance and Hyperliquid (data source: official documentation of Binance and Hyperliquid, data as of early 2025)

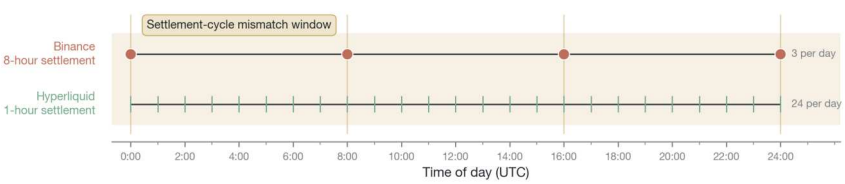


Figure 16-20. An analysis of the settlement-cycle mismatch (Data source: a conceptual schematic constructed by the author from the settlement parameters of Binance and Hyperliquid)

Figure 16-20 uses parallel timelines to compare the funding-rate settlement timing of Binance (8-hour settlement) and Hyperliquid (1-hour settlement): Hyperliquid responds to market changes at a higher frequency and rises first, whereas Binance's rate lags no-

ticeably because of its 8-hour "blunting window," and the region where the two curves are temporally offset marks the potential operating space for "settlement-window arbitrage."

Differences in the rate-calculation method and rate cap also widen the rate gap. Hyperliquid's rate calculation more purely reflects the long-short imbalance within its platform, and its extremely high rate cap allows it to show funding rates far above those of a CEX under extreme conditions, which directly creates a significant arbitrage space. Based on a high-frequency sample spanning 26 exchanges over several consecutive days, Zhivkov (2026) [12] found that the average annualized funding rate of BTC perpetual futures on Hyperliquid is systematically higher than on Binance, with a difference on the order of about 18 percentage points annualized (in that study's high-frequency sample, the annualized rates of the two were about 23% and 4.5%, respectively; note that this figure comes from that particular high-frequency observation window rather than the mean of a full quarter, and its specific numbers could not be fully re-verified through public channels, so it is best understood as an indicative order of magnitude of the structural difference). This structural rate difference provides arbitrageurs with a considerable source of profit and also reflects the fundamental difference in the two platforms' user composition and participant behavior. The causes of this structural rate difference are multidimensional. Besides the long-short imbalance caused by the higher proportion of retail and high-leverage traders in Hyperliquid's user base, there are at least two other important factors. First, the order-book depth of Hyperliquid is far below that of Binance (the resting order volume within a $\pm 2\%$ price range is typically only 10-20% of Binance's), which means a directional trade of the same size has a larger impact on the Hyperliquid perpetual premium. Second, the points and airdrop incentive mechanisms that Hyperliquid implemented during its growth phase may have artificially distorted trading behavior, with some traders deliberately maintaining high-leverage positions to earn airdrop points—an incentive-driven behavior that differs from natural supply and demand. The limitation in liquidity depth also means that arbitrage positions above a moderate size face significant slippage when executed on Hyperliquid, especially during periods of heightened market volatility. Therefore, although a rate difference of about 18 percentage points is extremely attractive on paper, there is a clear capacity ceiling on the size of the arbitrage position that can actually be executed with low friction.



Figure 16-21. A comparison of Binance and Hyperliquid BTC funding rates (indicative order of magnitude: drawn from the CEX-DEX structural rate difference described by Zhivkov 2026 [12]; about 23% versus 4.5%, a difference of about 18 percentage points, as an indicative magnitude, with the horizontal axis an indicative interval rather than the exact high-frequency sample window of that study)

Figure 16-21 compares the funding rate trajectories of Binance and Hyperliquid on BTC perpetual futures on an every-8-hours basis: the two curves move in the same direction, but Hyperliquid's rate level is higher than Binance's in most periods and its amplitude is larger, and when the difference between the two exceeds a threshold, an arbitrage window forms (the shaded area in the figure). Converted to an annualized basis, this systematic structural difference is on the order of about 18 percentage points (the about 23% versus about 4.5% mentioned above; see [12]), providing a source of profit for cross-platform rate arbitrage.

Executing CEX-DEX funding rate arbitrage, however, also faces distinctive challenges. The greatest friction comes from capital efficiency and operational risk: fund transfers between two CEXs are nearly instantaneous, whereas moving funds between a CEX and a DEX must go through a cross-chain bridge, and its delay, smart-contract security, and bridging-fee risks (see Section 16.9.1) directly erode the arbitrage profit space. This execution friction is the core reason a significant rate difference can persist for a long time between CEXs and DEXs, and it also constitutes a test of arbitrageurs' "infrastructure-service" capability. Only arbitrageurs with efficient cross-chain infrastructure and the ability to execute and manage risk rapidly can earn sustained profits in this market.

16.7.4 Financing-cost calibration

As a spontaneous market-correcting force, the core efficiency function of funding rate arbitrage is to "smooth" the funding rate surface proposed earlier, compressing rate outliers across exchanges and assets and thereby calibrating the implied financing cost of perpetual futures globally. When arbitrageurs short on rate "highlands" (such as Hyperliquid) and go long on rate "lowlands" (such as Binance), their behavior directly increases the short-side supply in the high-rate market and the long-side demand in

the low-rate market. These two forces act together to push overly high rates down and overly low rates up, ultimately driving the market's rate level toward an equilibrium value.

Empirical research provides strong evidence for this mechanism. Although execution frictions prevent the rate difference from being eliminated entirely, arbitrage activity significantly limits the magnitude and persistence of the difference's expansion. Zhivkov (2026) [12] found that in the observed sample, as much as 17% of observations exhibited an economically significant (greater than 20 basis points) CEX-DEX funding rate arbitrage opportunity; however, only about 40% of the top opportunities still generated positive returns after transaction costs and spread reversals, and as many as 95% of the opportunities ultimately ended in a forced exit. This shows that funding rate arbitrage is not a "riskless," costless return; its persistent positive return is compensation for the costs and risks that arbitrageurs bear in overcoming multiple frictions to provide the service of market integration.

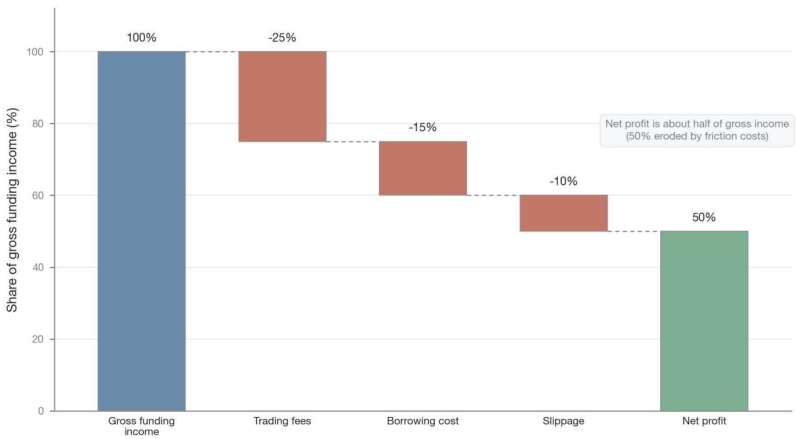


Figure 16-22. A decomposition of funding rate arbitrage returns (the cost-component proportions are illustrative values from the author's model, with parameters referencing the BTC perpetual-futures rate structures of Binance and Hyperliquid; the related arbitrage-opportunity and forced-exit proportions in the text are from Zhivkov 2026 [12])

Figure 16-22 shows the return and cost composition of the funding rate arbitrage strategy: the total rate income yields a net return only after a series of friction costs—trading fees, borrowing costs, and slippage—are deducted, and the net return can quickly turn negative when the rate difference narrows or slippage spikes. This shows that funding rate arbitrageurs provide the public good of "efficiency calibration" in a high-uncertainty environment and earn a risk premium for doing so.

There is also an important synergy between funding rate arbitrage and spot-perpetual basis arbitrage. Rate arbitrage brings the funding rates of exchanges worldwide into closer alignment, which provides basis arbitrageurs with a more stable and predictable cost environment. When the financing costs (funding rates) of all platforms are close, basis arbitrageurs can focus more on the price deviation between perpetual futures and spot, without worrying too much about the extra risk from rate fluctuations across

different platforms, which improves the efficiency and robustness of basis arbitrage—the core anchoring mechanism. Funding rate arbitrage thus constitutes a secondary calibration system that, by finely tuning the financing costs across markets, provides a solid foundation for the effective operation of the primary price-anchoring mechanism.

At a more macro level, the existence and effective operation of funding rate arbitrage reflect a trend of institutionalization and efficiency improvement in the crypto-derivatives market. When enough capable arbitrageurs are present in the market, the market's microstructure becomes more efficient, the price-discovery function becomes more complete, and participants' financing costs become fairer and more transparent. This is a core embodiment of this chapter's arbitrage infrastructure hypothesis: through their self-interested trading behavior, arbitrageurs in effect improve the efficiency of the entire market, so that the crypto-derivatives market gradually evolves into a more mature and more efficient financial ecosystem.

16.8 Cross-asset statistical arbitrage

Cross-asset statistical arbitrage maintains the internal consistency of the pricing system along the asset dimension. This strategy focuses on the dynamic balance of the relative relationships among different assets; by constructing market-neutral positions between highly correlated assets, it ensures that the pricing of related assets does not deviate from the intrinsic linkage jointly determined by fundamentals, technology adoption, and market narratives.

Unlike strategies that rely on deterministic no-arbitrage relationships, statistical arbitrage is essentially a probability-based game. It does not pursue riskless-ness on any single trade but captures temporary pricing deviations through a large number of trades, in the belief that these deviations will ultimately revert to their statistical mean. In the violently volatile and narrative-driven crypto market, the effectiveness of this strategy depends not only on complex mathematical models but, more fundamentally, on the distinctive institutional advantages that perpetual futures provide as a financial instrument. The institutional characteristics of perpetual futures have allowed the classic pairs-trading strategy to evolve into a new and more powerful form in the crypto world.

16.8.1 Pairs trading in perpetual futures

The core logic of pairs trading is to identify pairs of assets with a stable cointegration relationship and, when the spread deviates from the mean, to take opposing positions, betting on the mean reversion of the spread. This strategy is market-neutral and in theory bears no systematic beta risk.

In the crypto asset market, pairs trading has broad applicability. The most typical pair is Bitcoin and Ethereum. This combination comprises not only the two largest crypto assets by market capitalization but also represents the relative valuation of two different functional positionings—store of value and smart-contract platform. The fluctuation of their spread largely reflects shifting market preferences between the two narratives of store of value and decentralized-application platform. Besides BTC-ETH, the spread

dynamics of layer-1 (L1) public-chain competitor pairs (such as ETH and Solana) reflect the market's relative valuation of different public-chain ecosystems in terms of technical performance, developer activity, and user growth. The spread relationships of same-ecosystem token pairs (for example, different application tokens within the same layer-2 ecosystem, or protocol tokens with similar functions in DeFi, such as AAVE and COMP) are directly affected by internal ecosystem developments and the popularity of specific niches.

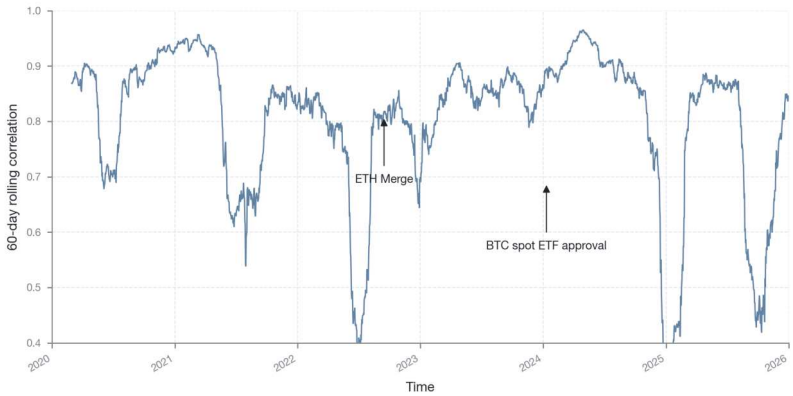


Figure 16-23. The evolution of the BTC-ETH rolling correlation (Data source: Binance exchange API data [10])

Figure 16-23 tracks the evolution of the 60-day rolling correlation coefficient between BTC and ETH: it remains in the high range of 0.7 to 0.9 in most periods (sharing significant systematic-factor exposure), but drops sharply—even briefly below 0.5—at key junctures such as major Ethereum upgrades, DeFi booms, or macro risk events. These moments of "correlation breakdown" are both the greatest source of risk for pairs trading and key signals for re-evaluating the hedge ratio and the cointegration relationship.

Gatev et al. (2006) [13], in a systematic study of pairs-trading strategies in the U.S. stock market, showed that the strategy still earns significant excess returns net of transaction costs, confirming the economic substance of the mean-reversion phenomenon. In the crypto market, the noise-trader risk described by DeLong et al. (1990) [14] (the risk that irrational traders' behavior may cause the spread to widen further before it converges) poses an additional challenge to pairs trading. The key to successfully constructing a pairs-trading strategy is to distinguish genuine statistical association from spurious, accidental correlation. Research shows that relying on high correlation alone is dangerous, because it may merely reflect two assets moving in the same direction under a common external factor (such as the macroeconomic environment) during a particular period—a relationship that can break at any time. A more robust statistical basis is cointegration. A cointegration relationship indicates that even if the price series of the two assets are individually non-stationary, some linear combination of them (the spread) is stationary, fluctuating around a long-run equilibrium value. This relationship implies a genuine, long-run economic link between the two, providing a more

reliable guarantee of mean reversion. Sophisticated statistical arbitrageurs therefore use cointegration tests (such as the Engle-Granger test or the Johansen test [15]) to screen trading pairs and combine them with indicators such as the Z-score to gauge the degree of the spread's deviation, thereby deciding the timing of entry and exit.

Methodologically, pairs trading in the crypto market has evolved from the classic static-regression framework to more refined dynamic-modeling methods. Early practitioners typically used the Engle-Granger two-step method to estimate the cointegration relationship, but this method implicitly assumes constant parameters and struggles to capture the rapid evolution of the correlation structure among crypto assets. The current frontier practice is to introduce state-space models: through the Kalman filter, the hedge ratio and the spread mean are estimated in real time, allowing the model to adapt to dynamic changes in market microstructure. For possible structural breaks (for example, when a public chain undergoes a qualitative change in its valuation relationship with a competing chain because of a major technical upgrade), the Markov regime-switching model [16] allows the mean and volatility of the spread to switch between different market states, giving traders a more robust risk-control framework. In addition, the vector error correction model, when multiple pairs of cointegrated assets exist, can simultaneously model the dynamic adjustment of multiple spreads, capturing the linkage effects within a network of assets. Interpretable machine-learning methods have also recently entered the feature analysis of crypto-market microstructure—for example, Bieganowski and Ślepaczuk (2026) [17] used gradient-boosting models and SHAP attribution to systematically characterize the microstructural features of the perpetual-futures limit order book and their predictive power for short-term returns, providing new methodological tools for signal construction and feature selection in statistical arbitrage. The choice among these methods is not a purely technical matter but directly determines a strategy's survivability in the crypto market's environment of high volatility and rapid narrative shifts.

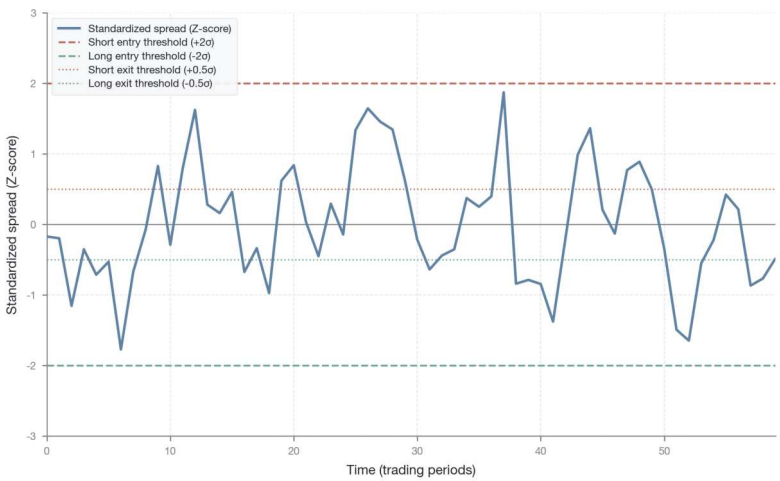


Figure 16-24. A schematic of spread reversion in pairs trading (Data source: constructed by the author)

Figure 16-24 schematically depicts the fluctuation and reversion of the spread of a cointegrated asset pair around its long-run mean: when the spread deviates from the mean by more than a set threshold (typically expressed as a multiple of the standard deviation via the Z-score), a trading signal is triggered—when the spread is too high, short the spread (short the relatively strong asset and long the relatively weak one), and when the spread is too low, long the spread. The figure uses threshold lines to mark the entry ($\pm 2\sigma$) and exit ($\pm 0.5\sigma$) triggers.

16.8.2 Differences between perpetual and spot statistical arbitrage

Although the concept of pairs trading can be applied to any market, implementing it in the perpetual-futures market offers a series of significant institutional advantages over the traditional spot market. These differences fundamentally change the strategy's risk-return profile and execution efficiency, making perpetual futures a near-ideal venue for statistical arbitrage, and pairs trading in particular.

First is the convenience of shorting. Shorting crypto assets in the spot market faces frictions such as high borrowing costs and limited quotas, whereas shorting perpetual futures is nearly frictionless, significantly lowering the execution threshold and cost of pairs trading.

Second, the funding rate mechanism adds a distinctive dimension of return or cost to pairs trading. For a pairs trader, the difference between the funding rates on the two legs directly constitutes an extra cash flow for the strategy. In the ideal case, if the funding rate on the long leg is negative and that on the short leg is positive, the trader can earn the funding rate on both legs at once, and this income can sometimes even exceed the profit from spread convergence. Conversely, an unfavorable rate difference becomes a continuous cost. The dynamic changes of the funding rate must therefore be treated as a core variable in model construction.

Third, the availability of high leverage is another important feature of the perpetual-futures market, and its effect cuts both ways. Pairs trading captures relatively small spread fluctuations, and high leverage can significantly amplify the returns from these tiny movements. Leverage, however, likewise amplifies risk. If the spread does not converge as expected but instead keeps diverging—the so-called "breakdown of the pair relationship"—high leverage will subject the position to enormous losses or even forced liquidation. The use of leverage must therefore be combined with a strict risk-management system. Particular caution is warranted regarding the asymmetric impact of the forced-liquidation mechanism, unique to perpetual futures, on statistical-arbitrage strategies. Even if the long-run cointegration relationship between two assets still holds, an extreme short-term divergence of the spread may trigger the forced liquidation of one or both legs, forcing the strategy to exit at the most unfavorable moment. More seriously, once one leg is liquidated, the originally market-neutral position instantly becomes a one-sided naked exposure, and this forced liquidation may further drive the spread's divergence, forming a positive-feedback liquidation cascade.

Finally, the crypto market's around-the-clock trading allows pairs-trading strategies to run at all hours and capture opportunities whenever they arise, unlike traditional markets constrained by opening and closing times. This also places higher demands on

the automation of the trading system and the real-time nature of risk monitoring. The differences between the two types of statistical arbitrage are shown in Table 16-4.

Dimension	Spot statistical arbitrage	Perpetual-futures statistical arbitrage
Ease of shorting	Restricted / high borrowing cost	Frictionless
Additional return dimension	None	Funding rate differential
Leverage availability	Limited	High leverage available
Trading hours	Limited	24/7
Settlement risk	T+2	Real-time
Capital efficiency	Low	High

Table 16-4. Perpetual-futures versus spot statistical arbitrage (Data source: constructed by the author)

16.8.3 Maintaining pricing consistency

The core function of cross-asset statistical arbitrage in the crypto market is to act as a decentralized regulating mechanism that continuously maintains the internal consistency of the entire asset-pricing ecosystem. Through the constant calibration of "relative value," it connects different assets into an interrelated, logically self-consistent value network. This efficiency function manifests at three levels.

First, it is a transmission mechanism for price discovery among assets. When one asset (such as BTC) rises in price because of some external information shock, a highly correlated asset (such as ETH) may not react immediately with an equal-magnitude move, causing the spread between the two to deviate from equilibrium. At this point, statistical arbitrageurs quickly step in, going long ETH perpetual futures and shorting BTC perpetual futures. This trading behavior amounts to "transmitting" the new information in the BTC market to the ETH market, increasing buying pressure on ETH and selling pressure on BTC, thereby accelerating ETH's "lagged rise" and BTC's "price correction" and ultimately pushing the spread between the two back toward equilibrium. In this process, arbitrageurs not only profit themselves but also, in effect, improve the information-transmission efficiency of the entire market.

Second, it integrates and reinforces cross-asset liquidity. Active pairs-trading activity simultaneously increases the trading volume and order-book depth of the perpetual futures of both underlying assets. A trader seeking to build a large position in ETH might find that placing the order directly would cause considerable market impact. But if active BTC-ETH pairs trading exists, the trader can indirectly affect ETH's price and liquidity by trading BTC. This liquidity "spillover effect" means that a single asset's liquidity is no longer isolated but can be supplemented and enhanced through the trading of related assets, thereby improving the depth and resilience of the entire market.

Third, it provides a dynamic reference frame for the valuation system of crypto assets as a whole. By observing the spread levels and fluctuation ranges of different asset pairs, market participants can intuitively grasp the relative strength of various assets

and market sentiment. For example, when the spreads of L1 public-chain pairs (such as ETH-SOL) narrow broadly, it may mean the market is re-evaluating the value of different public chains and the competitive landscape is intensifying. And when the pair spreads across the entire market widen broadly, it may signal rising risk-aversion sentiment, with investors inclined to hold leading assets (such as BTC) and sell other, less correlated assets. The spread data generated by statistical-arbitrage activity thus becomes a valuable indicator of market sentiment and risk appetite in its own right.

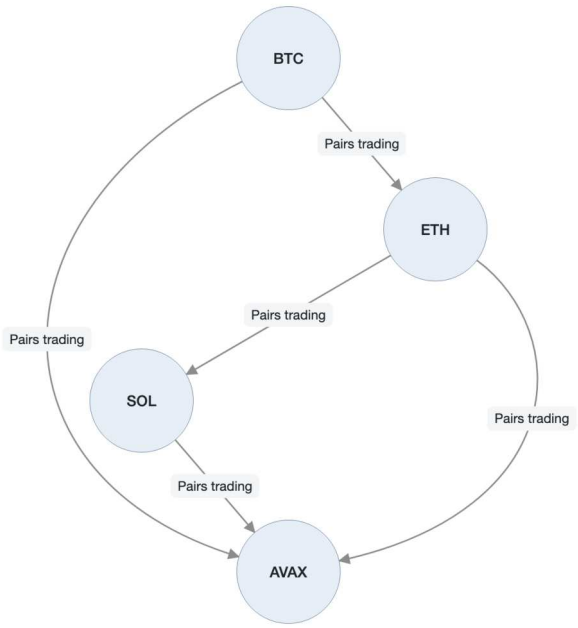


Figure 16-25. Relative-pricing maintenance in the asset-pricing ecosystem (Data source: a conceptual schematic constructed by the author)

Figure 16-25 uses a network topology to show the mechanism by which statistical arbitrageurs maintain relative pricing: each node represents a crypto asset, and the links between nodes represent the relative-pricing relationships that arbitrageurs maintain, in which mainstream assets such as BTC and ETH form the core pairs with the deepest liquidity and the most active arbitrage, while arbitrage coverage of long-tail assets is relatively limited—revealing how arbitrageurs weave isolated asset pricings into a mutually calibrated value network.

In sum, cross-asset statistical arbitrage is an inevitable result of the crypto market's maturation. It plays an indispensable role along the "asset dimension" of the efficiency spectrum, weaving isolated asset pricings into a dynamically balanced, logically self-consistent ecosystem through countless captures and corrections of tiny spreads. This strategy is not without risk, however, and its greatest intrinsic risk lies in the fragility of correlation itself. When the market undergoes a structural shift, historically formed statistical relationships may be broken entirely. In the crypto market, the

triggers of such breakdowns have a distinctive narrative-driven character: an "ETH killer" narrative can completely change the ETH-SOL spread dynamics within weeks, and a single social-media wave can instantly dissolve a long-stable pair relationship. This narrative-driven structural break is fundamentally different from the gradual relationship evolution caused by fundamental changes in traditional markets: it is faster, larger in magnitude, and less predictable, posing a fundamental challenge to statistical models based on historical data and causing the strategy to suffer major losses. This will be developed further in the discussion of arbitrage risk in Chapter 17.

16.9 CEX-DEX cross-tier arbitrage

In the arbitrage ecosystem of perpetual futures, cross-tier arbitrage between centralized and decentralized exchanges is the most active arbitrage domain, where the structural contradictions of the crypto market intertwine with frontier innovation. As two starkly different paradigms of financial infrastructure, CEXs and DEXs create, through their inherent architectural differences, a persistent and dynamically evolving efficiency gap. Arbitrageurs continuously monitor and exploit the spread opportunities that arise from differences in information latency, transaction costs, and consensus mechanisms. The core argument of this section is that CEX-DEX cross-tier arbitrage is not only the last of the six core arbitrage mechanisms but also an important analytical vantage point, revealing the deep tension between centralized and decentralized financial infrastructure, the trade-off between efficiency and fairness, and an ongoing arbitrage-driven contest of technology and mechanism.

16.9.1 The sources of the efficiency gap

The efficiency gap between the CEX and DEX perpetual-futures markets is not accidental but stems from fundamental differences in their technical architecture, information transmission, and trade-execution mechanisms. This difference is structural, and it defines the conditions under which arbitrage opportunities arise and the forms they take. As the efficiency analysis of Chapter 15 and Zhivkov (2026) [12] reveal, the price-discovery integration of the CEX ecosystem (measured by the Hasbrouck information share) is 61% higher than that of the DEX ecosystem, and over the study's sample period all statistically significant information flows were transmitted one-way from CEXs to DEXs, with no significant reverse causality detected. As DEX trading volume grows rapidly—Hyperliquid's daily volume repeatedly exceeded $10 billion in the second half of 2025—this one-way dominance may be evolving, and the price-discovery contribution of DEXs deserves continued attention, especially in DEX-first-listed assets and long-tail tokens. This significant efficiency gap constitutes the macro backdrop against which cross-tier arbitrage can exist. We can decompose the sources of this efficiency gap into the following core layers.

First, the most central source is oracle latency. Different DEX perpetual-futures protocols vary significantly in their degree of dependence on oracles, and they can be divided into three categories. The first is fully oracle-dependent, exemplified by early virtual automated market maker (vAMM) protocols, whose internal pricing is determined almost entirely by external oracle price feeds and whose arbitrage window is the largest. The

second is oracle-assisted, exemplified by Hyperliquid, whose price discovery is driven by the real-time order flow of an on-chain central limit order book (CLOB), with the oracle used mainly for mark-price calculation and liquidation triggers, so that its arbitrage window depends on the degree of synchronization between on-chain order flow and the CEX price. The third is hybrid, such as dYdX v4, which combines on-chain order-book data with an external oracle to construct the mark price. Although the architectures of these protocols differ significantly, they are all affected, to varying degrees, by oracle-update delay or the lag of on-chain price discovery. Updating external-world (mainly CEX) price information onto the chain takes time, however. This process involves data aggregation, network transmission, on-chain transaction packing, and consensus confirmation, and its delay ranges from a few hundred milliseconds to several seconds—or even minutes during network congestion. When the CEX price moves briefly because of major news or a large order, the DEX oracle price inevitably lags, forming a brief but deterministic window of price deviation, which constitutes the arbitrageur's main operating space.

Second, the inherent frictions of on-chain transactions are another key factor in the DEX's efficiency shortfall relative to the CEX. Executing a trade on a DEX requires the full process of broadcasting the transaction, entering the mempool, being packed into a block by validators, and finally waiting for block confirmation. This process is not only time-consuming (depending on the block time and finality of the underlying blockchain) but also requires paying a gas fee. By contrast, CEX trade execution occurs in the memory of its centralized servers, the matching engine can complete order matching in microseconds, and transaction costs are typically lower and more predictable. Therefore, even if an arbitrageur discovers an opportunity on a DEX, its execution faces the dual constraints of time delay and cost uncertainty.

Third, when arbitrage involves DEXs on different blockchains, cross-chain bridging delay and risk become a new obstacle. Moving funds between a CEX and DEXs deployed on different chains (such as Arbitrum, Solana, or Base) must go through a cross-chain bridge. This process can take minutes or even longer, and the bridge itself carries major security risks (according to Chainalysis, cross-chain bridge attacks caused losses of more than $2.5 billion between 2021 and 2023, including major incidents such as the Ronin bridge at $625 million, Wormhole at $320 million, and Nomad at $190 million). For CEX-DEX arbitrageurs, using a cross-chain bridge to move funds entails not only delay risk but also substantive smart-contract security risk, and this cost must be incorporated into the expected-return calculation of the arbitrage. This friction in capital flow prevents arbitrageurs from moving funds as quickly as they can between CEXs, thereby limiting the scale and efficiency of arbitrage activity.

Finally, the transparency of DEXs brings a distinctive paradox. Unlike the "opaque" order book of a CEX, the state of a DEX's liquidity pools, and even the pending transactions in the mempool, are fully public. This transparency, on the one hand, provides arbitrageurs with clear arbitrage targets, but on the other hand exposes their trading intentions to more technically advantaged competitors—MEV searchers. An arbitrage trade targeting oracle latency, once it enters the mempool, may be front-run by an MEV bot through a "sandwich attack," so that the original arbitrageur's profit is eroded or the transaction fails outright. This complex on-chain game further increases

the execution difficulty and risk of CEX-DEX arbitrage, forming a highly adversarial on-chain environment.

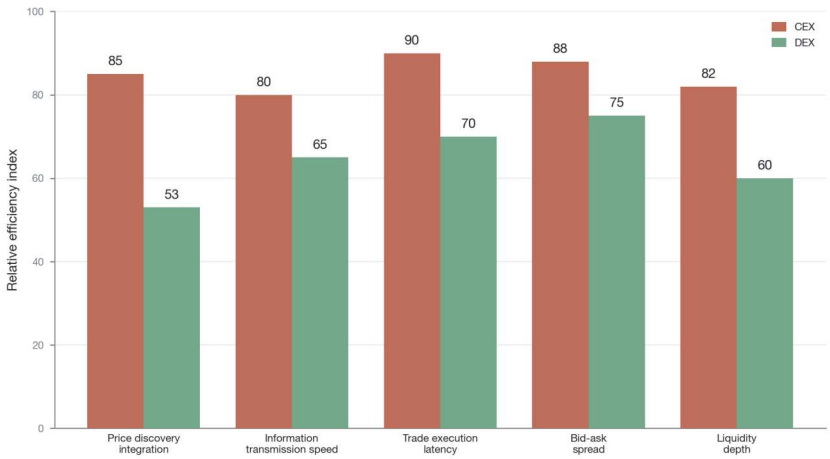


Figure 16-26. A comparison of CEX and DEX efficiency (for the "price-discovery integration" dimension, the CEX is about 61% higher than the DEX, based on Zhivkov 2026 [12]; the other dimensions are representative scores constructed by the author from Chapter 15 and Section 16.9)

Figure 16-26 intuitively shows the efficiency differences between CEXs and DEXs across several dimensions: whether in the integration of price discovery or the one-way dominance of information flow, CEXs display a clear advantage. These structural frictions allow the CEX-DEX efficiency gap to persist, constituting an emerging arbitrage domain that combines challenge and opportunity.

16.9.2 Oracle latency and OEV

The core of CEX-DEX cross-tier arbitrage lies in exploiting the structural efficiency gap described above to buy low and sell high between the two market tiers. Its most classic and common mode is oracle-latency arbitrage. Consider the following scenario: because of a piece of good news, the price of Bitcoin rises rapidly on Binance from \$50,000 to \$50,500. At that moment, however, the oracle price of a DEX perpetual-futures protocol deployed on Arbitrum remains at \$50,000. After monitoring this deviation, the arbitrageur (typically an automated bot) immediately executes the following operations.

The arbitrageur opens a Bitcoin perpetual-futures long position on the DEX at the "stale" oracle price of \$50,000 and, almost simultaneously, opens an equivalent Bitcoin perpetual-futures short position on Binance at the real-time price of \$50,500.

Having completed these two steps, the arbitrageur has constructed a market-neutral position. When the DEX oracle price finally updates to \$50,500, the arbitrageur can close the long on the DEX while closing the short on the CEX, locking in a spread profit of about \$500 (less transaction costs). The source of profit in this process is essentially the short-lived pricing error produced by the DEX's failure to reflect fair market value in time. This arbitrage profit, arising from the oracle-update mechanism, is defined in the DeFi space as oracle extractable value (OEV) [18].

As Daian et al. (2020) [19] revealed in their pioneering study, transaction ordering on decentralized exchanges itself contains extractable value. Weintraub et al. (2022) [20] further provided a systematic measurement of MEV in private transaction ordering, confirming the prevalence and scale of this value-extraction phenomenon. The legal characterization of MEV activity remains contested. In traditional finance, front-running clearly constitutes illegal conduct; on-chain MEV, because of the public nature of the blockchain mempool, is more ambiguous at the legal boundary. However, the criminal charges the U.S. Department of Justice brought in 2023 against Avraham Eisenberg for market manipulation (the Mango Markets incident) indicate that enforcement agencies are attempting to extend the legal framework for traditional market manipulation into the DeFi domain (Eisenberg was convicted by a jury in April 2024, but that criminal conviction was vacated by the court in May 2025, while the civil actions brought by the SEC and the CFTC remain ongoing; this back-and-forth itself shows that the legal boundaries of on-chain market manipulation are still contested and in formation). From the standpoint of efficiency analysis, it is necessary to distinguish "benign MEV" (such as pure oracle arbitrage, whose effect is to accelerate the convergence of the DEX price with the global market) from "malicious MEV" (such as a sandwich attack, whose effect is to extract value from end users); the two have opposite effects on market efficiency. This chapter uses "the net direction of the effect on market efficiency" as the criterion for this dichotomy—that is, whether the behavior drives prices toward global fair value without harming informed users; note that this is a simplified classification, and in reality there exists an intermediate zone that is hard to categorize cleanly. Oracle extractable value (OEV) is an important subset of MEV, referring specifically to value captured by exploiting the timing gap or content difference in oracle-data updates. In the example above, the $500 spread is a typical instance of OEV. By trading ahead of the oracle update, the arbitrageur successfully "extracts" value that should belong to the protocol or the liquidity providers.

This arbitrage can be divided into two types. Atomic arbitrage occurs on the same chain (for example, between a DEX's spot pool and another DEX's perpetual futures), where the arbitrageur can use a flash loan to complete all the "borrow-trade-repay" steps within a single atomic transaction; because the transaction either succeeds as a whole or fails and rolls back as a whole, almost no own capital is put at risk. CEX-DEX arbitrage, by contrast, belongs to non-atomic / cross-chain arbitrage: because it involves two independent systems (one off-chain, one on-chain), the transaction cannot be completed in a single atomic operation, and the arbitrageur must execute on one end first and then hedge on the other, with a time gap in between. If the market price moves adversely before the hedge is completed, or if the on-chain transaction fails because of network congestion, the arbitrageur faces risk exposure. The study by Zhivkov (2026) [12] also confirms that even when a significant spread exists, a large number of arbitrage opportunities still result in losses because of spread-reversal risk.

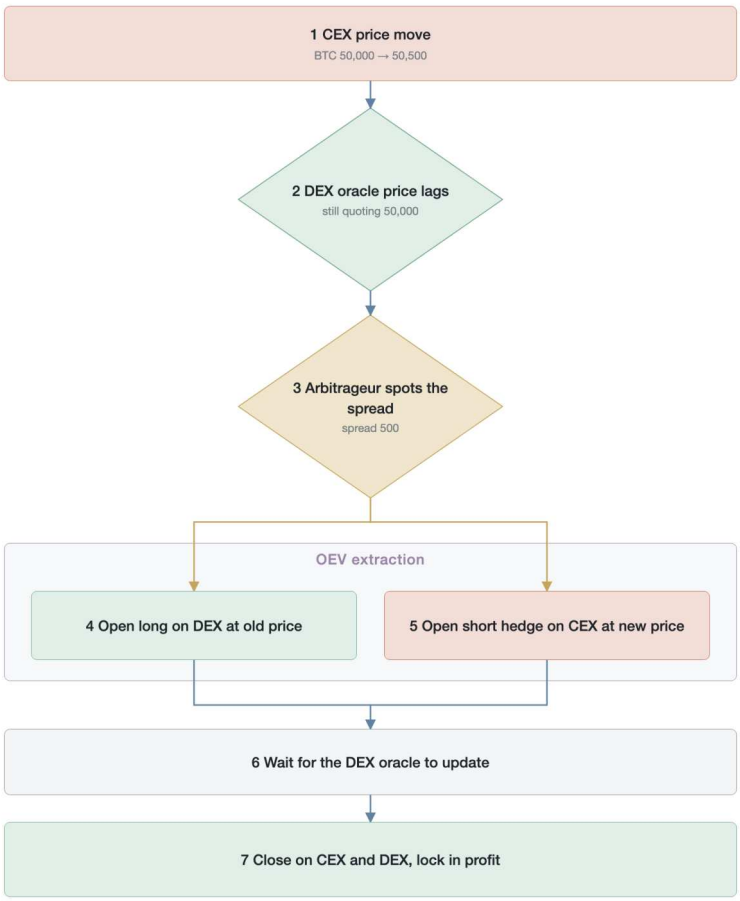


Figure 16-27. The CEX-DEX arbitrage workflow and OEV extraction (Data source: a mechanism schematic constructed by the author)

Figure 16-27 clearly depicts the complete flow of oracle-latency arbitrage and the arbitrageur's trade-off among return, cost, and risk: from discovering the spread to locking in the profit, each step involves a precise calculation of latency, cost, and execution certainty, at whose core is the extraction of OEV—which in turn introduces the institutional-design trade-off between efficiency and fairness.

16.9.3 The efficiency function and the LP-loss problem

At the macro level, the profit-seeking behavior of CEX-DEX arbitrageurs performs an important calibrating function for market efficiency. When the DEX price deviates from the global market because of oracle latency, arbitrageurs' intervention quickly applies opposing trading pressure (for example, buying when the price is too low and selling when it is too high), thereby pushing the DEX price back toward the CEX's fair value. From this angle, arbitrage activity accelerates the DEX's price-discovery process, narrows the information gap between market tiers, and makes the pricing of the entire

crypto market more unified and efficient. Arbitrageurs assume the price-transmission function between the CEX and DEX market tiers, and their activity ensures that the law of one price can still function, to some degree, in a friction-laden decentralized environment.

This efficiency improvement, however, comes with a corresponding cost. In DEXs based on the AMM or vAMM model, the counterparty to a trade is the passive liquidity provider (LP), and the arbitrageur's adverse-selection loss is passed directly to the LP. A distinction must be drawn, however: in a pure order-book DEX (such as Hyperliquid's core CLOB mode), liquidity is provided by active market makers, and their loss mechanism is essentially different from the impermanent loss of an AMM—active market makers can respond to adverse selection by canceling orders, adjusting quotes, and managing inventory, and what they face is the classic market-maker adverse-selection cost of traditional finance. The following analysis addresses mainly the AMM/vAMM model, but its core logic (the arbitrageur's gain comes at the expense of the liquidity provider's loss) holds in both models. When arbitrageurs profit from OEV, their profit comes directly from the LP's loss. This constitutes an adverse-selection loss more direct than the "impermanent loss" that LPs face, which can be called "arbitrage loss." Every successful oracle-latency arbitrage means the LP has traded with the arbitrageur at an unfavorable "stale price." If such arbitrage occurs frequently, the LP's annualized return will be severely eroded.

This gives rise to a tension at the level of ethics and mechanism design. On the one hand, arbitrage is a necessary mechanism for the market to move toward efficiency; on the other hand, it harms the interests of the LPs who provide the market's basic liquidity. If LPs choose to withdraw liquidity because of continuous arbitrage losses, the DEX's trading depth will fall and slippage will increase, ultimately perhaps causing the DEX's overall efficiency to decline rather than improve, entering a negative feedback loop of falling liquidity. The efficiency improvement brought by CEX-DEX arbitrage is therefore, to a large extent, internalized as a cost borne by the more fragile participants in the DEX ecosystem—the LPs. This raises a core question: can we protect LPs from excessive adverse-selection loss without suppressing arbitrage, the very mechanism that delivers efficiency? The answer to this question has driven the continuous evolution of DEX protocols' anti-arbitrage design.

16.9.4 The evolution of DEX anti-arbitrage design

To address the OEV and LP-loss problems, DEX perpetual-futures protocols have advanced a series of iterative upgrades at the levels of technology and mechanism design. The core goal of this evolution is to keep narrowing the efficiency gap with CEXs while preserving decentralization, thereby compressing arbitrageurs' profit space or redistributing arbitrage profit to the protocol and LPs. This evolutionary path can be roughly divided into the following stages.

First is the iteration of oracle technology. First-generation DEXs often used oracles with low update frequency and high latency, leaving a large window for arbitrage. A new generation of "low-latency" oracles, represented by Pyth Network, has successfully shortened price latency to the level of a few hundred milliseconds by having data

publishers submit price updates directly on-chain while placing the aggregation process off-chain, greatly compressing traditional oracle-arbitrage opportunities [21]. Going further, projects represented by API3 have proposed the concept of OEV Share—that is, auctioning the right to update the oracle so as to capture part of the OEV and return it to the DEX protocol or LPs, achieving a paradigm shift from "fighting arbitrage" to "coexisting with arbitrage and sharing the profit" [22].

Second is anti-MEV design at the protocol level. To cope with various MEV attacks, including OEV, DEX protocols have begun to explore more complex trade-execution mechanisms. For example, an "intent-driven" architecture allows users to express only a trading intent (such as "I want to swap 1 ETH for as much USDC as possible") and leaves the complex task of finding the optimal execution path to a network of professional "solvers." Solvers compete for the right to execute through mechanisms such as batch auctions and can settle multiple trades within a single block, thereby eliminating MEV problems such as front-running. In addition, techniques such as delayed execution and encrypted mempools aim to hide transaction details until they are finally confirmed, so as to protect users and LPs from adverse-selection loss.

Finally, and most fundamentally, evolution is heading toward a radical transformation of DEX perpetual-futures protocols at the level of their underlying architecture. The latest generation of DEXs, represented by dYdX v4 and Hyperliquid, has abandoned the traditional model of building applications on general-purpose smart-contract platforms (such as the Ethereum mainnet or L2 rollups) and instead developed "app-chains" designed specifically for high-performance trading.

As an independent Cosmos app-chain with its own validator network, dYdX v4 can achieve a block time of about 1 second and near-instant transaction finality. Its core design distributes the matching and maintenance of the order book across the memory of its validator nodes, with each validator independently maintaining an in-memory copy of the order book; consensus on trade results is reached through CometBFT consensus, and only the consensus-confirmed fills are written to the blockchain. This "distributed off-chain matching" architecture is essentially different from a CEX's single-server architecture: it retains Byzantine fault tolerance and is superior in security to a centralized scheme, but its performance is still limited by consensus latency—thereby achieving high throughput and low latency [23]. Hyperliquid adopts a more radical "unified-state architecture"; its self-developed HyperBFT consensus mechanism can achieve sub-second transaction-confirmation latency, and it places the trading layer and the smart-contract layer under the same consensus, eliminating cross-layer communication delay so that on-chain state can be accessed instantly, fundamentally shortening oracle latency [24]. The degree of Hyperliquid's decentralization, however, must be assessed with caution. As of the end of 2025, its validator set remained highly concentrated (the number of active validators is far smaller than on Ethereum or the mainstream chains of the Cosmos ecosystem), and its core code was not fully open-source. The JELLY token market-manipulation incident in March 2025 (which caused a peak unrealized loss of about $12 million to $13.5 million in the liquidity vault and forced validators into a controversial manual intervention, voting to force the liquidation of specific positions) exposed the fragility of such app-chains in security governance. Although the validators' manual intervention curbed the loss's expansion in the short

term, it also raised fundamental questions about whether a "decentralized exchange" is worthy of the name. These security trade-offs show that in pursuing CEX-grade performance, app-chain DEXs inevitably compromise on the degree of decentralization and the transparency of governance.

These architectural innovations have brought the new generation of DEXs very close to CEXs in performance, significantly compressing the arbitrage space created by technological latency. Although the CEX-DEX efficiency gap can never disappear entirely because of the physical limits of blockchains (such as the speed of light and the global distribution of nodes), its width is being continuously compressed by technological progress and fierce arbitrage competition. A new equilibrium is forming: at this equilibrium point, the CEX-DEX spread will converge to an extremely narrow level that only covers the execution costs and risk premia of the top arbitrageurs (the Tier 1 and Tier 3 participants with the lowest latency, lowest costs, and strongest MEV-game capabilities). For ordinary participants, significant riskless arbitrage opportunities will no longer exist, marking the maturation of the "institutionalization" and "specialization" of this frontier arbitrage domain.

16.10 Arbitrage and market quality

The preceding six sections (16.4-16.9) deconstructed, one by one at the micro-mechanistic level, the operating logic, sources of return, and efficiency contributions of the six core arbitrage strategies. Each strategy performs an irreplaceable corrective function along a specific dimension of the efficiency spectrum. These micro-level analyses, however, have not yet answered a more macro question: how does the collective effect of arbitrage activity systematically reshape the overall quality of the market? This section lifts the analysis from the single strategy to the aggregate level of the arbitrage ecosystem, examining the systematic impact of arbitrageurs' collective behavior on market microstructure along four dimensions: spread compression, liquidity integration, price-discovery leadership, and rate stability.

The six arbitrage mechanisms do not operate in isolation; they form a dynamic network of mutual synergy and constraint. Along the synergy dimension, basis arbitrage (①) and funding rate arbitrage (③) are complementary: basis arbitrageurs collect the funding rate by continuously holding positions, and their behavior itself constitutes a form of rate arbitrage; while rate arbitrageurs' smoothing of cross-platform rates provides basis arbitrageurs with a more stable cost environment. Cross-exchange arbitrage (②) lays the premise for the execution of all the other strategies: only when the prices of the world's major exchanges remain highly consistent are the basis and rate signals based on a single exchange's quotes reliable. Along the substitution dimension, triangular arbitrage (⑤) and basis arbitrage (①) partly overlap in function: when the CME futures basis and the perpetual-futures funding rate converge, the profit space of triangular arbitrage is compressed, and arbitrageurs turn to pure basis strategies; and vice versa. Along the competition dimension, when Tier 0 participants such as Ethena execute basis arbitrage at scale and systematically depress the funding rate, the rate-arbitrage space of Tiers 2 and 4 narrows accordingly, reflecting the profit competition among different strategies. CEX-DEX cross-tier arbitrage (⑥) is relatively independent

in strategy logic, and its profit space is mainly constrained by the evolution of technological infrastructure. Along the risk-transmission dimension, however, it is tightly interconnected with the other strategies: the FTX collapse showed that a risk event on the CEX side can be rapidly transmitted to DEXs through mechanisms such as fund freezes, oracle-price anomalies, and panicked spread widening, so that independence of strategy logic does not imply isolation of risk. Understanding these inter-strategy interaction dynamics is key to grasping the overall equilibrium state of the arbitrage ecosystem.

Arbitrage activity is often simply regarded as a zero-sum game, but this view overlooks the systemic positive externalities that arbitrage behavior generates. From the standpoint of market microstructure, arbitrageurs' collective behavior constitutes the core force reshaping market quality. As an endogenous market-regulating force, arbitrageurs' collective action, in eliminating spreads, systematically compresses transaction costs, integrates fragmented liquidity, and leads the effective transmission of price discovery. The four impacts examined below reveal its role as a core driver of market quality.

16.10.1 Spread compression

The bid-ask spread is a core indicator of market liquidity and transaction cost. A narrow spread means that buyers and sellers can trade at prices closer to fair market value, lowering the costs of all market participants. Arbitrage activity, especially high-frequency cross-exchange arbitrage, is one of the most direct and important mechanisms for compressing the bid-ask spread.

Its operating logic is rooted in competitive pressure. In a fragmented market, high-frequency arbitrageurs (Tier 1), at microsecond speed, execute "quote-picking" operations across multiple exchanges (that is, buying the sell orders of a low-price exchange and selling the buy orders of a high-price exchange), exerting continuous competitive pressure on market makers. To avoid having their quotes systematically picked off, market makers are forced to tighten their bid-ask spreads to a level close to the globally unified market price (the specific spread-compression trend was presented in detail in Section 16.5.4). This competitive mechanism ultimately benefits all end users.

Empirical research clearly displays the negative correlation between arbitrage activity and the bid-ask spread. We can quantify the intensity of arbitrage activity by constructing an "arbitrage-activity index"—for example, using indicators that more directly reflect arbitrage behavior, such as the frequency of synchronized opposing trades across exchanges and the share of identified delta-neutral positions in total open interest—rather than outcome variables such as the speed of spread convergence, so as to avoid the circular-reasoning risk of mistaking the effect of arbitrage for its cause. As the figure below shows, when the arbitrage-activity index rises, the average bid-ask spread in mainstream perpetual-futures markets shows a clear downward trend. This relationship persists across market states, indicating that arbitrage is a structural force maintaining a low-transaction-cost environment.

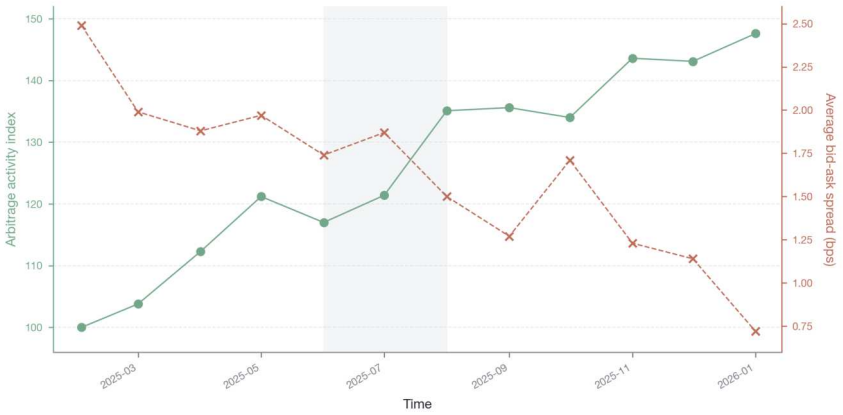


Figure 16-28. The dynamic relationship between arbitrage activity and the bid-ask spread (data source: Binance exchange API data [10]; arbitrage activity is measured by direct proxies such as the frequency of synchronized opposing trades and the share of delta-neutral positions, rather than the speed of spread convergence, to avoid circular reasoning)

Figure 16-28 shows the dynamic relationship between the arbitrage-activity index and the average bid-ask spread in mainstream perpetual-futures markets. Here arbitrage activity is measured comprehensively by direct proxies such as the frequency of synchronized opposing trades across exchanges and the share of identified delta-neutral positions, rather than the speed of spread convergence, to avoid circular reasoning. The two curves show a clear negative correlation: the spread narrows in step when activity rises and widens when activity weakens (as during the liquidity ebb triggered by a major market event), indicating that arbitrage is a structural force maintaining a low-transaction-cost environment.

Ultimately, this arbitrage-driven spread-compression mechanism passes the reduction in transaction costs to every corner of the market. Both the retail trader making small trades and the institution executing large orders benefit from a tighter, more efficient pricing environment. In pursuing their own profit, arbitrageurs in effect supply the entire market with an important public good: lower transaction friction.

16.10.2 Liquidity integration

Besides compressing spreads, arbitrage activity improves market quality along another key dimension: integrating order-book depth. In a highly fragmented environment such as the digital asset market, liquidity is dispersed across hundreds of independent exchanges. For a trader wishing to execute a large order, the order-book depth of a single exchange may be insufficient to complete the trade without significant price impact. The presence of arbitrageurs, however, connects these isolated liquidity pools into a "virtual global liquidity pool."

The core of this concept is that arbitrageurs' cross-market activity allows a change in one exchange's liquidity to be rapidly transmitted to other exchanges (Section 16.8.3 has already argued for a similar liquidity-spillover mechanism from the cross-asset perspective). When a large buy order appears on exchange A, consuming sell-side depth

and pushing the price up, arbitrageurs buy on lower-priced exchanges B and C and sell on exchange A, both bringing exchange A's price back to equilibrium and replenishing the depth that was consumed. In effect, the liquidity of exchanges B and C is transmitted to exchange A, so that the "effective depth" of any single exchange is far greater than the depth visible on its order book.

Through smart order-routing systems, trade aggregators can split a large order across multiple exchanges for simultaneous execution, drawing on the virtual global liquidity network that arbitrageurs have built. An empirical comparison shows, by way of illustrative estimation, that a $10-million-scale order executed on a single exchange might cause a price impact of about 20 basis points, whereas spreading it across multiple mainstream exchanges through an aggregator might reduce the total price impact to about 5 basis points (the specific figures depend on the underlying asset, market state, and order-book depth). This effect stems from arbitrageurs serving as the transmission mechanism of liquidity, ensuring the consistency of global prices and the interconnection of liquidity.

As the figure below shows, in a market without an efficient arbitrage mechanism, the liquidity of each exchange is isolated. But when arbitrageurs actively participate, not only does the depth of each exchange itself increase because market makers participate more actively, but, more importantly, the liquidity of all exchanges is integrated into a virtual global liquidity pool far larger than the sum of its parts. In the dispersed market structure of the digital asset market, which lacks central clearing and a unified trading venue, arbitrageurs assume the core function of liquidity integration and are the key transmission mechanism connecting hundreds of independent liquidity pools.

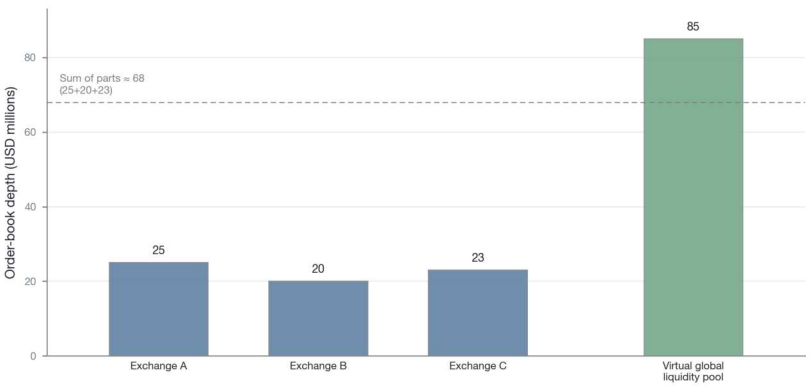


Figure 16-29. The integrating effect of arbitrage on order-book depth (Data source: constructed by the author)

Figure 16-29 uses a single panel to compare the independent order-book depth of individual exchanges with the virtual global liquidity pool after arbitrage integration: the first three bars represent the limited depth of three exchanges when they are independent of one another (the liquidity-island state), and the rightmost bar represents the virtual global liquidity pool that arbitrageurs weave together through cross-market activity, whose total depth is significantly higher than the simple sum of the individual exchanges' depths (as the reference line shows)—indicating that efficient arbitrage not

only thickens each exchange's depth through market makers' more aggressive quoting but also integrates dispersed liquidity into a whole whose total depth far exceeds the sum of its parts.

16.10.3 Price-discovery leadership

As Chapter 15 showed, drawing on multiple empirical studies (including the information-share analysis of price discovery in ether spot and derivatives markets by Alexander et al. (2020) [25]), perpetual futures lead the price-discovery process most of the time (with a contribution of about 70-80%). Arbitrageurs serve as the key medium of bidirectional information transmission in this dynamic.

In a normal market environment, the perpetual-futures market, by virtue of its structural advantages (as analyzed in earlier sections), becomes the preferred venue for informed traders to express their views. When new information about an asset's fundamentals (whether macroeconomic data, project developments, or regulatory news) appears, these traders build positions first in the most liquid perpetual-futures market, causing the perpetual-futures price to move first. At this point, spot-perpetual basis arbitrageurs quickly step in. If the perpetual-futures price rises, they sell perpetual futures while buying an equivalent amount of spot to lock in the basis profit. This process "transmits" the price signal of the perpetual-futures market to the spot market, prompting the spot price to follow upward until the basis returns to a normal level. Along this typical path, arbitrageurs are the agents through which perpetual futures lead spot.

Price-discovery leadership, however, is not fixed. Under the shock of certain special events or structural capital-flow changes, leadership can flip. A typical example is the approval of and large-scale capital inflows into the U.S. spot Bitcoin ETFs in early 2024. In this case, the information source (the large-scale buying demand) acts directly on the spot market (mainly custodial exchanges such as Coinbase). The large ETF-creation demand translates into direct purchases of spot Bitcoin, driving the spot price up first. At this point, basis arbitrageurs perform the reverse operation: observing that the spot price is higher than the perpetual-futures price, they buy perpetual futures while selling spot (or, if holding spot, selling it directly). This behavior transmits the strong demand signal of the spot market to the perpetual-futures market, causing the perpetual-futures price to follow upward passively. In this scenario, the price-discovery share of the spot market rises significantly, briefly approaching that of perpetual futures, while arbitrageurs play the role of transmitting the spot signal in reverse to the derivatives market.

Viewing arbitrageurs as a bidirectional transmission mechanism for price signals is therefore key to understanding the dynamic changes in price-discovery leadership. They do not serve the dominance of perpetual futures in one direction only but mechanically respond to any spread that deviates from equilibrium. Wherever the price signal originates, arbitrageurs ensure it is propagated to every tier of the market in the shortest time. As the figure below shows, although perpetual futures dominate most of the time, the leadership share changes dynamically, reflecting shifts in the main channels of market information inflow across periods. Arbitrageurs constitute the core execution

mechanism in this dynamic transmission system.

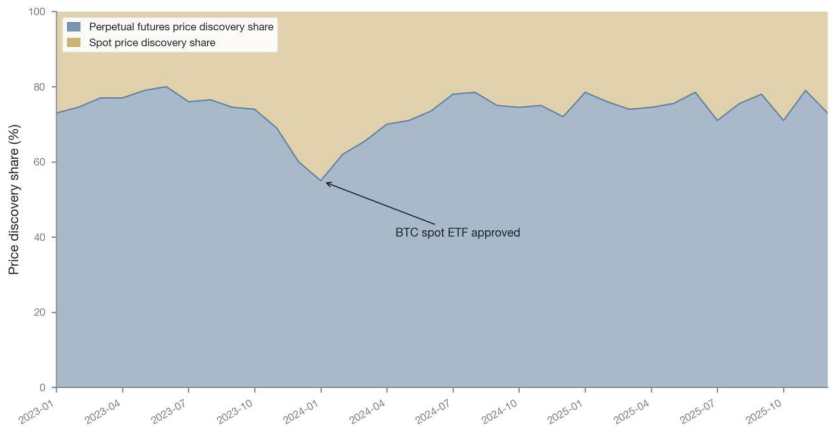


Figure 16-30. The dynamics of price-discovery leadership between perpetual futures and spot (data source: a schematic constructed by the author, synthesized from the information-share analysis of Chapter 15, including Alexander et al. 2020 [25])

Figure 16-30 tracks the evolution of the price-discovery leadership share between the perpetual-futures and spot markets (the vertical axis is the perpetual futures' share of information): during most of the observation period, the perpetual share remains above 70%, confirming its status as the dominant venue for price discovery; but at key junctures such as the approval of the spot Bitcoin ETFs and the large capital inflows into spot in early 2024, the perpetual's leadership share falls markedly and the spot share rises correspondingly, briefly approaching the perpetual's. These dynamic switches confirm the core role of arbitrageurs as a bidirectional transmission mechanism for price signals.

16.10.4 The rate-stabilization function

As the empirical analyses of Sections 16.4.4 and 16.7.4 show, the funding rate mechanism is itself merely a set of rules, and its actual effectiveness depends entirely on the presence of a group of arbitrageurs of sufficient scale, activity, and efficiency to perform price-correction operations. When the funding rate deviates substantially from equilibrium because of a shift in market sentiment, arbitrageurs collect the rate by constructing opposing delta-neutral positions, and their large-scale building of opposing positions itself constitutes corrective pressure on the perpetual-futures price, pushing the funding rate back toward the equilibrium level.

We can understand the decisive role of arbitrageurs as the "mechanism stabilizer" through a thought experiment: suppose a perpetual-futures market with no arbitrageurs. When extreme sentiment arises in the market, the funding rate might rise to an extreme level (or fall to an extremely low negative value) and stay there for a long time, because there is no strong enough counterparty force to offset the one-directional pressure of speculators. In this case, the perpetual-futures price could deviate substantially from the spot price for a long time, and its function as an effective hedging and

price-discovery tool would be severely impaired. Arbitrageurs are the core participants that drive this set of rules to function effectively.

As the figure below shows, there is a nonlinear mapping between arbitrage intensity and the level of the funding rate. When the funding rate is in a relatively moderate range, arbitrage intensity stays at a certain level. But when the funding rate rises sharply because of market imbalance, arbitrage intensity grows exponentially, as the higher potential return draws in more arbitrage capital. This response mechanism ensures that the funding rate is always subject to a strong constraint toward convergence to the equilibrium level, preventing it from deviating without bound. Arbitrageurs are therefore not only the core driving force of market efficiency but a necessary condition and intrinsic component that allows the institutional design of perpetual futures to hold together and function effectively.

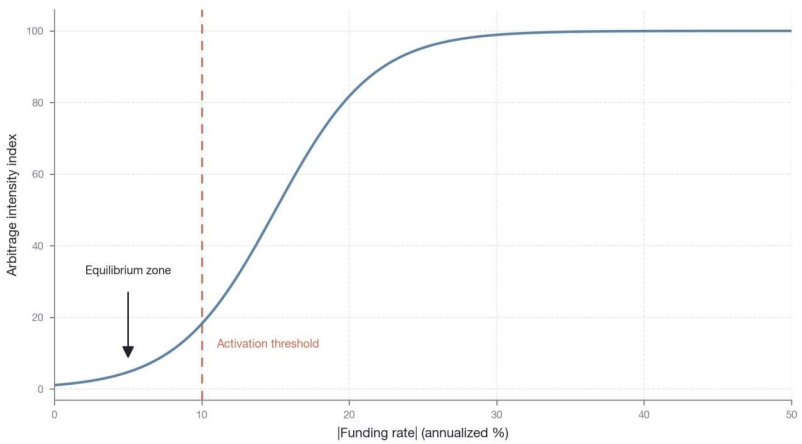


Figure 16-31. The mapping between the funding rate and arbitrage intensity (Data source: an analytical model constructed by the author)

Figure 16-31 reveals the nonlinear mapping between the funding rate (annualized) and arbitrage intensity (measured by the scale of newly added delta-neutral positions): when the rate is in a moderate range (such as 5-15% annualized), arbitrage intensity stays at a baseline level, but once the rate breaks through a certain threshold because of market imbalance, it jumps exponentially, reflecting the strong attraction of high returns for arbitrage capital. This convex response mechanism ensures that the funding rate is always subject to a constraint toward convergence to the equilibrium level, preventing it from deviating without bound in an extreme direction.

16.11 The arbitrage-efficiency transmission model

The analysis of the preceding ten sections has systematically presented a complex arbitrage ecosystem. From the theoretical framework to the stratification of participants, from the six core arbitrage strategies to their reshaping of market microstructure, we have assembled a detailed picture. In particular, Section 16.10 displayed, along four dimensions (spread compression, liquidity integration, price-discovery leadership, and

rate stability), the empirical impact of arbitrage activity on market quality. If these analyses are not integrated into a unified theoretical framework, however, they still lack systematic explanatory power. This section aims to establish a "transmission model" that abstracts the empirical impacts observed in Section 16.10 into three progressive theoretical mechanisms—price calibration, information aggregation, and liquidity integration—thereby explicitly connecting micro-level arbitrage behavior with macro-level market efficiency and answering the central question of the whole chapter: how does arbitrage push the perpetual-futures market toward efficiency? These three mechanisms map clearly onto the four empirical dimensions of Section 16.10: spread compression and rate stability fall under the "price calibration" layer, price-discovery leadership falls under the "information aggregation" layer, and liquidity integration corresponds separately to the "liquidity integration" layer.

The efficiency transmission of arbitrage is not a single, linear process but a multitiered, multidimensional complex network. We can decompose it into three progressive, mutually reinforcing layers.

The first layer is price calibration, the most direct and most observable efficiency-transmission path of arbitrage: arbitrageurs' buying and selling act directly on the price deviation, compressing it to near zero (the specific mechanisms of the three types of calibration—basis, cross-exchange, and funding rate—are detailed in Sections 16.4, 16.5, and 16.7, respectively). Compared with the latter two layers, the transmission mechanism of the price-calibration layer is easier to observe directly and to verify quantitatively. In the efficiency-spectrum framework, the price-calibration layer operates mainly on the "market-tier" and "space" dimensions, ensuring that prices remain consistent across different markets and exchanges.

The second layer is information aggregation. In pursuing profit, arbitrageurs not only compress spreads but also integrate the information dispersed across different markets, different exchanges, and different contract types into a unified, effective price signal: when new information is first priced in one market, arbitrageurs' cross-market trading ensures it is transmitted to other markets within seconds, bringing the same asset's global prices into alignment (the bidirectional transmission path of its price discovery is detailed in Section 16.10.3). Arbitrageurs are the core accelerating mechanism of this information-transmission process. In the efficiency-spectrum framework, the information-aggregation layer operates mainly on the "information type" and "time" dimensions, ensuring that the same information is reflected across different markets at increasingly similar speeds.

The third layer is liquidity integration, the layer of the three transmission mechanisms with the broadest coverage and the most lasting impact on market structure. Arbitrageurs (especially Tier 1 high-frequency traders), while executing arbitrage, add liquidity to the market through continuous two-sided quoting, and, more importantly, integrate the "local liquidity" of individual exchanges into a "virtual global liquidity pool" through cross-market activity, so that large orders can be split across multiple exchanges for execution via smart order routing (the full discussion of this integration mechanism appears in Section 16.10.2). In the efficiency-spectrum framework, the liquidity-integration layer operates mainly on the "asset" dimension, ensuring that the

liquidity of mainstream assets is far higher than that of long-tail assets, thereby forming a clear liquidity gradient.

These three transmission mechanisms do not exist independently but reinforce and promote one another. Price calibration lays the foundation for information aggregation, because only when prices can be effectively calibrated can information be accurately reflected; information aggregation in turn drives liquidity integration, because when participants believe that prices across different markets are consistent, they are more willing to trade across markets, strengthening the formation of virtual global liquidity; and liquidity integration further strengthens price calibration, because deeper liquidity means lower execution costs for arbitrage trades, allowing even smaller spreads to be eliminated by arbitrage.

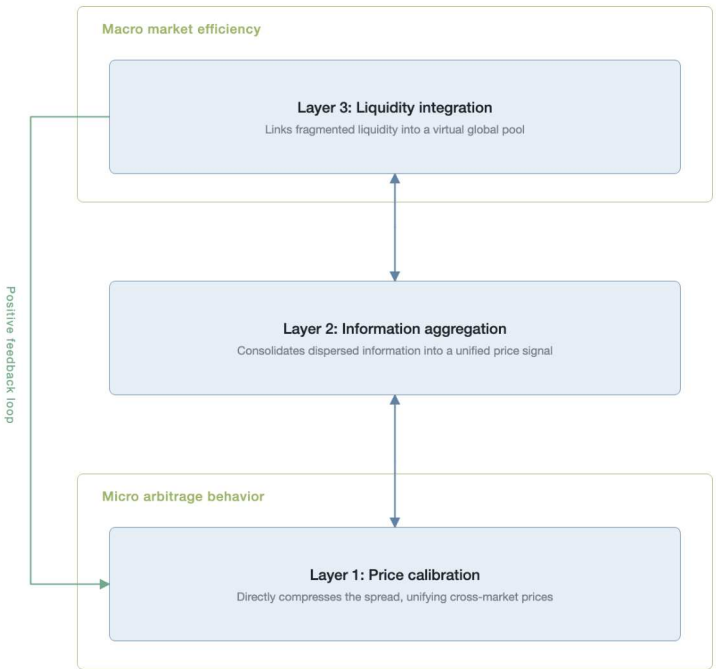


Figure 16-32. The three-layer model of arbitrage efficiency transmission (Data source: a conceptual schematic constructed by the author)

Figure 16-32 uses a tiered structure diagram to display the three-layer model of arbitrage efficiency transmission: the bottom layer is price calibration (most direct and most observable), the middle layer is information aggregation (integrating dispersed information into a unified price signal through cross-market trading), and the top layer is liquidity integration (with the broadest coverage and the most lasting impact on market structure). The three layers form a mutually reinforcing positive-feedback relationship via bidirectional arrows: price calibration lays the foundation for information aggregation, information aggregation drives liquidity integration, and liquidity integration in turn lowers the execution cost of price calibration.

Building on the three-layer transmission mechanism, we can further establish a mapping between arbitrage intensity and market efficiency. The efficiency-spectrum framework established in Chapter 15 provides a multidimensional tool for describing market efficiency, but it is itself static and descriptive. The core innovation of this section is to use arbitrage intensity as a mediating variable, transforming this static efficiency-spectrum framework into a dynamic, causal explanatory model.

We can express this mapping with the following functional relationship:

$$\text{efficiency}(t, a, i, s) \approx g\big[\text{arbitrage intensity}(t, a, i, s)\big]$$

where t denotes the time scale, a denotes the asset, i denotes the information type, and s denotes the market state (here i refers to the information-type dimension and has a different meaning from the i used as an exchange subscript in Sections 16.5 and 16.7). This function indicates that at any specific coordinate of the efficiency spectrum (determined by the four dimensions of time, asset, information type, and market state), the level of market efficiency is determined by the arbitrage intensity at that coordinate. As noted in the discussion of circular reasoning in Section 16.10.1, the dependent variable "efficiency" here must be measured by indicators orthogonal to the spread (such as the Hasbrouck information share, the spread-convergence half-life, and market depth) rather than by the speed of spread convergence, or the mapping would degenerate into a tautology.

The mapping above implicitly assumes a stable macro-financial environment. In fact, the state of global dollar liquidity constitutes an important moderating variable: in a loose liquidity environment, arbitrage capital is abundant and financing costs are low, so the same level of arbitrage intensity can produce a higher efficiency-transmission effect; whereas during a liquidity contraction (such as a Federal Reserve balance-sheet-reduction cycle), the opportunity cost of arbitrage capital rises and its risk tolerance contracts, so that even with unchanged arbitrage strategies and technology, the transmission efficiency declines systematically. The crypto-market liquidity crisis of 2022 illustrated this vividly. A more complete mapping should therefore take the conditional form: efficiency $\approx$ g[arbitrage intensity | macro liquidity state], where macro liquidity moderates transmission efficiency by affecting the cost and risk appetite of arbitrage capital.

This mapping exhibits pronounced heterogeneity across different quadrants of the efficiency spectrum. Consider two extreme scenarios. The first is the quadrant of "high-frequency, mainstream coin, standardized information, normal market." In this quadrant, arbitrage intensity reaches its highest level. High-frequency arbitrageurs are most active here, all six arbitrage strategies are feasible and competition is fierce, and market participants are numerous and well-capitalized. As a result, market efficiency in this quadrant approaches the theoretical ceiling. Spreads are extremely narrow (typically on the order of 1-2 basis points for BTC perpetual futures on mainstream exchanges), price discovery is extremely fast (millisecond scale), and liquidity is extremely deep. The second extreme scenario is the quadrant of "low-frequency, small-cap coin, complex on-chain information, crisis market." Here arbitrage intensity is extremely low. The liquidity of small-cap coins is itself limited, the complexity and verification cost of on-chain information are high, and the risk premium in crisis periods is enormous, so

that even simple arbitrage strategies struggle to profit. As a result, market efficiency in this quadrant is far below that of the former. Spreads are wide (up to the order of 50-100 basis points for illiquid long-tail assets), price discovery is slow (minute scale or longer), and liquidity is thin.

This mapping model has a clear policy implication: if we wish to improve market efficiency, the key lies not in "calling for a more efficient market" or "imposing stricter rules" but in reducing the friction and cost of arbitrage. This spans several levels. At the technical level, it means improving interconnection among exchanges and reducing the latency of cross-market trading. At the institutional level, it means unifying the settlement cycles and rate structures of different exchanges to lower the institutional friction of arbitrage. At the regulatory level, it means securing the legal standing of arbitrageurs, protecting their property and trading rights, and preventing policy uncertainty from raising arbitrage costs. Any improvement in these areas would directly raise arbitrage intensity and, in turn, improve market efficiency.

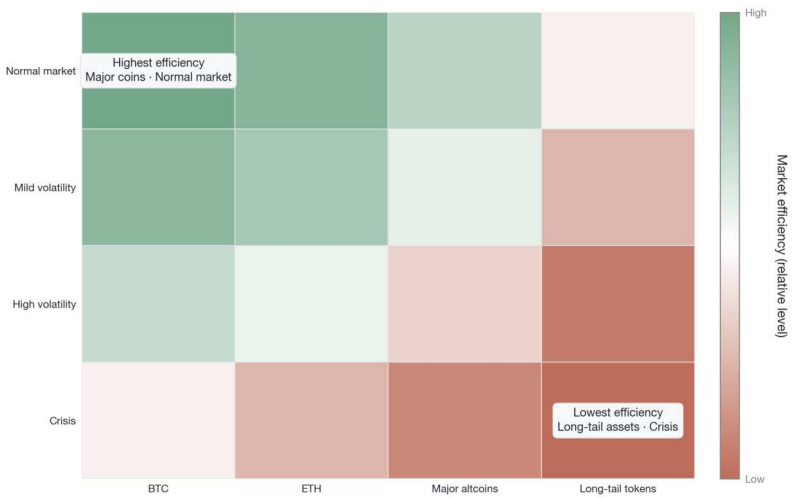


Figure 16-33. The mapping between arbitrage intensity and the efficiency spectrum (data source: a conceptual model constructed by the author; arbitrage intensity is characterized comprehensively by indicators such as the speed of cross-exchange spread convergence, the rate of change of basis-arbitrage-related open interest, and the mean-reversion half-life of the funding rate)

Figure 16-33 simplifies the multidimensional coordinate space of the efficiency spectrum into a two-dimensional projection and marks two extreme quadrants: in the "high-frequency-mainstream-coin-normal-market" quadrant (the upper-left region), arbitrage intensity and market efficiency both reach their highest, and both spreads and information latency are compressed to the limit; whereas in the "low-frequency-long-tail-asset-crisis-market" quadrant (the lower-right region), insufficient arbitrage intensity causes efficiency to fall significantly below that of the former. The gradient distribution between the two quadrants intuitively reflects the decisive influence of the arbitrage cost structure on the level of market efficiency.

This seemingly simple mapping model, however, harbors an intrinsic paradox—one

that was already revealed in the classic paper of Grossman and Stiglitz (1980) but is especially pronounced in the crypto market.

The self-limiting nature of arbitrage stems from a basic economic principle: when a profit opportunity is fully exploited, the opportunity vanishes. The inflow of arbitrage capital compresses the basis and the rate, driving returns back to the equilibrium level that "just covers arbitrage costs" (the dynamic-equilibrium concept itself is presented in Section 16.1.4, and the reflexive-loop example of Ethena's large-scale shorting systematically depressing the funding rate appears in Section 16.4.3). This process can be observed in all six arbitrage types: the fall in returns prompts some capital to exit, the spread and the rate then widen again, and a new round of capital is drawn in, so that the market forms a persistent "efficiency oscillation" around some equilibrium level.

This self-limiting mechanism aligns closely with the adaptive markets hypothesis proposed by Lo (2004) [26]. The adaptive markets hypothesis holds that market efficiency is not an end state but a perpetually ongoing dynamic process. Market participants continuously adapt to the environment, discovering and exploiting new arbitrage opportunities, but this process itself eliminates those opportunities, pushing the market toward a new equilibrium. This cycle has no end point, and market efficiency has no fixed "ceiling" but fluctuates continuously under different conditions.

From this perspective, the "arbitrage failures and market anomalies" that Chapter 18 will discuss no longer seem mysterious. Those seemingly long-lasting market anomalies in fact often reflect a rise in arbitrage costs or a temporary shortage of arbitrage capital. When arbitrage costs rise for some reason (such as leverage constraints, higher risk premia, or technical failures), the spreads that arbitrage had eliminated reappear. In most cases, this is not a market "failure" but a new equilibrium under a new cost structure. Two starkly different situations must be distinguished, however: in a normal cost adjustment, the spread does re-stabilize at a new level; but in an extreme liquidity crisis, the widening of the spread may trigger self-reinforcing positive feedback: arbitrageurs are hit by margin calls or forced liquidation, and their forced unwinding further drives the spread wider, causing more arbitrageurs to be liquidated. In such a margin spiral, the system does not converge toward a "new equilibrium" but may keep diverging until external intervention or bankruptcy clearing. The equilibrium assumption of arbitrage's self-limiting nature holds only when the system has not entered the liquidity-spiral region, and because the crypto market lacks a lender-of-last-resort mechanism and a unified clearing system, its threshold for entering this dangerous region may be lower than in traditional financial markets.

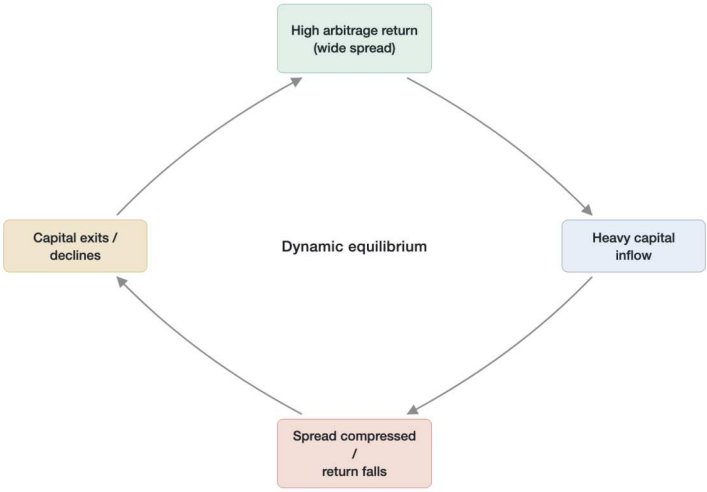


Figure 16-34. The self-limiting nature of arbitrage and adaptive equilibrium (data source: a conceptual schematic constructed by the author, with parameters referencing the adaptive markets hypothesis framework of Lo 2004 [26])

Figure 16-34 uses a cycle diagram to depict the dynamic equilibrium between arbitrage intensity and arbitrage returns: rising arbitrage returns attract new capital and drive arbitrage intensity up, higher arbitrage intensity compresses the spread and the rate and causes returns to fall, and falling returns in turn prompt some capital to exit, so the spread widens again and returns rise once more. The system ultimately oscillates around a dynamic-equilibrium point, and the spread level corresponding to that equilibrium point is the efficiency ceiling the market can maintain, as well as the minimum compensation needed to incentivize arbitrageurs to keep providing infrastructure services.

In sum, perpetual-futures arbitrage, compared with arbitrage in traditional markets, displays irreplaceable efficiency contributions along three dimensions. First, the "flow anchoring" paradigm gives the arbitrage force continuity in the time dimension, so that price-correction pressure is no longer concentrated around the maturity date of traditional futures but is distributed evenly across the entire trading cycle through the funding rate mechanism. Second, as a real-time quantitative indicator of market sentiment, the funding rate gives arbitrage behavior a sentiment-calibration function: arbitrageurs' opposing position-building against extreme rates, in correcting the price deviation, also sends the market a clear signal about excessive leverage and risk accumulation—something the traditional futures market lacks. Third, perpetual-futures arbitrageurs simultaneously cross multiple dimensions—spot and derivatives, CEX and DEX, different exchanges, and different assets—serving as a cross-dimensional intermediary for information and liquidity transmission in a highly fragmented market structure. These three contributions make perpetual-futures arbitrage an irreplaceable endogenous mechanism in the efficiency system of the crypto market.

16.12 Chapter summary

Focusing on the arbitrage mechanisms of the perpetual-futures market, this chapter has established a complete analytical framework from micro behavior to macro efficiency. At the theoretical level, we proposed the "arbitrage infrastructure hypothesis," redefining the arbitrageur's role from the "captor of riskless profit" of traditional theory to the "active provider of market infrastructure." In a crypto market that lacks a central clearinghouse and unified regulation, arbitrageurs in fact assume three key infrastructure functions: the cross-market price-transmission bus (partly substituting for a central clearing institution along the price-linkage dimension, but not assuming its concentrated management of counterparty credit risk or its default-fund functions), the anchor of the perpetual-futures price, and the integrator of global liquidity. At the same time, the "flow anchoring" paradigm of perpetual futures, compared with the "point convergence" of traditional futures, fundamentally reshapes the logic of arbitrage, transforming it from a one-off trade into an ongoing position-management activity.

At the empirical level, this chapter systematically deconstructed six core arbitrage mechanisms—spot-perpetual basis arbitrage, cross-exchange arbitrage, spot-futures-perpetual triangular arbitrage, funding rate arbitrage, cross-asset statistical arbitrage, and CEX-DEX cross-tier arbitrage—and revealed how they jointly advance market efficiency through the three-layer transmission path of price calibration, information aggregation, and liquidity integration. The multi-tiered participant structure of the arbitrage ecosystem (from protocol-level arbitrageurs to high-frequency traders, from systematic quantitative funds to on-chain MEV searchers) forms a self-regulating dynamic-equilibrium system, whose intrinsic self-limiting mechanism aligns closely with Lo's adaptive markets hypothesis.

Looking ahead, the evolution of the arbitrage ecosystem will unfold along two main lines. The first is the further protocolization of infrastructure. As on-chain derivatives protocols mature and cross-chain interoperability improves, the efficiency gap between CEXs and DEXs will be compressed further, and the "arbitrage-as-a-service" model represented by Ethena may spur more innovations that package arbitrage strategies into composable financial products. The second is the shaping effect of the regulatory framework. Different jurisdictions' regulatory policies for crypto derivatives will directly affect the friction-cost structure of arbitrage: reducing cross-market trading latency, unifying settlement cycles, and safeguarding the legal standing of arbitrageurs would all directly raise arbitrage intensity and, in turn, improve market efficiency.

The analysis in this chapter, however, also reveals that arbitrage infrastructure is not omnipotent. The self-limiting nature of arbitrage means that market efficiency will never reach the theoretical ceiling; and under extreme market stress, the arbitrage force may temporarily fail because of margin constraints, forced liquidation, or counterparty risk. These "arbitrage failure" scenarios and the market anomalies they trigger will be the central topics of Chapters 17 and 18.

References

[1] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets.

Journal of Financial Economics, *135*(2), 293–319. https://doi.org/10.1016/j.jfineco.2019.07.001

[2] Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35–55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

[3] Mitchell, M., Pulvino, T., & Stafford, E. (2002). Limited arbitrage in equity markets. *The Journal of Finance*, *57*(2), 551–584. https://doi.org/10.1111/1540-6261.00434

[4] He, S., Manela, A., Ross, O., & von Wachter, V. (2022). *Fundamentals of perpetual futures*. SSRN. https://doi.org/10.2139/ssrn.4301150

[5] Gromb, D., & Vayanos, D. (2010). Limits of arbitrage: The state of the theory. *Annual Review of Financial Economics*, *2*(1), 251–275. https://doi.org/10.1146/annurev-financial-073009-104107

[6] Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, *70*(3), 393–408. https://www.jstor.org/stable/1805228

[7] DefiLlama. (n.d.). *DeFi TVL dashboard*. Retrieved March 10, 2026, from https://defillama.com/

[8] The Block. (2024, February 23). *Ethena captures 5% of ether perpetual futures open interest*. The Block. https://www.theblock.co/post/278802

[9] Token Terminal. (n.d.). *Token Terminal: Fundamentals for crypto*. Retrieved March 10, 2026, from https://tokenterminal.com/

[10] Binance. (n.d.). *Binance Futures market data API*. Retrieved March 10, 2026, from https://www.binance.com/en/futures

[11] CME Group. (n.d.). *Bitcoin futures and options on futures*. Retrieved March 10, 2026, from https://www.cmegroup.com/markets/cryptocurrencies/bitcoin/bitcoin.html

[12] Zhivkov, P. (2026). The two-tiered structure of cryptocurrency funding rate markets. *Mathematics*, *14*(2), 346. https://doi.org/10.3390/math14020346

[13] Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, *19*(3), 797–827. https://doi.org/10.1093/rfs/hhj020

[14] DeLong, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, *98*(4), 703–738. https://doi.org/10.1086/261703

[15] Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, *59*(6), 1551–1580. https://doi.org/10.2307/2938278

[16] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357–384. https://doi.org/10.2307/1912559

[17] Bieganowski, B., & Ślepaczuk, R. (2026). *Explainable patterns in cryptocurrency microstructure*. arXiv. https://arxiv.org/abs/2602.00776 https://doi.org/10.2139/ss

rn.6159346

[18] Pandya, P., Moser, M., & Sofia, G. (2024, August 30). *An introduction to oracle extractable value (OEV)*. Chorus One; Superscrypt. https://chorus.one/reports-research/an-introduction-to-oracle-extractable-value-oev

[19] Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., & Juels, A. (2020). Flash boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. In *2020 IEEE Symposium on Security and Privacy (SP)* (pp. 910–927). IEEE. https://doi.org/10.1109/SP40000.2020.00040

[20] Weintraub, B., Ferreira Torres, C., Ritter, C., & Breidenbach, L. (2022). A flash(bot) in the pan: Measuring maximal extractable value in private transaction ordering. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (pp. 3765–3776). ACM.

[21] Pyth Network. (n.d.). *Pyth Developer Hub*. Retrieved March 5, 2026, from https://docs.pyth.network/

[22] Greene, J., Shahid, A., Benligiray, B., & Vanttinen, H. (2022). *Oracle extractable value (OEV) through order flow auctions* (v1.0.1). API3 DAO. https://github.com/api3dao/oev-litepaper

[23] dYdX Foundation. (n.d.). *dYdX documentation*. Retrieved March 5, 2026, from https://docs.dydx.xyz/

[24] Liu, J.-H. (2025, December 13). *Technical architecture comparison: Hyperliquid, dYdX, and Lighter.xyz*. Medium. https://medium.com/@gwrx2005/technical-architecture-comparison-hyperliquid-dydx-and-lighter-xyz-2fd005854a7e

[25] Alexander, C., Choi, J., Massie, H. R., & Sohn, S. (2020). Price discovery and microstructure in ether spot and derivative markets. *International Review of Financial Analysis*, *71*, 101506. https://doi.org/10.1016/j.irfa.2020.101506

[26] Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, *30*(5), 15–29. https://doi.org/10.3905/jpm.2004.442611

The Risks of Arbitrage — Why Risk-Free Arbitrage Is Not Risk-Free

In May 2022, the crypto market underwent a profound systemic crisis that offers instructive empirical material for understanding the true nature of the risks in perpetual futures arbitrage. In the preceding months, the funding rate on Bitcoin perpetual futures had remained persistently positive, and many basis arbitrageurs had built delta-neutral positions on that footing—buying spot and shorting an equivalent notional amount of perpetual futures—to capture the basis and collect funding payments. Throughout this chapter, the basis is defined as the perpetual futures price minus the spot price: a positive basis denotes a perpetual premium, and a negative basis denotes a perpetual discount. In a low-volatility environment, the strategy performed steadily and was widely regarded as a low-risk source of yield. The representative scenario below—which combines the characteristic features of several historical crises, including May 2022—illustrates the strategy's fragility under extreme market conditions (the specific figures are illustrative; a review based on public data appears in Section 17.9.1).

Within a few days, however, market conditions reversed sharply. Over a 72-hour window, the funding rate plunged from roughly +0.03% per 8 hours to −0.10% per 8 hours, and arbitrageurs' positive cash flow abruptly turned into a persistent drain on capital. At the same time, the steep fall in the Bitcoin price triggered margin calls, while the unrealized gains on the short perpetual-futures leg could not, under the settlement mechanism, effectively serve as margin—a structural contradiction. When arbitrageurs tried to close out under the combined pressure of funding rate reversal and margin calls, market liquidity had already dried up severely, and the basis had widened from −50 basis points before the crisis to −500 basis points. A strategy that was, in theory, fully hedged produced a realized loss of 15% to 30% within a single week. Over the same period, the systemic collapse of the Terra/UST algorithmic stablecoin system reinforced the lesson: every "risk-free" strategy that depends on a particular institutional design will expose its intrinsic fragility under extreme market stress.

These events reveal a fundamental fact that runs through the entire crypto-finance system: under the current institutional and technical structure, no genuinely "risk-free arbitrage" exists. This is the symmetric counterpart to the Arbitrage Infrastructure Hypothesis (AIH) advanced in Chapter 16 (see Section 16.1.3). Chapter 16 positioned

arbitrageurs as the central custodians of market efficiency; this chapter focuses on the structural risks they face—for when arbitrageurs collectively retreat in the face of intolerable risk, the transmission mechanism of market efficiency breaks down systemically.

To answer systematically the question of where the limits of arbitrage lie, this chapter develops an analytical framework it calls the *arbitrage constraint topology*.[1] Starting from the sources of risk, the framework decomposes perpetual futures arbitrage into seven core constraint dimensions: margin and forced liquidation; funding rate reversal; counterparty and platform risk; execution and timing risk; liquidity depletion; smart contract and oracle risk; and regulatory and compliance risk. The chapter analyzes the internal mechanism of each dimension in turn and shows how, under extreme market conditions, these dimensions undergo *risk resonance* that drives a nonlinear collapse in arbitrage capacity. That arbitrage returns remain persistently positive can be explained, to a large degree, as the joint product of constraints (which limit both the scale of and the appetite for arbitrage) and risk compensation (the pricing of tail and resonance risk). Dimensions such as margin and funding rate reversal supply an earnable risk premium, whereas dimensions such as regulation and market access sustain spreads primarily by shrinking the pool of eligible capital. Because the two operate through different mechanisms, they should not be lumped together as a single "premium."

17.1 The theoretical framework of arbitrage risk

In the idealized picture drawn by financial theory, arbitrage is the central mechanism through which a market restores equilibrium. In that framework, arbitrageurs step in the instant a price deviation appears, eliminate the mispricing through "risk-free" trades, and earn a profit in the process. Chapter 16 described precisely how this mechanism weaves a complex web of efficiency in cryptocurrency markets, and in the perpetual futures ecosystem in particular. A central theoretical question nonetheless remains: if arbitrage is truly "risk-free," why do arbitrage opportunities persist—and in some periods even grow unusually rich? Why is the market's efficiency gap never fully closed?

The answer lies not in any deficiency of arbitrageurs' skill or in market irrationality, but in a more fundamental structural constraint: in the real world, no absolutely "risk-free" arbitrage exists. Every seemingly flawless arbitrage is conducted not in a theoretical vacuum but embedded within a complex constraint structure composed of capital, time, technology, and rules. This section reframes the analysis—from the textbook illusion of "risk-free" arbitrage to a more explanatory theory of arbitrage risk built around the notion of *constraint*. The analysis shows that arbitrageurs' capacity to correct mispricings is systematically limited by structural constraints along multiple dimensions. Understanding the sources, mechanisms, and interactions of these constraints is the key to understanding the upper bound on efficiency in perpetual futures markets.

[1]The term *topology* is borrowed here to emphasize the connective relationships and structural invariance among the risk dimensions, rather than in the strict sense of mathematical topology. The framework is concerned with how the relational structure among risk factors remains stable, or shifts abruptly, across different market states.

17.1.1 The theory of the limits of arbitrage

The three core assumptions of "pure arbitrage" in classical finance—unlimited capital, instantaneous and costless execution, and certain convergence—are all systematically violated in perpetual futures markets. Capital is constrained by margin requirements, execution is subject to latency and slippage, and the absence of an expiry date in perpetual futures breaks, at the most fundamental level, the certainty of price convergence.

It is precisely this pronounced gap between theory and reality that gave rise to an important branch of financial economics: the theory of the *limits of arbitrage*. The seminal contribution is the paper by Shleifer and Vishny (1997) [1]. They observed that real-world arbitrage—especially arbitrage carried out by professional fund managers deploying other people's capital—faces two core risks. The first is *fundamental risk*: even if an asset's price ultimately reverts to its fundamental value, new adverse information may arrive during the reversion and widen the deviation further. The second is *noise-trader risk*: markets contain large numbers of irrational noise traders whose collective sentiment or behavior can keep pushing prices away from fundamental value in the short run, making the mispricing more severe. The core insight of Shleifer and Vishny is that the greatest threat to an arbitrageur is not an error in judging an asset's ultimate value, but the risk of being liquidated by a temporary, large mark-to-market loss before prices revert. When arbitrageurs must report performance to their capital providers, the pressure of such short-term losses is amplified and can force them to close positions at the least opportune moment, turning a theoretically profitable strategy into a realized loss.

Building on Shleifer and Vishny, the theory of the limits of arbitrage has been extended in several important directions. DeLong, Shleifer, Summers, and Waldmann (1990) [2] were the first to build a formal model of noise-trader risk, proving that even when arbitrageurs' valuation judgments are entirely correct, the persistent presence of noise traders can sustain price deviations over the long run. Xiong (2001) [3] further analyzed the fragility of convergence traders during funding-liquidity crises, showing how leveraged arbitrageurs can be forced to close positions at the worst possible moment under the pressure of margin calls. Liu and Longstaff (2004) [4] rigorously proved that even when a spread is certain to converge eventually, the optimal arbitrage strategy must still limit position size substantially, because uncertainty along the intermediate path can lead the arbitrageur to be liquidated before convergence occurs. These contributions provide a solid academic foundation for analyzing the constraints on perpetual futures arbitrage: the absence of an expiry date amplifies the dilemma in the Liu–Longstaff model—from "convergence known, path uncertain" to the far more severe challenge of "convergence itself uncertain."

This framework offers a key resolution of the Grossman–Stiglitz paradox. The paradox holds that if markets are fully informationally efficient, no one has an incentive to gather information, because information cannot be turned into profit; yet if no one gathers information, how can the market become efficient at all? The theory of the limits of arbitrage identifies precisely the frictions that prevent arbitrageurs from fully eliminating market inefficiency. Because arbitrage faces capital constraints, execution risk, and uncertainty about price convergence, arbitrageurs cannot—and are unwill-

ing to—arbitrage every mispricing without limit. Arbitrage returns can therefore be viewed as compensation for bearing these "limits of arbitrage" risks. The arbitrage opportunities that persist, never fully eliminated, in turn motivate professional investors to discover and study them, thereby sustaining the market in a dynamic equilibrium of "partial efficiency" [5].

17.1.2 The distinctive risk profile of perpetual futures

Chapter 16 argued for the institutional advantages of perpetual futures, but these same institutional features are also the source of risk. Every institutional advantage of perpetual futures corresponds to a potential risk exposure.

The expiry date of a traditional deliverable futures contract gives arbitrageurs a guarantee of certain price convergence and sets a clear temporal boundary on their risk exposure.

Because perpetual futures have no expiry date, the basis may, in theory, never converge. The arbitrageur's core risk shifts from "misjudging direction" to "running out of capital before the spread converges." Risk exposure changes from a finite horizon to a potentially unlimited one, which fundamentally reshapes the paradigm of risk management.

Accompanying this is uncertainty in the source of returns. As a random variable, the funding rate can flip negative rapidly when market sentiment reverses, turning the strategy's source of return into a persistent cost. This risk—the reversal of the return source—is the core structural risk facing perpetual futures arbitrageurs.

This analysis does not imply that the perpetual futures price will deviate from spot without limit. The funding rate mechanism itself constitutes a *soft convergence force*: when perpetual futures trade at a sustained premium, the funding that longs pay to shorts makes holding the premium position economically more expensive over time, which endogenously restrains the deviation from widening indefinitely. Under a balanced structure of market participants, moderate leverage, and active market makers, this anchoring mechanism generally works effectively. Yet the soft convergence force operates within clear boundary conditions: when market sentiment becomes highly synchronized—for example, in a panic-driven, one-sided rush to short—and the long-short balance becomes severely skewed, the funding rate's penalty may be insufficient to offset the expected profit from directional speculation, and this anchoring mechanism (the funding rate's *streaming anchor*, as described in Chapter 16) can fail systemically. It is precisely within such failure windows that the basis shifts from mild mean reversion to sustained, one-directional divergence—the core risk examined in this section.

Table 17-1 contrasts the essential differences between deliverable futures and perpetual futures across the dimensions of arbitrage risk.

Risk dimension	Deliverable futures arbitrage	Perpetual futures arbitrage
Convergence certainty	Forced convergence at expiry	No certain point of convergence
Holding period	Finite (until expiry)	Theoretically unlimited
Stability of return source	Holding to expiry locks in a certain convergence return	Funding rate fluctuates randomly, and its direction can change
Margin risk	Finite-horizon exposure	Open-ended exposure
Risk of return-source reversal	Low (the basis will eventually converge)	High (the funding rate can reverse for extended periods and by large magnitudes)

Table 17-1. Comparison of arbitrage-risk dimensions between deliverable futures and perpetual futures (Data source: compiled by the author)

Table 17-1 reveals the defining characteristic of perpetual futures arbitrage risk: the risk borne by "time" in traditional arbitrage is transferred to "capital," and the higher arbitrage return is precisely the necessary compensation for this structural risk.

17.1.3 The arbitrage constraint topology

This chapter proposes the *arbitrage constraint topology*: a list of seven principal risk categories organized around the institutional features and market structure of perpetual futures, whose overall structure is shown in Figure 17-1. These seven constraints form an open list that is additive and overlapping rather than a strictly mutually exclusive and collectively exhaustive (MECE) taxonomy. The dimensions are not partitioned by a single uniform criterion (some by the object of risk, others by cause); a particular risk—such as the collateral basis risk discussed in Section 17.10—may, depending on context, be placed within an existing dimension or treated as a supplementary one; and some risks, such as maximal extractable value (MEV) and API-key compromise, may straddle several dimensions. The framework aims to integrate seemingly isolated risk factors into an interconnected logical network, thereby providing a clear map for the identification, quantification, and management of risk.

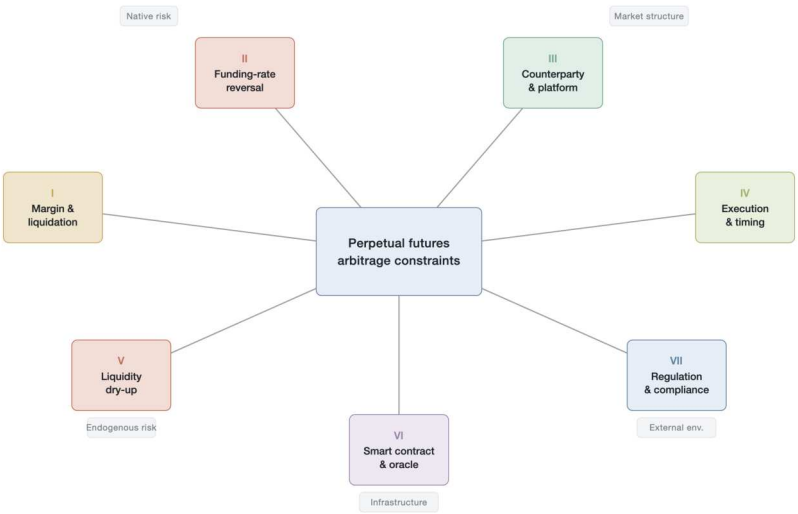


Figure 17-1. Overview of the arbitrage constraint topology (Data source: structural schematic drawn by the author; the seven constraints form an open, additive, and overlapping list rather than a mutually exclusive and collectively exhaustive set)

Figure 17-1 presents the seven constraint dimensions in a star-shaped radial structure, together with their assignment to five categories—native, market-structure, endogenous, infrastructure, and external-environment (the specific classification of each dimension appears in Table 17-2 and below). The seven spokes are joined through a central hub, revealing that the constraints are interdependent rather than isolated from one another.

These seven dimensions make up the list of risks that limit the force of arbitrage (their classification and institutional roots are summarized in Table 17-2, and the transmission relationships among them appear in Figure 17-26). They are as follows.

Dimension	Constraint	Core mechanism	Institutional root	Detailed discussion
I	Margin and forced liquidation	Leverage magnifies exposure, leading to insufficient margin and triggering forced liquidation, which terminates arbitrage positions at the worst possible moment.	High-leverage design; the mark price mechanism	Section 17.2
II	Funding rate reversal	The funding rate, a source of return, turns negative, so the cost of carrying a position shifts from a positive return to a persistent cash outflow.	The intrinsic randomness of the funding rate anchoring mechanism	Section 17.3
III	Counterparty and platform	An exchange collapse, an API outage, or a freeze on funds makes one "leg" of an arbitrage position vanish instantly, destroying the hedge structure.	Market fragmentation; the absence of a central counterparty	Section 17.4
IV	Execution and timing	Trade legs across multiple platforms and contracts cannot be perfectly synchronized, producing slippage, partial fills, and price deviation.	Physical latency; network congestion; differences in exchange performance	Section 17.5
V	Liquidity depletion (in its extreme form, the "liquidity black hole")	Under extreme market volatility, liquidity evaporates rapidly, so arbitrageurs cannot exit at reasonable prices and may even be caught in a stampede.	The endogeneity of arbitrageurs' own behavior; collective action	Section 17.6
VI	Smart contract and oracle	For on-chain perpetual futures, code vulnerabilities, oracle manipulation, or cross-chain bridge attacks can cause asset losses or render a strategy ineffective.	The rigidity of "code is law"; dependence on external data	Section 17.7
VII	Regulation and compliance	Sudden changes in laws and regulations can block certain trading routes, spike compliance costs, or even force a platform to halt service.	The uncertainty and fragmentation of the global regulatory environment	Section 17.8

Table 17-2. The arbitrage-constraint taxonomy and its core mechanisms (Data source: compiled by the author)

The taxonomy is designed to follow a layered logic from the inside out. The margin-

and-forced-liquidation and funding-rate-reversal dimensions are the *native risks* of the perpetual futures institution: they arise directly from its core mechanism design and are the fundamental risks that distinguish it from other derivatives. The counterparty-and-platform and execution-and-timing dimensions are *market-structure risks*, determined by the current reality of crypto markets—many coexisting exchanges and the absence of a unified clearing infrastructure. The liquidity-depletion dimension is an *endogenous risk*: it is not imposed from outside but generated by the collective behavior of the arbitrageur population, reflecting positive-feedback effects within the system. The smart-contract-and-oracle dimension is a *technical-infrastructure risk*, applying mainly to decentralized derivatives protocols. Finally, the regulation-and-compliance dimension is an *external-environment risk*, representing the macro-level uncertainty emanating from the real-world legal system.

The key insight of this topological framework is that it reveals the common origin of the high returns and high risks of perpetual futures arbitrage. The two most central risks—the margin-and-forced-liquidation and funding-rate-reversal dimensions—are almost absent, or present only very mildly, in traditional futures arbitrage. It is precisely to compensate for these additional and profound risks that the market must offer perpetual futures arbitrageurs a higher expected return. This is consistent with the AIH of Chapter 16: arbitrageurs' excess returns are not costless gains but a "service fee" collected, in their role as providers of market infrastructure, for bearing systemic risk.

17.1.4 The temporal structure of risk

Having constructed the seven-dimensional spatial topology of risk, we must also introduce the dimension of time to observe how this structure evolves dynamically. Arbitrage constraints are not static: their intensity and their correlations change dramatically as the market regime shifts. As shown in Figure 17-2, the temporal structure of risk can be divided into three principal states: normal, stressed, and resonant.

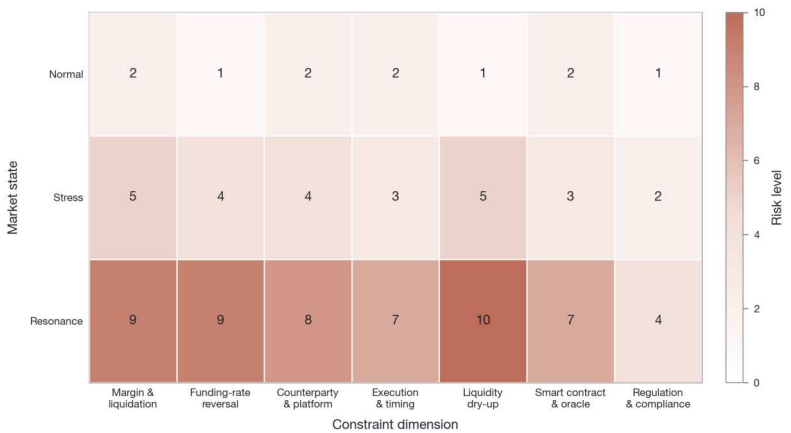


Figure 17-2. A classification framework for arbitrage risk (Data source: drawn by the author; the values shown are the author's risk ratings on an ordinal 1–10 scale, not empirical data)

In a *normal* market, the seven constraints are independent of one another and low in intensity. The funding rate fluctuates steadily around a small positive value, market liquidity is ample, price movements are mild, and platforms operate stably. In this environment, risk is identifiable, measurable, and largely manageable. Arbitrageurs can hedge most of it through careful calculation and diversification; arbitrage proceeds smoothly, and market efficiency is maintained at a high level. Arbitrage here more closely resembles a systematic, quantitative operation.

When the market suffers a shock—say, a sudden and large price pullback—the system enters the *stressed* state. Several constraint dimensions now tighten markedly. Violent price swings cause the pressure in the margin-and-forced-liquidation dimension to spike, and many positions face margin calls or the threat of forced liquidation. Market panic may drive the funding-rate-reversal dimension, so that positive cash-flow returns narrow or even turn negative. At the same time, a surge in trading volume can trigger risk in the execution-and-timing dimension, and slippage and latency become severe. In the stressed state, risks are no longer isolated: they begin to interconnect, though a severe chain reaction has not yet formed. Arbitrageurs' profit margins are compressed, they are forced to cut position sizes to survive, and overall market efficiency begins to decline noticeably.

The highest-risk state—and the core of this chapter's theoretical contribution—is the *resonant* state. When the market suffers an extreme shock, such as a black-swan event, the system may jump from the stressed state to the resonant state (as shown in Figure 17-3). Now the constraints in several dimensions, or even all seven, are activated simultaneously; they no longer add up in a simple linear fashion but form destructive positive-feedback loops that amplify one another, causing arbitrage capacity to contract nonlinearly and abruptly (the dimension-by-dimension transmission chain and the empirical review appear in Section 17.9). Arbitrageurs, who normally serve as market stabilizers, now face pressure on their own survival and may even become a force that intensifies the market collapse. It is precisely during risk resonance that the transmission mechanism of market efficiency is severely obstructed and a cluster of extreme price anomalies (examined in detail in Chapter 18) emerges.

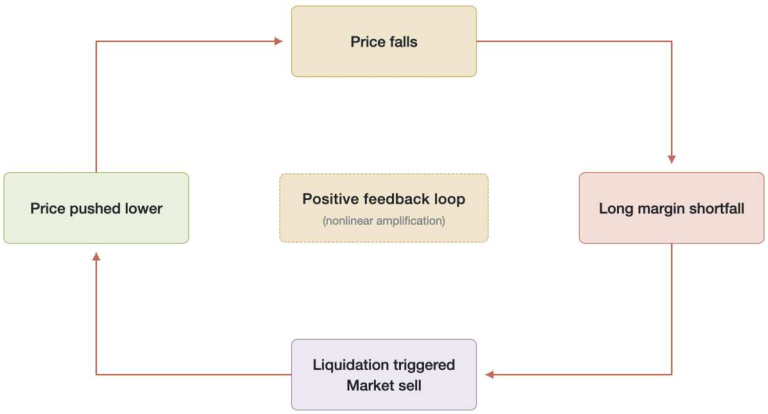


Figure 17-3. Schematic of the price–margin–liquidation positive-feedback loop in the risk-resonance state (Data source: conceptual mechanism schematic drawn by the author, not empirical data)

Understanding this temporal structure of risk—from independence in the normal state, to tightening in the stressed state, to nonlinear amplification in the resonant state—is the key to understanding why crypto markets appear highly efficient most of the time yet reveal extreme fragility at critical moments. It marks the boundary of the arbitrage mechanism's power to maintain efficiency, and it provides the dynamic, macro-level backdrop for each specific constraint dimension explored below.

To make this three-state framework operational, one must identify observable indicators of regime transition. Candidate indicators include the z-score of the funding rate (its deviation from the historical mean, in standard deviations), the ratio of market-wide liquidation volume to open interest, the intraday coefficient of variation of order-book depth, and the dispersion of the basis across exchanges. When several of these indicators simultaneously cross their respective historical-percentile thresholds, this can be read as an early-warning signal that the system is transitioning from normal to stressed, or from stressed to resonant. This approach is consistent in methodological spirit with the Markov regime-switching model of Hamilton (1989) [6], although a formal estimation of a three-state model for crypto markets still awaits future empirical research.

17.2 Margin and forced liquidation

The margin regime is a core component of the perpetual futures design. Through leverage, it magnifies capital efficiency and lets traders participate in the market at a scale far exceeding their principal. Yet the margin regime is also the most direct constraint on arbitrage. It turns a theoretically delta-neutral strategy into a practical problem bounded by capital management and the ability to forecast risk. The health of the margin account directly determines whether an arbitrage position can survive, and the forced-liquidation mechanism sets a hard boundary condition on the failure of an arbitrage strategy. This section examines four core aspects of the margin-and-forced-liquidation dimension, showing how it fundamentally limits the scale of arbitrage and,

under specific conditions, gives rise to systemic market risk.

17.2.1 How margin constrains the scale of arbitrage

In perpetual futures markets, the margin regime directly determines the maximum position size an arbitrageur can deploy, constituting the first hard ceiling on the scale of arbitrage. Analyzing spread data across major exchanges, Makarov and Schoar (2020) [7] found that the capture rate of cross-exchange arbitrage opportunities is far below theoretical expectations, and that many viable spreads go unexecuted because capital cannot be redeployed across platforms in time. This finding demonstrates the systematic constraining effect of the margin regime on the force of arbitrage.

A rational arbitrageur must keep effective leverage far below the maximum the exchange permits, preserving an ample safety buffer against potential mark-to-market losses (see Figure 17-4). Although mainstream perpetual futures platforms commonly offer nominal leverage as high as 100x or even 125x, a prudent arbitrageur, for reasons of survival, typically keeps effective leverage well below that ceiling—say, between 3x and 10x—to retain a sufficient margin of safety. In practice, the leverage available to arbitrageurs is further and systematically suppressed by exchanges' tiered-margin regimes. On major exchanges, margin requirements are not a fixed ratio but rise in steps with the notional size of the position. Consider an illustrative tier structure (the exact tiers vary by exchange): at a low notional value, the initial margin rate may be as little as 1% (corresponding to roughly 100x maximum leverage), but as notional value rises into the millions of dollars, the initial margin rate may climb to 5% (roughly 20x), and at still larger sizes may require more than 10%. This means that the effective leverage ceiling faced by large arbitrage institutions is systematically lower than that of small participants, and their capital efficiency is markedly suppressed by the tier system. Tiered margin thus imposes another structural constraint on scaling arbitrage: the larger the position, the higher the marginal margin requirement and the lower the return on capital. On this illustrative basis, an arbitrageur with $1 million of capital may prudently run a position of only about $3 million to $10 million—far short of the $100 million that would be theoretically possible in an unconstrained case. In this way, the margin regime substantially compresses an individual arbitrageur's capacity to deploy capital.

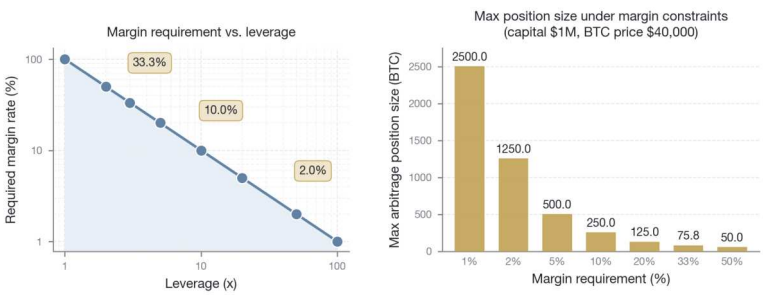


Figure 17-4. The relationship between leverage and margin requirements (Data source: drawn by the author)

Figure 17-4 reveals the key trade-off between leverage and the liquidation buffer: the higher the leverage, the narrower the price cushion before forced liquidation. At 5x leverage, the price must move roughly 20% against the position to trigger liquidation; at 20x, an adverse move of only about 5% is enough to wipe it out. A rational arbitrageur must carefully weigh capital efficiency against the probability of survival.

This constraint affects different types of arbitrage strategy very differently. For a basis trade conducted within a single exchange—for example, holding a spot long and a perpetual futures short on the same platform simultaneously—an exchange that offers portfolio margin can net the risks of the two positions and thereby lower the overall margin requirement. For the great majority of strategies that must be executed across platforms, however, the margin constraint is magnified several times over. Cross-exchange basis arbitrage (Section 16.5), for instance, requires establishing positions on two separate exchanges, so sufficient margin must be pre-funded on both. Funding rate arbitrage (Section 16.7) likewise faces margin pressure on both ends. For more complex multi-leg strategies, such as triangular arbitrage (Section 16.6), capital may need to be deployed across three or more trading pairs or platforms, and total capital demand is the sum of the margin requirements of all legs rather than the margin on the net exposure. CEX-DEX arbitrage (Section 16.9) involves two entirely different systems of capital commitment—a centralized margin account on the centralized-exchange (CEX) side and an on-chain collateral pool on the decentralized-exchange (DEX) side—which further compounds the complexity of capital management. This fragmentation of the margin regime is examined more fully in Section 17.2.4.

17.2.2 The cascading-liquidation mechanism

Brunnermeier and Pedersen (2009) [8] revealed the deep linkage between margin constraints and market liquidity: when falling asset prices tighten margin, traders are forced to sell, which pushes prices down further and triggers additional margin calls—a *margin spiral*. This mechanism is especially pronounced in perpetual futures markets.

If the margin regime is a static constraint on capital, forced liquidation is a dynamic mechanism that amplifies risk. It is not merely the failure of one trader's risk management but a market mechanism with pronounced negative externalities and positive-feedback characteristics. Under extreme conditions, a single large forced-liquidation event is fully capable of triggering a market-wide wave of liquidations—a *liquidation cascade*.

The mechanism of a liquidation cascade is a textbook positive-feedback loop. When market prices fall sharply, the margin levels of many highly leveraged long positions drop below the maintenance margin requirement, and the positions are taken over by the exchange's liquidation engine. To close these risky positions as quickly as possible, the liquidation engine typically dumps them onto the market with market orders or aggressive limit orders (see Figure 17-5). This description reflects a simplified model of early exchange liquidation mechanisms. The liquidation engines of modern major exchanges have evolved into multi-layered designs. The first layer is progressive deleveraging: when the margin ratio falls to a warning line, the system forcibly closes only part of the position to bring leverage back to a safe level, rather than liquidating the whole

position at once. The second layer is algorithmic execution: for large positions, the engine uses iceberg orders or a time-weighted average price (TWAP) strategy to execute in tranches and reduce market impact. The third layer is market-maker takeover: some exchanges maintain a liquidation-takeover engine that allows designated market makers to absorb liquidated positions at prices better than the bankruptcy price. In the DEX arena, liquidation mechanisms take different forms: dYdX v4 uses Dutch-auction liquidations, while GMX relies on a keeper network to trigger liquidations and carries an inherent delay. Although these buffering mechanisms work effectively in most market swings, in a systemic crisis—when liquidation volume grows exponentially over a short interval—every line of defense in the layered design can be breached one by one, and the process still degenerates into large-scale, market-impact selling. In an environment where liquidity has already contracted markedly, these large sell orders rapidly consume the depth on the bid side and intensify downward price pressure. The further fall in price then trips the liquidation lines of a new batch of leveraged positions, prompting still larger market selling, and the cycle repeats to form a self-reinforcing downward spiral. In the process, market liquidity is rapidly drained, the bid-ask spread widens sharply, and subsequent liquidations must execute at worse prices—compounding the destructive power of the entire cycle.

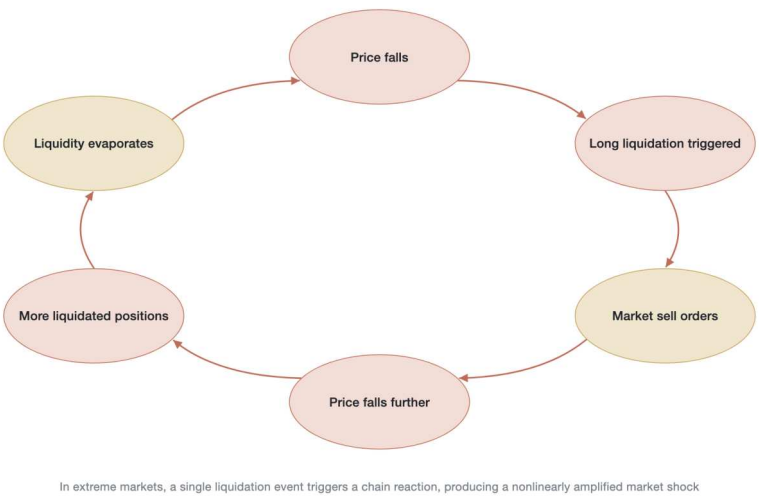


Figure 17-5. Schematic of the positive-feedback-loop mechanism of a liquidation cascade (Data source: mechanism schematic drawn by the author, not empirical data)

For arbitrageurs, the impact of a liquidation cascade is especially significant. A standard basis arbitrageur's position—a spot long plus a perpetual futures short—is, in theory, market-neutral, so price movements alone should not affect its net value. When a liquidation cascade occurs, however, this theoretical neutrality often breaks down. On the one hand, a steep price decline can shrink the value of the spot asset; if that spot is also pledged as collateral for other positions, it directly increases the arbitrageur's overall margin pressure. On the other hand, a more immediate risk arises: even if the arbitrageur's perpetual futures short is sitting on unrealized gains as prices fall, the liq-

uidation of a long position held on the same exchange—belonging, say, to an unrelated statistical-arbitrage strategy—may exhaust all the margin in the account and force the profitable short to be closed as well, or vice versa. A more dangerous scenario still is when the liquidation of an arbitrageur's position on Exchange A leaves the overall strategy no longer delta-neutral, exposing naked directional risk on the other leg at Exchange B. If Exchange B then also experiences violent price swings through market contagion, the arbitrageur may face the extreme case of large losses on both sides at once.

The mark price mechanism—praised in Chapter 15 for preventing manipulation and protecting traders—can also produce unintended adverse effects in a liquidation cascade, giving rise to what may be called a *latency trap*. The mark price is far from a simple moving average. On Binance, for example, it is the median of three candidate prices: two are computed from the spot index price plus a moving average of the funding-rate basis, and the third is the bid-ask midpoint of the contract. This design means a funding-rate component is embedded in the mark price, so that under extreme conditions the mark price itself may deviate from spot. In analyzing a liquidation cascade, one must distinguish two deviations of different natures. The first is the lagged deviation of the mark price relative to the spot index price, which reflects the inertia of the smoothing mechanism and is the direct cause of the latency trap. The second is the deviation of the perpetual futures traded price relative to the mark price, which reflects the quality of price discovery and the state of immediate liquidity within the perpetual futures itself. The causes and risk implications of the two deviations are entirely different. Because the mark price changes more smoothly than the latest traded price on any single exchange, in a normal market it effectively filters out local, transient price aberrations and prevents traders from being liquidated by a malicious "wick." When the market moves sharply in one direction, however—for example, in a cliff-edge price drop—the smoothing property of the mark price inevitably lags the true, rapidly changing market price. Many positions have therefore already reached their liquidation line in real terms, but because the mark price has not yet "caught up," the liquidation is delayed. On the surface, this delay gives traders a buffer, but in practice it can allow liquidation pressure to accumulate further. When the market keeps falling and finally drags the mark price across the liquidation threshold, the positions whose liquidation was delayed have grown to an enormous quantity; they are then thrown onto the market simultaneously, delivering a single, far more violent shock and amplifying the destructive power of the liquidation cascade exponentially.

17.2.3 The temporal constraint of no expiry date

A widely circulated market adage holds that "the market can remain irrational longer than you can remain solvent." (The line is often misattributed to Keynes, but no definitive source appears in his writings; the earliest traceable formulation is by the investment analyst A. Gary Shilling in the 1980s.) In the world of perpetual futures arbitrage, this maxim can be rewritten precisely as: "The basis can remain dislocated longer than an arbitrageur's margin can last." Behind it lies a fundamental difference in risk structure between perpetual futures and traditional deliverable futures—the open-ended risk exposure created by the absence of an expiry date, the *time trap*.

For a traditional deliverable-futures arbitrageur, the strategy carries an implicit and ultimate guarantee of certainty: the delivery date at expiry. However violently the basis between futures and spot swings during the holding period—driven by sentiment, supply-demand imbalances, and the like—once the delivery date arrives, the design of the delivery mechanism ensures that the two prices must converge. As long as the arbitrageur has enough margin to "ride out" the finite interval from opening the position to the delivery date, the arbitrage return is ultimately certain. The expiry mechanism gives the arbitrageur a guarantee that the strategy will eventually converge.

Perpetual futures shatter this certainty completely. Because they have no delivery date, their price is anchored entirely by the "streaming" mechanism of the funding rate rather than by a fixed "endpoint." This mechanism is the core advantage that lets perpetual futures track the spot price continuously, but from a risk perspective it constitutes a significant structural hazard. With no expiry date to force final convergence, the basis between the perpetual futures price and the spot price can, in theory, deviate forever. For a basis arbitrageur—buying spot and shorting perpetual futures, say—if the basis (perpetual price minus spot price) moves from negative to ever more deeply negative, the short perpetual leg keeps generating unrealized losses. Because no definite "future date" guarantees that the basis will revert, the arbitrageur must bear these unrealized losses on their own margin for an indefinite period. The consumption of capital along the time dimension is continuous and cumulative (see Figure 17-6). The cost of capital accumulates linearly over time and may ultimately exhaust the entire margin, forcing the arbitrageur out through liquidation before the basis finally reverts.

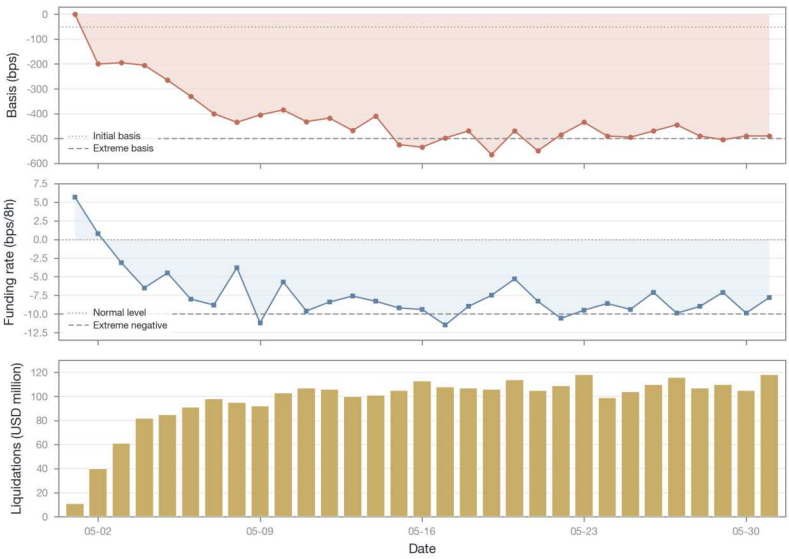


Figure 17-6. The evolution of the perpetual futures basis, funding rate, and liquidation volume during the market turmoil of May 2022 (Data source: representative series sketched by the author; the magnitudes—the basis widening from about −50 bps to below −500 bps and the rate turning deeply negative—are consistent with the text, and are not day-by-day empirical data)

Historical data have repeatedly confirmed the practical impact of this temporal con-

straint. In the market turmoil triggered by the LUNA/UST collapse in May 2022, the basis on Bitcoin perpetual futures widened sharply within a few weeks from around −50 basis points to below −500 basis points. For basis arbitrageurs who had entered beforehand, this meant their short positions came under enormous unrealized-loss pressure. At the same time, extreme market panic flipped the funding rate from positive to deeply negative, so they had not only to top up margin against the unrealized losses but also to keep paying steep funding to the longs. The source of return became a persistent cost, and margin was eroded from both sides. Many strategies that were robust under normal conditions exhausted all their margin reserves under this dual pressure of a widening basis and negative funding. Successive episodes of market extremes have repeatedly set new records for the duration and magnitude of basis dislocation. Together, these extremes constitute the "tail scenario" that arbitrageurs must take seriously and prepare for when designing their margin-management strategies. Their core risk is never a misjudgment of direction—the strategy is neutral by construction—but the exhaustion of capital before the mean reversion they firmly believe in actually occurs.

17.2.4 Cross-exchange margin fragmentation

Unlike traditional financial markets, where portfolio margining and netting are achieved through a central clearinghouse, today's cryptocurrency markets lack a comparable mechanism for optimizing capital efficiency across platforms. The market is highly fragmented: each centralized exchange is an independent, closed clearing system (see Figure 17-7) with its own account structure, risk engine, and clearing process. This means a profitable position on one exchange cannot automatically offset the margin requirement of a losing position on another. For traders engaged in cross-exchange arbitrage, this is a significant constraint on capital efficiency.

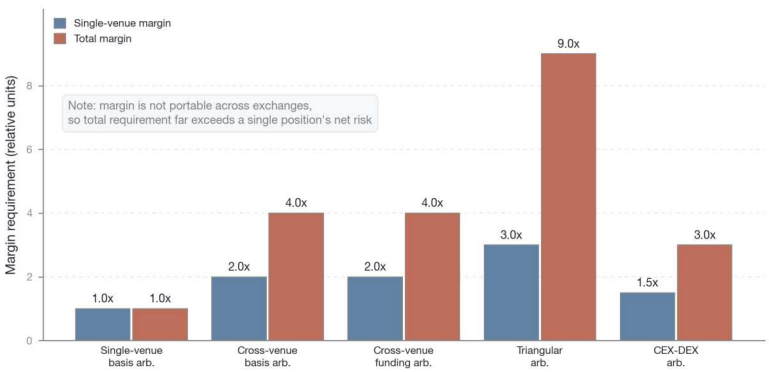


Figure 17-7. Schematic of the structure of cross-exchange margin fragmentation (Data source: structural schematic drawn by the author; the relative multiples are illustrative values, not empirical data)

As the figure shows, in a fragmented market structure an arbitrageur must maintain a separate margin account on each independent exchange, and gains and losses cannot flow automatically between them. When one end incurs an unrealized loss and requires additional margin, the unrealized gain at the other end cannot be called upon

in real time; the arbitrageur must redeploy funds manually through on-chain transfers or a withdrawal-and-deposit process—a procedure exposed to serious timing risk when markets are moving violently.

Suppose an arbitrageur finds that the BTC perpetual futures price on Binance is higher than the BTC spot price on Coinbase. The arbitrageur must simultaneously buy BTC spot on Coinbase and short BTC perpetual futures on Binance. Although the net exposure of the combined position is close to zero, sufficient funds must be deposited on both Coinbase and Binance: Coinbase needs cash to buy the spot, and Binance needs margin to open the short. If the market price rises, the Binance short generates an unrealized loss and requires additional margin; meanwhile, although the Coinbase spot is appreciating, that unrealized gain cannot be recognized by Binance's margin system and cannot automatically cover Binance's margin shortfall. The arbitrageur must manually withdraw from Coinbase and transfer to Binance—a process that is both slow and, when markets are moving violently, fraught with operational risk. To keep the strategy running robustly, therefore, the arbitrageur is forced to pre-fund each participating exchange with a large amount of "idle" margin far exceeding what the actual net risk requires, held in reserve for contingencies. As a result, total capital committed may be several times the margin that the net exposure alone would require, which markedly lowers the return on capital and directly limits the total scale of arbitrage opportunities the trader can pursue.

In recent years, some exchanges have introduced so-called unified margin accounts or portfolio-margin features, which let users pledge multiple crypto assets in the account as unified collateral and allow gains and losses to offset one another across different coins and product lines (spot, perpetual futures, options). This alleviates capital fragmentation *within* an exchange to some degree and improves capital efficiency on a single platform. For the heart of crypto-market arbitrage, however—*cross-exchange* arbitrage—the problem remains unsolved. As long as exchanges remain independent clearing entities, cross-platform margin cannot be pooled. This is a core competitive advantage of large, well-capitalized trading institutions—top high-frequency trading firms and market makers—over small and midsize arbitrageurs. Their vast capital lets them pre-fund enormous sums across all major exchanges, enabling them to capture the fleeting cross-exchange opportunities that demand the highest capital efficiency—a domain ordinary participants can scarcely reach.

In practice, professional arbitrage firms have developed tools that partly mitigate cross-exchange margin fragmentation. Crypto prime-brokerage providers, such as FalconX and Hidden Road, offer cross-exchange netting and credit lines that let institutional clients adjust their effective margin levels across platforms without physically moving funds. Off-exchange settlement networks, such as Copper ClearLoop, reduce the need to shuttle funds by custodying capital with an independent third party while the client trades across multiple exchanges. Yet these intermediary services introduce a new counterparty risk of their own. As the impact of the FTX collapse on institutions that used its settlement infrastructure revealed, reliance on intermediary services can become a new node for transmitting systemic risk in a crisis.

17.2.5 A quantitative framework for the margin constraint

The margin constraint is not merely a qualitative risk factor but a mathematical problem that can be quantified precisely. For a rational, risk-averse arbitrageur, the objective is not to maximize the return on any single trade but to seek long-term, steady capital growth subject to a given probability of survival (that is, of not being liquidated). This naturally raises a central quantitative question: given the capital base, the margin requirements, and the expected distribution of future market volatility (particularly basis volatility), what is the optimal size of the arbitrage position?

Building such a quantitative framework requires the joint consideration of several key parameters. The first is the exchange's margin rules, including the initial margin rate and the maintenance margin rate; these two ratios directly define the upper bound on leverage and the threshold that triggers liquidation. The second is the arbitrageur's own total capital and risk appetite; as shown in Figure 17-8, the latter can be expressed as the maximum drawdown one is willing to bear or an acceptable probability of liquidation (for example, wishing to keep the probability of being liquidated within the next year below 1%). The third, and most central, input variable is the estimate of the future volatility of the target spread (such as the basis). This estimate cannot rest on the historical average alone but must fully account for the *fat tails* of the distribution—the fact that the likelihood of extreme, low-probability events is far higher than a normal distribution predicts. Drawing on tools designed to characterize the tail of a distribution, such as extreme value theory, rather than simply applying a normality assumption, is essential to computing a sufficiently robust margin level.

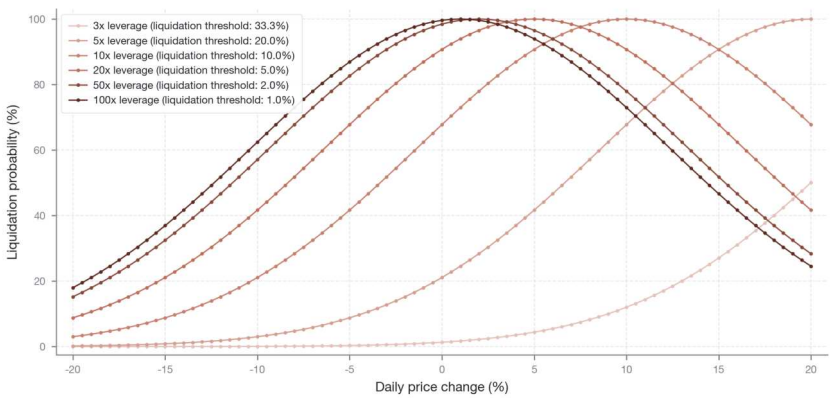


Figure 17-8. The relationship between the magnitude of intraday price movement and the probability of liquidation across leverage levels (Data source: drawn by the author; an illustration of modeling based on extreme value theory, not empirical data)

The figure illustrates the relationship between the magnitude of intraday price movement and the probability of liquidation across leverage levels. The higher the leverage, the smaller the price move needed to trigger liquidation (the liquidation threshold is only about 1% at 100x leverage, versus about 33.3% at 3x); hence, for the same market volatility, a highly leveraged position has a markedly higher probability of liquidation than a low-leverage one. Modeled with extreme value theory, the curve highlights the fragility of highly leveraged arbitrage positions in the face of tail price movements—the

fundamental reason a rational arbitrageur must deliberately hold effective leverage far below the ceiling.

In a simplified model, consider a basis-arbitrage strategy whose risk stems mainly from an adverse widening of the basis. The daily volatility of the basis, σ, can be estimated from historical data or from options-implied volatility. If the arbitrageur uses L-times leverage, the initial margin rate is $1/L$. An adverse move in the basis large enough to cause liquidation is approximately $(1/L) \times (1 - \text{MMR/IMR})$—that is, IMR − MMR, where IMR is the initial margin requirement and MMR the maintenance margin requirement. Given the probability distribution of basis movements, one can compute the probability that, on any given day, the basis moves beyond this liquidation threshold. Conversely, given an acceptable single-day liquidation probability P_target (0.01%, say), one can back out the maximum safe leverage L_safe permissible under that risk constraint. The optimal arbitrage position size is then total capital multiplied by L_safe.

The core insight of this quantitative framework is that the risk-adjusted optimal position is typically far smaller than the position that could be built in an "unconstrained" ideal world. The force of arbitrage is systematically and endogenously "discounted." This provides a solid microfoundation for the theory of the limits of arbitrage set out in Section 17.1. Precisely because every rational arbitrageur imposes this kind of self-constraint, the market's efficiency gap is never fully or instantly closed, and arbitrage returns therefore persist as compensation for margin risk. The framework also makes clear the dynamic relationship between risk and efficiency: when market volatility intensifies or exchanges raise margin requirements, the safe leverage L_safe that every arbitrageur computes falls systematically, the market-wide total of arbitrage positions contracts, and market efficiency declines predictably. This mechanism is the microfoundation for how a tightening of constraints maps directly into a lower upper bound on market efficiency, echoing the *efficiency spectrum* framework established in Chapter 15.

The framework above implicitly assumes that basis movements are independent and identically distributed across periods. Empirical research shows, however, that basis series exhibit pronounced autocorrelation and volatility clustering: extreme deviations are often not isolated single-day events but successive shocks lasting several days. A more robust analytical tool than the single-day liquidation probability is therefore the *first-passage-time distribution*—the time it takes for the margin balance to fall from its initial level below the liquidation threshold for the first time. This approach captures path-dependent risk and provides a more prudent basis for setting the margin buffer.

17.3 Funding rate reversal

The funding rate mechanism gives basis arbitrageurs a persistent source of positive carry. Yet it is here that the intrinsic contradiction in the perpetual futures design is most fully revealed. The funding rate that serves as the source of arbitrage return is not itself a fixed parameter but a random variable that swings sharply with market sentiment, supply and demand, and price expectations. When market conditions reverse, the

anchoring mechanism abruptly changes direction—from a source of profit to a persistent drain on capital. This chapter calls this phenomenon *funding rate reversal*.

To understand the mechanism of funding rate reversal precisely, it helps to recall how the rate is computed. The standard funding rate has two components: the interest rate component (usually fixed at 0.01% per 8 hours, reflecting the interest differential between holding stablecoins and holding crypto assets) and the premium component (which measures the deviation of the perpetual futures price from the spot index). The final funding rate is typically further bounded by a clamp function to prevent the rate from diverging without limit under extreme conditions. Different exchanges set different rate bounds: Binance, for example, caps each period's rate within the $[-2\%, +2\%]$ band, and Bybit's limits are broadly comparable. These bounds seem immaterial in a normal market, but under extreme conditions they constitute a key institutional constraint: they effectively cap the loss from any single funding period and thus limit, to a degree, the instantaneous destructive power of a rate reversal. Even with the single-period rate capped, however, extreme negative rates over many consecutive periods can still erode margin fatally through their cumulative effect—precisely the *tail risk of funding-rate polarization* discussed below.

This section examines a risk dimension unique to perpetual futures arbitrage. Unlike traditional deliverable-futures arbitrage, in which the basis converges deterministically over time, the funding rate faced by perpetual futures arbitrageurs is fundamentally uncertain. When the funding rate on which the arbitrageur relies for return reverses, it not only drives profit to zero but can inflict substantial losses in a short period. We begin with the *negative funding rate trap* latent in a positive-carry strategy and, through a quantitative analysis of historical funding-rate data, reveal the tail risk of its polarization. We then take the Ethena protocol as a case study to examine the systemic fragility exposed when scaled-up, protocol-level arbitrage meets sustained negative funding. Finally, the section shows how funding rate reversal and the margin pressure discussed in the previous section form a cross-amplifying effect—a risk resonance of considerable destructive power—providing the first concrete illustration of why arbitrage capacity collapses nonlinearly in extreme markets.

17.3.1 The negative funding rate trap

The return model of the basis-arbitrage strategy rests on a key implicit assumption: that the funding rate will remain positive. When the market environment changes drastically, this assumption fails.

At that point, the *negative funding rate trap* is sprung. In a panic sell-off, investors rush into the perpetual futures market to short as a hedge or a speculation, demand for short positions surges, and the perpetual futures price falls to a deep discount to spot. By the logic of the funding-rate calculation, the rate quickly turns negative in order to balance the long and short sides. The cash flow that once served as a source of return now reverses: it is the shorts—the basis arbitrageurs—who must pay the longs. The positive-cash-flow structure the arbitrageur built not only stops producing income but turns into a persistent negative cash flow.

The problem does not end there. Funding rate reversal usually coincides with a sharp

fall in the underlying asset's price, which subjects the basis arbitrageur to double or even triple pressure. First, negative funding is a direct and continuing operating cost that erodes the position's capital. Second, although the short perpetual futures leg earns unrealized gains as prices fall, the equivalent spot holding suffers a real loss. The symmetry of this profit-and-loss settlement depends heavily on the margin-account mode the arbitrageur uses. Under isolated margin, each position's profit and loss is accounted for separately, and the short's unrealized gain indeed cannot be automatically released as available margin for other positions. Under cross margin, the unrealized profit and loss of all positions on the same exchange jointly determine the account's available balance, and the short's unrealized gain can offset the spot-side pressure to some degree. Under a unified margin account (such as the portfolio-margin feature some exchanges offer), profit-and-loss hedging across product lines is more efficient still. Yet however optimized the within-exchange account mode may be, the settlement asymmetry in the cross-exchange case always persists—and it is this that is the core vulnerability through which margin pressure is amplified in a crisis. When an arbitrageur's spot loss on Exchange A cannot be offset by the unrealized gain on perpetual futures at Exchange B, the spot loss is reflected immediately in account equity, whereas the unrealized gain on the perpetual futures leg cannot be fully released as available margin until the position is closed. This mismatch in settlement timing significantly increases margin pressure, as discussed in Section 17.2. Finally, when the arbitrageur, unable to bear the continuing negative-funding cost and margin pressure, tries to close out, they often find that market liquidity has dried up, producing enormous exit slippage, and the basis may already have widened far beyond expectations. A mathematically flawless delta-neutral strategy thus turns, under real-world constraints, into an adverse position that steadily consumes capital.

17.3.2 The tail risk of funding-rate polarization

Understanding the negative funding rate trap cannot stop at the qualitative level; a quantitative analysis is needed to grasp the true magnitude of the risk. Because the funding rate is a derivative of a market price, the statistical distribution of its time series reveals the source of its risk. Unlike an idealized random process that follows a normal distribution, the funding rate on perpetual futures exhibits pronounced leptokurtosis and fat tails, meaning that extreme events—rates far above or far below the average—occur with a probability well beyond what a normal distribution predicts.

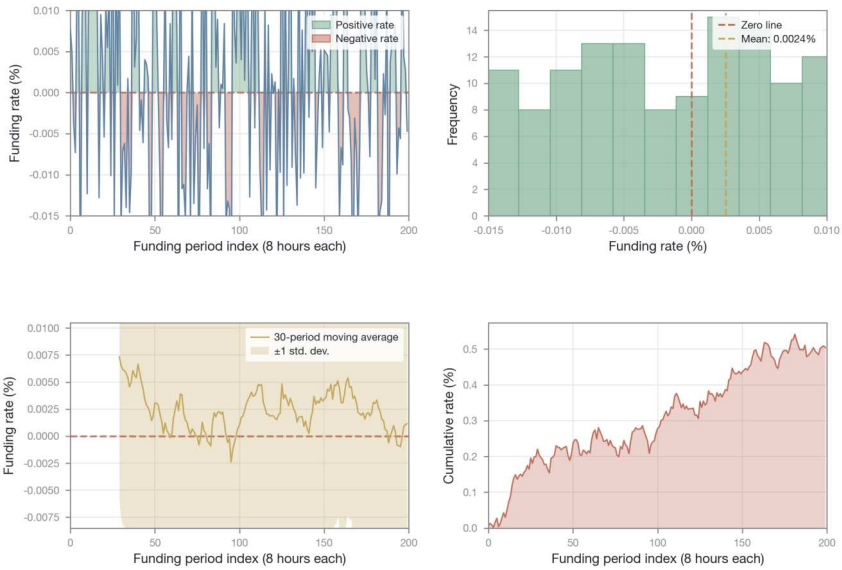


Figure 17-9. The statistical distribution and fat tails of the funding rate on BTC perpetual futures (Data source: drawn by the author; simulated/illustrative data based on extreme value theory, not empirical; for the fat-tail phenomenon shown, see Coinglass [9])

The four panels of Figure 17-9 together display the leptokurtosis and fat tails of the BTC funding rate: the rate clusters mostly near zero, while the extreme negative-rate events in the tail of the distribution pose the gravest threat to a basis-arbitrage strategy.

This fat-tailed distribution means that relying on the historical average rate to assess a strategy's expected return is seriously misleading. The historical mean masks the destructive potential of extreme events. A strategy that appears to yield 10% annualized may, in a single extreme negative-funding episode lasting several days, lose months' or even a year's worth of profit. Stress testing therefore becomes an indispensable part of an arbitrageur's risk management. The essence of a stress test is to simulate the maximum loss an arbitrage position might face, and its ability to survive, under extremely adverse market conditions.

A simple stress-test framework can be built as follows. Consider a basis-arbitrage position at 10x leverage, with an initial margin requirement of 10% and a maintenance margin requirement of 5%. The question is how long the position can avoid forced liquidation when it meets a run of extreme negative funding. As a stress-test example, assume an extreme negative-funding scenario (the parameter magnitudes can be referenced to historical runs of negative funding, but the figures below are assumed values set for illustration): an average funding rate of −0.05% over 21 consecutive settlement periods (that is, 7 days). Over those 7 days, the arbitrageur's total cost is `21 × −0.05% = −1.05%`. For a 10x-leveraged position, a 1.05% loss of notional principal consumes 10.5% of the initial margin (10%). Extending the extreme assumption to 42 consecutive settlement periods (14 days), the cumulative loss reaches 2.1% of notional principal and consumes 21% of the initial margin, materially eroding the margin buffer and driving the position toward the maintenance-margin line. This calculation

makes clear that, in an extreme rate-reversal scenario, the fragility of a highly leveraged arbitrage position is sharply amplified.

This leads to an important qualification of the mean-reversion property of the funding rate discussed in Chapter 16. Chapter 16 noted that the presence of arbitrage causes the funding rate to revert to its mean over the long run. This is correct in theory, but for a real arbitrageur in the market it can be a dangerous oversimplification. The key problem is that the speed and path of mean reversion are highly uncertain. As the widely circulated adage warns, "the market can remain irrational longer than you can remain solvent." Before the funding rate reverts to its long-run mean, a leveraged arbitrageur may already have been liquidated for want of margin. A quantitative assessment of tail risk, and a capital buffer sized accordingly, is therefore the decisive dividing line between the professional arbitrageur and the amateur.

17.3.3 Ethena and sustained negative funding

Once basis arbitrage is run at scale and packaged into a protocol, the risk of funding rate reversal takes on a systemic character. The Ethena protocol packages the "hold spot, short perpetual" basis trade into the synthetic dollar USDe, and the yield to sUSDe holders derives almost entirely from positive funding on perpetual futures.

The fragility of this model is laid bare when the funding rate reverses. When market sentiment turns and the rate goes negative, the protocol's source of return instantly becomes a cost of carry. This triggers a clear chain of risk transmission: negative funding lowers the sUSDe yield, and rational holders begin to redeem USDe. Through a behavioral-finance lens, this redemption process is likely to be highly nonlinear: when the sUSDe yield declines mildly, most holders tend to wait and see (the joint effect of anchoring and the disposition effect); but once the yield falls below some psychological threshold—below zero, say, or below the rate on a traditional money-market fund—it can trigger the self-fulfilling run described by Diamond and Dybvig (1983) [10], and redemptions escalate in an instant from a trickle to a flood. The protocol is then forced to close its short positions on the market to meet redemptions; the large-scale unwinding pushes the perpetual futures price up and drives the funding rate more negative still; and more holders redeem in panic. This positive-feedback loop—negative funding $\rightarrow$ redemptions $\rightarrow$ unwinding $\rightarrow$ still more negative funding—is the core systemic risk facing protocol-level arbitrage.

In addition, Ethena's collateral mix includes liquid staking derivatives such as stETH, which introduces an extra risk layer independent of the funding rate: in an episode of extreme market panic—such as the Celsius/3AC events of June 2022—stETH can trade at a marked discount to ETH, further eroding the value of Ethena's collateral and compounding the shock from funding rate reversal. A detailed analysis of this risk layer appears in Section 17.10.

During the market correction of April 2024, this transmission mechanism received an initial confirmation: on-chain data from Ethena Labs (2024) [11] show a marked increase in USDe redemptions, though this did not escalate into a full-blown run at the time. A complete analysis of Ethena's risk exposure—including the "too big to unwind" paradox created by its concentration, the item-by-item mapping onto the seven-dimensional

constraint topology, and a detailed review of the April 2024 stress test—is developed in Section 17.10.

17.3.4 The cross-amplification of funding and margin

If funding rate reversal and margin pressure are challenges the arbitrageur meets along two separate dimensions, then when both dimensions are activated at once, the destructive power they generate is not a simple linear sum but a cross-amplifying effect with the character of resonance. This resonance is the key to understanding why arbitrage strategies collapse nonlinearly in extreme markets, and it is the first concrete instance of the *risk resonance model* examined in depth later in this chapter.

The mechanism of this cross-amplification can be analyzed by reconstructing a typical market-crash scenario.

A sharp market decline triggers the initial pressure. The market meets sudden bad news, and prices begin to fall steeply. For an arbitrageur holding a "buy spot, short perpetual" basis position, the delta-neutral structure begins to be tested. Although the short perpetual leg earns unrealized gains as prices fall, the spot holding is suffering a real loss. As noted earlier, because of the asymmetry in the settlement mechanism, the spot loss directly reduces account equity, whereas the unrealized gain on the perpetual leg often cannot be fully counted as available margin until the position is closed. The arbitrageur's margin level therefore begins to fall, bringing initial pressure from the margin-and-forced-liquidation dimension.

On top of this, panic drives the funding rate to reverse. As panic spreads, large numbers of investors rush into the perpetual futures market to short, and the funding rate flips quickly from positive to negative. The funding-rate-reversal dimension is now activated. The arbitrageur's former source of return has instantly become a continuing cost, and with every funding settlement period the account's margin is passively consumed.

When these two pressures act together, the cross-amplification of the double pressure emerges. The arbitrageur is now trapped by the superposition of two pressures. On the one hand, the continuing price decline steadily widens the loss on the spot side and keeps pressing on margin. On the other, the persistently negative funding rate keeps consuming margin. These two originally independent sources of pressure now form a vicious cycle: negative funding accelerates the consumption of margin, so the arbitrageur approaches the maintenance-margin line faster and is more easily liquidated; at the same time, the increased margin pressure sharply reduces the arbitrageur's capacity to withstand the shock of continued negative funding. A position that could originally have endured weeks of negative funding may, against the backdrop of a large spot-price decline, be liquidated within a few days.

This cross-amplification markedly shortens the arbitrageur's "survival time." A position that could hold up under pressure along a single dimension becomes extraordinarily fragile when the two pressures resonate. It captures the core idea of the limits-of-arbitrage theory: the arbitrageur's risk stems not from any logical flaw in the strategy but from the fact that, under real-world constraints, they "exhaust the capital needed to hold the position before the basis converges or the market returns to rationality."

The resonance of funding rate reversal and margin pressure is precisely the catalyst and accelerant of this capital-consumption process, revealing how, in a complex financial system, risk propagates across dimensions and ultimately produces systemic instability. This lays an important intuitive foundation for the more comprehensive risk resonance model built in Section 17.9.

17.4 Counterparty and platform risk

In the risk topology of perpetual futures arbitrage, if margin and funding-rate risks are structural constraints endogenous to the contract's design, then counterparty and platform risk is the fragility of the infrastructure on which arbitrageurs themselves depend. Traditional financial markets manage and dissolve counterparty risk through the institutional innovation of the central counterparty (CCP), letting traders concentrate on market risk itself. In the current cryptocurrency market structure, however, there is no industry-wide central counterparty. Each centralized exchange is an independent, externally unaudited, self-regulating clearing center and asset custodian. This fragmented structure means that any cross-exchange arbitrage position is not merely a judgment about a market spread but an implicit trust that two (or more) platforms will remain financially sound, operationally fair, and technically available. When that basis of trust dissolves, an arbitrage strategy—however mathematically perfect—fails instantly.

This section examines three core manifestations of platform risk: the outright collapse of an exchange, temporary functional failures (such as a withdrawal freeze), and, as the last line of defense, insurance-fund depletion and the auto-deleveraging mechanism. Together, these risks constitute the highest-risk "broken-leg" scenario in arbitrage trading—when one leg of a hedge fails through platform risk and the other is left fully exposed to market swings.

17.4.1 Lessons from the FTX collapse

The collapse of FTX in November 2022 gave the entire crypto market its most painful and profound case study of platform risk. For arbitrageurs who relied on cross-platform operations, the FTX collapse was no distant piece of industry news but a direct blow to their trading positions. It vividly illustrated how, when one of the platforms a strategy depends on suddenly fails, a carefully constructed hedge can turn in an instant into a fully exposed, one-sided bet.

Consider a typical basis arbitrageur's situation. During a broadly bullish period, the arbitrageur buys 10 BTC of spot on Binance while shorting BTC perpetual futures on FTX at an equivalent notional. This is a textbook delta-neutral strategy: movements in the spot price are fully hedged by the profit and loss on the perpetual futures, and the arbitrageur's return comes mainly from the persistently positive funding rate on FTX. When FTX abruptly announced a suspension of user withdrawals on November 8, however, the risk character of this strategy changed fundamentally.

As panic spread, the BTC price began to plunge. The value of the arbitrageur's spot long on Binance shrank rapidly, producing a large unrealized loss. Ordinarily, that loss

would be hedged by the unrealized gain on the short perpetual position at FTX. But by now FTX account information was inaccessible, and the short's "profit" was merely a number on a screen—it could not be closed, still less withdrawn. The arbitrageur could neither transfer the FTX gain to Binance to cover the spot loss nor close the short on FTX and sell the spot on Binance to terminate the whole strategy. As a result, the short leg meant to serve as a hedge was effectively "dead," leaving only the Binance spot long to bear all the market's downside alone. A "risk-free" arbitrage strategy degenerated in an instant into a naked long position.

FTX's subsequent bankruptcy proceedings and regulatory filings revealed the root of the problem. According to charges by the Commodity Futures Trading Commission (CFTC) and bankruptcy filings, the platform had a shortfall in customer assets on the order of $8 billion, and its affiliated trading firm, Alameda Research, had misappropriated about $10 billion of customer funds [12]; early media reports likewise indicated that at least $1 billion to $2 billion of customer funds was unaccounted for [13]. This meant that even though the arbitrageur's short position showed a profit on paper, the money was long gone from the exchange's vault. Platform risk here revealed its most extreme form: the outright default of the counterparty—the exchange itself.

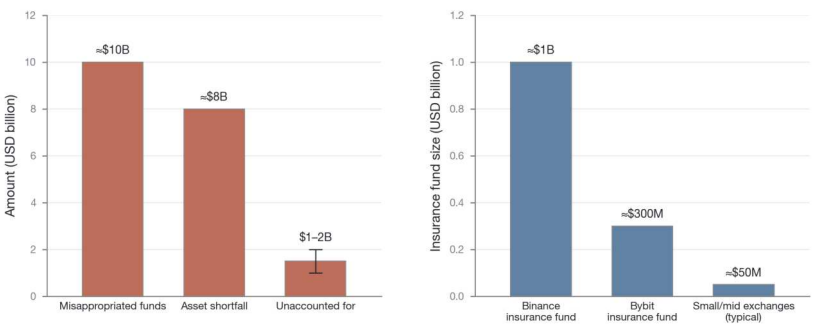


Figure 17-10. Comparison of counterparty-risk characteristics across exchanges (Data source: compiled by the author)

The left panel of Figure 17-10 depicts the enormous asset shortfall at the time of the FTX collapse (its magnitude is given above [12]), against which any hope of recovering funds from the platform faced serious obstacles. This warns all market participants, and arbitrageurs in particular, that they must treat an exchange's solvency risk as a core consideration. Unlike traditional financial institutions, the great majority of crypto exchanges are financially opaque and lack independent third-party audits, and their "proof of reserves" often fails in practice to reflect their true liabilities in full.

More important, the FTX episode was not a unique black swan in crypto history. As Figure 17-11 shows, platform-risk events—exchange failures, hacks, and regulatory shutdowns—have been common over the past several years. From the collapse of Mt. Gox to the recent withdrawal suspension at BlockFills, these events together form a clear pattern: platform risk is a persistent, far-from-rare endogenous risk in the structure of crypto markets. For cross-exchange arbitrageurs, this means the success of

a strategy depends not only on the convergence of a spread but also on whether every trading platform they rely on survives this game of attrition.

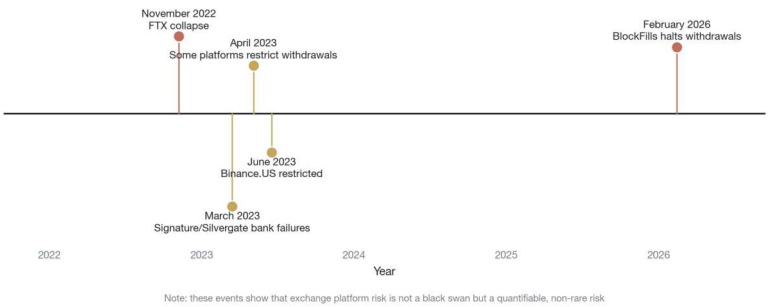


Figure 17-11. A timeline of platform-risk events at crypto exchanges (2022–2026) (Data source: compiled by the author from public reports; for the FTX collapse, see [12]; for the restrictions on Binance.US, see [14])

The assessment of platform risk should therefore not be a binary "present-or-absent" judgment but a probabilistic evaluation grounded in the historical frequency of events, an exchange's operational transparency, its corporate-governance structure, and the regulatory environment of its jurisdiction. Every cross-exchange arbitrage is an act of dispersing capital across multiple trading platforms of differing credit risk, and the failure of any one of them can cause the entire strategy to fail badly.

17.4.2 Withdrawal restrictions and frozen funds

Compared with the extreme case of an outright exchange collapse, a more common and more insidious platform risk is a temporary functional failure, the most typical of which is a suspension of withdrawals. Amid violent market swings or under regulatory pressure, an exchange may unilaterally announce a suspension of user withdrawals of all or some assets. Although such an action does not necessarily mean the platform is near bankruptcy, its destructive power is just as severe for an arbitrageur to whom every second counts. It creates a temporary "liquidity trap" within the arbitrageur's funding network, severing the flow of capital between platforms and thereby dissolving the basis for cross-market hedging.

Return to the cross-exchange arbitrage setting. What lets an arbitrageur maintain risk neutrality across markets is, at its core, the ability to move funds freely between platforms to meet the margin requirements of different positions. When the market price rises, for instance, the spot long leg of a cross-exchange basis position earns an unrealized gain while the short perpetual leg incurs an unrealized loss and may face a margin call. The arbitrageur must then withdraw part of the profit from the exchange holding the spot and transfer it to the exchange holding the perpetual futures to keep the short from being liquidated. The free flow of funds is the core precondition for sustaining the entire hedge.

Once the exchange holding the perpetual futures announces a withdrawal suspension,

however, this critical channel is blocked. Even if the arbitrageur is well capitalized overall, they cannot inject funds into the account that is generating losses. This creates a structural dilemma: a trader who is well capitalized globally may nonetheless be forcibly liquidated on a single platform because of a local liquidity shortage. In June 2023, after the U.S. Securities and Exchange Commission (SEC) sued Binance.US, the platform announced a suspension of its dollar withdrawal channel [14], which directly broke the path of arbitrage strategies that relied on it for fiat deposits and withdrawals. Similarly, many exchanges have over the years briefly suspended deposits and withdrawals of particular tokens for reasons such as technical upgrades, wallet maintenance, or responses to hacks—each a direct threat to strategies that rely on those tokens for arbitrage.

Beyond restrictions the exchange imposes deliberately, *funds-in-transit* risk is another form of liquidity trap. When the network is extremely congested, an ordinary on-chain transfer may take tens of minutes or even hours to confirm. During this in-transit period, the arbitrageur can neither use the funds nor recall the transaction. The market price may already have moved violently, and a position that needed additional margin may long since have been liquidated. This delay, rooted in the performance bottlenecks of the blockchain network itself, further compounds the uncertainty of moving funds across platforms.

The core harm of withdrawal restrictions and frozen-funds risk, therefore, is that they break the assumption that "capital is a single, unified whole" and split the arbitrageur's capital into mutually isolated "risk islands." On each island, every position must bear its margin pressure alone, without support from profitable positions on other platforms. This sharp decline in capital efficiency forces the arbitrageur to hold excess margin far above normal levels on every platform, systematically reducing the scale and efficiency of capital deployment across the entire arbitrage industry—and explaining why many seemingly profitable spreads are so slow to be fully closed.

17.4.3 Insurance-fund depletion and auto-deleveraging

In the risk-management system of a perpetual futures exchange, the insurance fund and the auto-deleveraging (ADL) mechanism are the last two lines of defense against the shock of mass liquidations. Both, however, were designed to protect the exchange from bankruptcy rather than to protect individual traders. Under extreme conditions, when the insurance fund is depleted and auto-deleveraging is triggered, platform risk appears in a more paradoxical form: it no longer punishes the losers but achieves a cold "socialization of losses" by forcibly closing the winners' positions. For arbitrageurs, this means the most solid, profitable leg of their strategy may be precisely the one that ends up bearing the market's risk of last resort.

The insurance fund's core function is to absorb *bankruptcy losses*—losses on a forcibly liquidated position that exceed all of its margin. When such a shortfall arises, the uncovered portion is borne by the insurance fund. The fund is capitalized mainly from two sources: the exchange's own injected capital, and the "surplus" generated by liquidation orders that fill at prices better than the bankruptcy price. As shown in the right panel of the earlier figure, the insurance funds of major exchanges typically

range from tens of millions to hundreds of millions of dollars. In normal market swings, this is enough to cover sporadic bankruptcy losses. In a systemic flash crash or short squeeze, however, when thousands of positions are liquidated densely within a short interval, the total loss they generate can rapidly drain the entire insurance-fund pool.

When the insurance-fund balance falls to zero or a preset danger threshold, the auto-deleveraging mechanism is activated. The logic of auto-deleveraging is direct (see Figure 17-12): since there is no longer any common fund to fill the losses, the system must find those in the market who are "able" to bear them. Who is most able? Those holding profitable positions in the prevailing market direction. Using a complex ranking algorithm, the system sets deleveraging priority by a combined ordering of the profit-and-loss ratio and effective leverage. The ranking algorithms differ markedly across exchanges: Binance ranks by the product of profit percentage and a leverage percentile; Bybit ranks by the product of profit-and-loss percentage and effective leverage, so that rank = PnL% × effective leverage; and OKX's ranking logic differs again. These differences mean that the same arbitrageur may face very different ADL risk on different exchanges: a position that ranks near the top and is highly likely to be deleveraged on one exchange may rank lower and be relatively safe on another. This differentiated deleveraging risk is an important factor the arbitrageur must weigh when choosing platforms and allocating positions. Starting from the highest-ranked, the algorithm then forcibly closes the winners' profitable positions one by one. These positions fill not at the current market price but at the bankruptcy price of the original losing position that caused the shortfall. In this way, part of the winners' profit is used to cover the losers' shortfall, ensuring the exchange operates with zero loss.

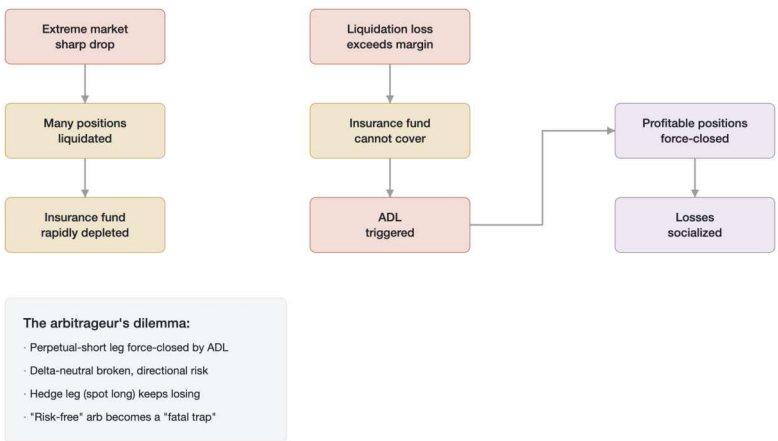


Figure 17-12. Schematic of the auto-deleveraging trigger mechanism and profit-ranked deleveraging (Data source: drawn by the author, referencing Bybit [15])

The figure clearly shows the trigger path of the auto-deleveraging mechanism and its serious impact on arbitrageurs. For a basis-arbitrage strategy running in a falling market (long spot, short perpetual futures), the short perpetual leg is generating large unrealized gains. To the auto-deleveraging system, this highly profitable short is the

perfect "counterparty" to cover the losses from the many longs being liquidated across the market. The short is therefore highly likely to be selected and forcibly closed by the ADL system. As a result, the arbitrageur not only fails to realize the expected profit but falls once again into the broken-leg predicament: the short leg that hedged the risk vanishes, leaving only the spot long that is generating continuing losses. A strategy that should have been robust in a falling market ends up suffering a major loss at the hands of auto-deleveraging.

The existence of auto-deleveraging fundamentally challenges the intuition that "to be profitable is to be safe." It reveals a counterintuitive rule of crypto derivatives markets: in an extreme crisis, the first to be sacrificed may not be the weakest but those holding the largest unrealized gains. By virtue of their market-neutral design, arbitrageurs often end up among the few winners in a chaotic market—which is precisely what makes them the ADL system's top targets for "requisition." This asymmetry of risk is a core variable that every perpetual futures arbitrageur must understand deeply and incorporate into their risk model.

17.4.4 The broken-leg scenario

Combining the three platform risks above—the outright collapse of an exchange, the temporary freezing of operational functions, and, finally, forced closure through auto-deleveraging—we can construct a unified risk framework: the *broken-leg* scenario. Its core idea is that for any hedge built on multiple platforms and multiple positions (multiple "legs"), the most fragile link is the platform that keeps those legs operating in coordination. Any platform event that causes one leg to fail involuntarily unbalances the entire strategy and exposes it to raw market risk.

The essence of broken-leg risk stems from the deeply rooted "asset fragmentation" and "risk siloing" of the crypto market structure. Unlike traditional financial markets, where assets are held centrally at institutions such as the Depository Trust & Clearing Corporation (DTCC), crypto assets are scattered across hundreds of independent exchange wallets. This means that in a cross-exchange arbitrage, the spot-long leg and the perpetual-short leg are entirely separate, both physically and legally. When one exchange runs into trouble, the other has neither the obligation nor the ability to "see" or "compensate" the trader's losses elsewhere.

The broken-leg scenario can be graded into three levels according to the severity of the platform risk.

A Level 1 break (a functional break) is triggered by problems such as withdrawal restrictions, funds-in-transit delays, or an API disconnection. Here the arbitrage position itself has not disappeared, but the trader has lost the ability to move margin between platforms and rebalance risk. The coordination among the strategy's legs is temporarily interrupted, and the position can be passively liquidated for want of margin in one location.

Above this, a Level 2 break (a forced break) is triggered by the auto-deleveraging mechanism. Here the strategy's profitable leg is forcibly closed by the system: not only is the hedge structure destroyed, but this comes at the cost of sacrificing realized profit.

The trader is exposed to directional risk and loses gains that should have been locked in.

The most destructive is a Level 3 break (a fundamental break), triggered by events such as an exchange collapse, a hack, or a regulatory shutdown. Here all assets held on that exchange may be permanently lost, producing the most direct and complete loss of principal.

These three levels affect different types of arbitrage strategy differently. For arbitrage conducted purely within a single exchange (such as a spot-versus-perpetual basis trade on the same exchange), platform risk is concentrated and single-point: the trader need only assess the reliability of that one platform. For the great majority of arbitrage types, however—cross-exchange spread arbitrage, funding rate arbitrage, and even CEX-DEX arbitrage spanning CEXs and DEXs—the risks are additive. The trader must face not only the CEX's human-management, operational, and solvency risks but also the DEX's smart-contract-vulnerability, oracle-manipulation, and on-chain-congestion risks. The more platforms a strategy spans, the more potential "break points" it has, and the fragility of the entire arbitrage structure grows exponentially.

A mature arbitrageur's risk model must therefore include a quantitative assessment of the probability of a broken leg. Such an assessment should weigh multiple dimensions: the exchange's reputation rating, the ratio of insurance-fund size to total open interest, the historical frequency of withdrawal freezes, the history of ADL triggers, technical stability, and regulatory clarity. Ultimately, arbitrage returns also embed a substantial premium for bearing the fragility of platform infrastructure. In perpetual futures trading, assessing and managing platform risk matters more than judging market direction.

17.5 Execution and timing risk

In the frictionless world of theoretical finance, arbitrage is portrayed as an elegant, instantaneous process: the moment a price dislocation appears, the arbitrageur simultaneously buys the undervalued asset and sells the overvalued one at zero cost and zero delay, locking in a risk-free profit. The reality of perpetual futures markets, however, differs markedly from this idealized model. As this chapter has repeatedly stressed, the gulf between theory and practice is made up of a series of demanding constraints, and execution and timing risk is one of the most microscopic yet consequential of them. It reveals a key fact: in a real trading environment built of fiber-optic cable, microwave towers, and globally distributed servers, "almost simultaneous" is never the same as "simultaneous." A delay of milliseconds—or even microseconds—is enough to turn a mathematically perfect arbitrage strategy into a gamble fraught with uncertainty.

The core thesis of this section is that every arbitrage strategy involving multiple trade "legs"—whether across maturities, markets, or assets—rests its profitability on a fragile assumption: that all legs can be executed synchronously at the expected prices and times. When that assumption breaks, risk is never far behind. Execution risk is not a single risk but a complex matrix composed of slippage, latency, and partial fills. It springs from the fragmented structure of the market, the technical heterogeneity of different trading venues, and the ever-intensifying race for speed in the high-frequency

arena. Understanding this constraint dimension is the key to understanding why so many theoretically available arbitrage opportunities cannot be exploited in reality, and why market efficiency always confronts a nearly insurmountable "last mile."

17.5.1 Slippage, latency, and partial fills

The success or failure of a multi-leg arbitrage strategy depends on the precision of execution. Any deviation from the expected execution directly erodes, or even wholly offsets, the anticipated arbitrage profit. Such deviation takes three principal forms—slippage, latency, and partial fills—which together constitute what is called *inter-leg risk*.

Slippage risk refers to the adverse gap between the actual average fill price of a large order and the best quote at the time the order was placed. When liquidity is thin or the market is highly volatile, a market order may "punch through" several price levels in the order book, producing an average fill far worse than expected. For a strategy such as arbitrage, whose profit margins are usually thin, a few basis points of slippage can turn it from profitable to unprofitable. A cross-exchange opportunity with an expected return of 5 basis points, for example, is doomed the instant it executes if each side incurs 3 basis points of slippage.

In practice, slippage is only one component of execution cost. A complete arbitrage profit-and-loss analysis must cover four core cost categories. First, trading fees: on major exchanges the taker fee is typically 0.04%–0.06%, so two-sided execution consumes total fees of roughly 0.08%–0.12%. Second, slippage and market-impact costs, whose magnitude, as discussed above, depends heavily on the market state and the position size. Third, the time cost of funding: within the window from spotting the opportunity to fully establishing the position, the rate may already have changed. Fourth, the cost of moving funds: for a cross-exchange strategy this includes the gas fees of on-chain transfers, and for a CEX-DEX strategy it includes bridging fees and the potential time cost of funds in transit. Take a typical cross-exchange basis trade: assuming an expected basis return of 10 basis points, two-sided fees of about 10 basis points, and two-sided slippage of about 4 basis points, the net return is only about -4 basis points. This explains why a small basis is actually unprofitable once all costs are counted, and it provides a microfoundation for the existence of the limits of arbitrage.

Latency risk arises from the time it takes from spotting an arbitrage opportunity to having the order filled in the exchange's matching engine. Though measured in milliseconds or even microseconds, this interval is, in a high-frequency environment, long enough to allow significant price movement. It spans several stages: signal processing, algorithmic decision-making, network transmission, and internal processing at the exchange (see Figure 17-13). In a market where prices change from instant to instant, any delay can cause the opportunity to vanish before the order arrives. Aquilina, Budish, and O'Neill (2022) [16] showed that, in traditional financial markets, the latency-arbitrage race has reached the microsecond scale. Although crypto markets still lag in overall speed, the core logic is the same: execution speed is the decisive variable in a strategy's success or failure.

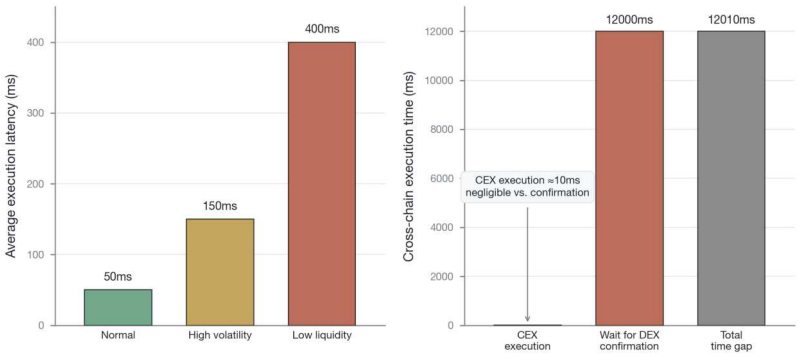


Figure 17-13. Comparison of latency across exchanges and networks (Data source: drawn by the author; the latencies are representative illustrative values, not empirical data)

As the figure shows, execution latency is closely tied to the market environment: under normal conditions, average latency stays in the tens of milliseconds, but when the market enters high volatility or liquidity depletion, the matching engine and network channels come under strain, and latency spikes to hundreds of milliseconds. This both magnifies the risk of price movement during execution and directly worsens slippage, forming a vicious cycle.

Partial-fill risk is one of the most destructive situations in a multi-leg strategy. When an arbitrageur tries to execute two or more legs simultaneously, the order on one side may fill completely while the order on the other fills only partially or fails outright. This turns a risk-hedged, delta-neutral position into a naked, one-sided exposure in an instant. In a basis trade, for example, if the short perpetual leg fills successfully but the spot-buy leg fails amid violent price swings, the arbitrageur is directly exposed to a fall in the asset's price, and the position's character shifts from arbitrage to unprotected directional speculation.

Together, these three risks form the core of inter-leg risk. Quantifying it requires precisely measuring the average time gap and price deviation between the fills of the legs under different market states. For a CEX-DEX arbitrage strategy, execution on the CEX side may complete within a few milliseconds, but execution on the DEX side is bound by the block-confirmation time of the target blockchain and may take several seconds or longer. This structural time asymmetry exposes the arbitrageur to enormous uncertainty, leaving them fully exposed to price movements in the CEX market throughout the long window of waiting for the DEX transaction to confirm.

17.5.2 The coordination problem of triangular arbitrage

If the execution risk of a two-leg arbitrage is linear, the coordination problem of triangular arbitrage is markedly magnified. Section 16.6 discussed the pricing mechanism of triangular arbitrage; this section focuses further on its execution coordination. Note that the "spot–futures–perpetual triangular arbitrage" defined in Section 16.6 is a variant across the term structure, whereas this section uses currency triangular arbitrage to illustrate the coordination problem: it captures the internal inconsistency of pric-

ing by trading consecutively across three asset pairs (for instance, selling A/B, buying A/C, and selling C/B). Both are triangular arbitrage, but they span different types of boundary. The success of this strategy requires the three trades to be completed in a near-perfect timed relay, and a break in any link leads to serious adverse consequences.

The execution window for triangular arbitrage is very brief, typically lasting only milliseconds. This demands that the arbitrage system not only complete three separate trades at high speed but also monitor each leg's execution status in real time and dynamically adjust the order parameters of the last two legs based on the fill of the first. It is a highly complex coordination task (see Figure 17-14). When the system detects an opportunity among BTC/USDT, ETH/BTC, and ETH/USDT, for instance, it immediately sends the first-leg trade. Within the tens of milliseconds from placing that order to confirming its fill, however, the market may already have changed. If the fill price or quantity of the first leg deviates from expectations, the mathematical basis of the entire arbitrage chain is already shaken. The algorithm must then decide in an instant: abort the strategy and accept the small loss and exposure the first leg may bring, or execute the remaining two legs in the hope that subsequent market movement will make up the loss—which may equally lock in a larger loss.

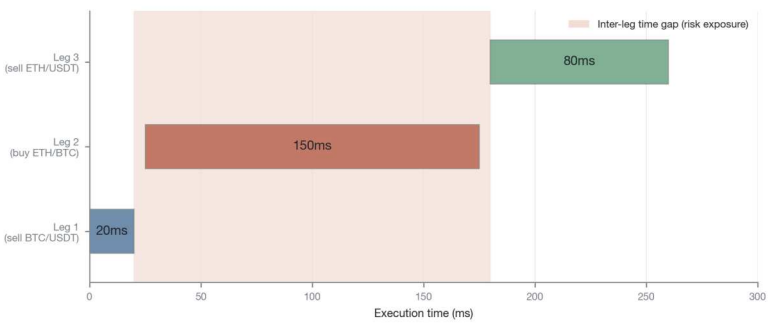


Figure 17-14. Schematic of the timing of three-leg triangular-arbitrage execution and the inter-leg risk-exposure window (Data source: drawn by the author; the execution times of the legs are illustrative values, not empirical data)

As the figure shows, throughout the window between the fill of the first leg and the fill of the third, the arbitrageur's position is incompletely hedged and exposed to market swings; the longer this window, the greater the probability that the strategy fails.

Recent academic research provides strong empirical evidence for this difficulty. A study of high-frequency data from Binance found that although theoretical triangular-arbitrage opportunities appear frequently (4,879 were identified during the study period), once real-world frictions—transaction costs, the limited depth of the order book, and execution latency—are taken into account, the net profit of these opportunities is negligible, tending toward zero [17] (see Figure 17-14). This indicates that centralized cryptocurrency exchanges are already highly efficient at the microstructural level, an efficiency shaped precisely by the fierce execution race among countless high-frequency arbitrageurs. After transaction costs and execution frictions are counted, these the-

oretically available opportunities are viable only for top participants with the lowest latency and the best order-execution algorithms; for ordinary market participants, they are unattainable in practice.

As crypto markets converge with traditional finance, this coordination problem grows more complex. An arbitrage chain involving a Bitcoin spot ETF—arbitraging among the ETF, CME futures, and perpetual futures, say—introduces a new "time-window mismatch" risk. Perpetual futures trade 24/7, whereas CME futures and ETFs have defined opening and closing hours. This mismatch in trading sessions means the arbitrageur cannot hedge the position at any arbitrary moment, accumulating enormous overnight risk while the traditional venues are closed—no longer a purely execution problem but a structural, institutional obstacle.

17.5.3 CEX-DEX cross-chain execution risk

CEX-DEX arbitrage (see Section 16.9) takes execution risk into an entirely new dimension, because it straddles two very different technical paradigms—centralized and decentralized. The CEX execution environment is closed, fast, and deterministic, whereas DEX execution is open, comparatively slow, and probabilistic. This structural difference is the core source of CEX-DEX arbitrage risk.

The most conspicuous risk is the vast asymmetry in execution speed. On a CEX, a trade typically confirms within milliseconds. On a DEX, the transaction must wait to be packed into a block and confirmed on-chain. This time varies by blockchain (see Figure 17-15), and two levels must be distinguished: soft confirmation (the transaction is provisionally accepted by the sequencer or validators) and hard finality (a definitive confirmation that the transaction cannot be rolled back). On Solana, a single slot is produced in roughly 400 milliseconds, but optimistic confirmation takes about 1.5 seconds and full finality (confirmation across 32 slots) about 12–13 seconds. On mainstream Layer-2 networks such as Arbitrum and Optimism, a sequencer's soft confirmation can complete within seconds, but withdrawing assets to L1 requires a challenge period as long as 7 days (optimistic rollups) or a proof-verification period of several hours (ZK-rollups). For a CEX-DEX arbitrageur, the key decision is which level of confirmation to accept: waiting only for soft confirmation shortens the execution window but bears the risk that a block reorganization rolls the transaction back, whereas waiting for hard finality provides certainty but greatly lengthens the window of risk exposure [18]. This means that after the arbitrageur completes one leg on a CEX, they must bear the market exposure of the CEX-side position for seconds—or even a dozen-odd seconds—while awaiting on-chain confirmation of the DEX-side transaction.

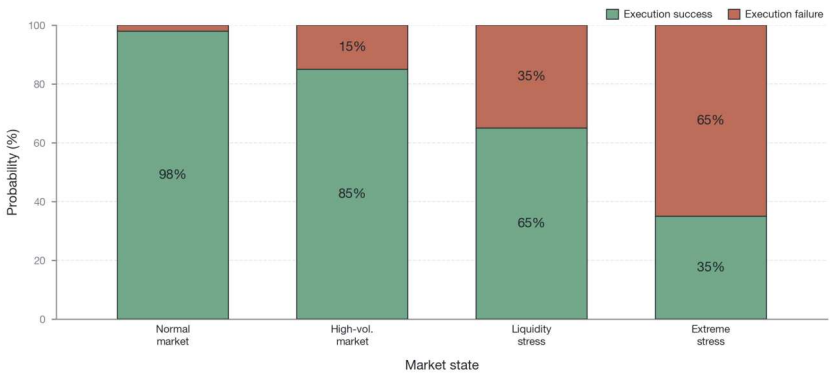


Figure 17-15. The execution success rate of CEX-DEX trades under different market conditions (Data source: drawn by the author; the success rates are illustrative values assumed by the author, not empirical; for mechanism background, see Daian et al. [18])

Worse still, the success of a DEX transaction is not guaranteed: as the figure shows, its execution success rate falls markedly as the market state deteriorates (during high volatility or network congestion, gas prices spike and competition for block space raises the probability of failure). A modeling study of Ethereum CEX-DEX arbitrage shows that the probability of a DEX transaction executing successfully is the key decision variable determining whether an arbitrageur enters [18] (see Figure 17-15). A higher risk of failure systematically deters arbitrageurs, causing the CEX-DEX spread to widen persistently and market efficiency to decline.

Beyond latency and failure risk, the DEX environment introduces a distinctive MEV risk. Because of the blockchain's transparency, all pending transactions—sitting in the mempool—are publicly visible before confirmation. This creates opportunities for professional MEV participants known as *searchers*. When a CEX-DEX arbitrageur's transaction appears in the mempool, a searcher can immediately recognize the profit opportunity and *front-run* it by paying a higher gas fee, completing the arbitrage ahead of the original arbitrageur and causing the latter's transaction to fail as prices change. Alternatively, they can mount a more sophisticated *sandwich attack*, placing one trade in each direction immediately before and after the arbitrageur's transaction to manipulate the price and drain all the profit from the original trade. In this environment, the CEX-DEX arbitrageur is racing not only against the market price but also against MEV searchers in a competitive game under information asymmetry. Recent empirical research jointly conducted by researchers at Flashbots, the Ethereum Foundation, and others [19] shows that CEX-DEX arbitrage remains one of the largest sources of MEV on Ethereum to this day—its transactions occupy less than 2% of block space yet contribute more than 15% of block value—which is ample evidence of how intense the competition is.

Beyond these risks, the volatility of gas fees adds enormous cost uncertainty to CEX-DEX arbitrage. In active markets, the gas price on Ethereum can multiply several times over within minutes. A trade that was profitable can turn into a loss because of an unexpectedly high gas fee. Although a flash loan can achieve atomic arbitrage within a single transaction on a single blockchain—so that all legs either succeed together or

fail and revert together, eliminating inter-leg risk—this mechanism cannot be applied to cross-chain or CEX-DEX settings. As soon as funds must cross a trust boundary, atomicity ceases to exist, and execution risk necessarily appears.

17.5.4 High-frequency competition and technical risk

Along the execution-and-timing dimension, competition takes its ultimate form in an endless *arms race* over speed. High-frequency trading is a class of strategy that relies on a sub-millisecond speed advantage to capture tiny price spreads. For high-frequency trading firms, a lead of even a millionth of a second can mean millions of dollars in annualized returns. This competition drives the continual iteration of trading technology, but it also markedly raises the barrier to entry for arbitrage and gives rise to new, systemic technical risks.

The core of this arms race is cutting latency. To gain a physical edge, high-frequency trading firms spare no expense. A typical strategy is *colocation*—placing one's own trading servers in the same physical data center as the exchange's servers to minimize network latency (see Figure 17-16). In traditional financial markets, this service is costly. In crypto markets, though the structure differs, server hosting near core exchanges is likewise a scarce resource. Going further, to connect exchanges in different geographic locations (Chicago and New Jersey, say), high-frequency trading firms invest in dedicated microwave transmission networks, because microwaves propagate through air about 40% faster than light through fiber. The construction and maintenance of such infrastructure routinely costs hundreds of thousands, or even millions, of dollars a year [20].

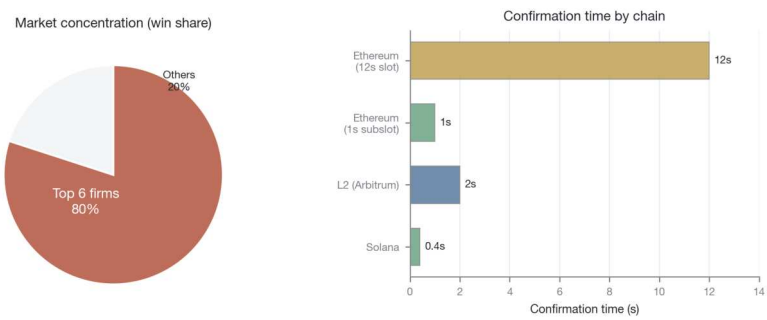


Figure 17-16. Market concentration in the latency-arbitrage race and confirmation times across blockchains (left: the top six firms account for more than 80% of race wins and losses, per Aquilina, Budish, and O'Neill [16]; right: confirmation times per chain are public parameters, compiled by the author)

Algorithmic optimization is another key to the race: high-frequency firms burn their trading logic into hardware chips such as FPGAs to eliminate microsecond-scale latency. The direct consequence is a highly concentrated market—an authoritative study of the London Stock Exchange found that latency-arbitrage races occur on average once per minute per stock, are usually decided within 5–10 microseconds, and are won and lost

by the top six firms more than 80% of the time [16] (see Figure 17-16). The great majority of tiny arbitrage opportunities are thus captured by the fastest participants, leaving very limited opportunities for everyone else.

Yet the relentless pursuit of execution speed introduces a new systemic fragility: heavy dependence on complex technical systems markedly increases operational risk. When the high-frequency trading infrastructure—the whole apparatus of algorithms, proprietary networks, and API connections—suffers even a small malfunction, the consequences can be severe. A wrong algorithm parameter, an unexpected network outage, or an unplanned update to an exchange's API can lead an unmonitored system to build up an enormous erroneous position within seconds. The 2012 episode in which Knight Capital lost $440 million in 45 minutes to a faulty trading algorithm and was ultimately driven into insolvency is a classic case of this risk. In crypto markets, which trade 24/7 and are more volatile, the risk of such technical failures can only grow. For the institutions that invest the most in the speed race, the core source of risk is precisely the complex technical systems on which they depend.

Beyond algorithmic and hardware failure risk, automated arbitrage systems face a distinctive security threat: API-key compromise. Arbitrage bots need exchange API keys to execute trades automatically, and these keys are usually granted full trading permissions. A key may be stolen through an accidental public commit to a code repository, a phishing attack, or a supply-chain vulnerability in a third-party tool. Once a key falls into an attacker's hands, they can use the arbitrageur's account to execute malicious trades—buying a low-liquidity token at an extreme price, for example, to complement a counterparty position of their own—causing not only direct financial loss but also, through the anomalous trading pattern, potentially tripping the exchange's risk controls and freezing the account. In 2023, several incidents of bulk account theft caused by key leaks at third-party service providers together produced losses of tens of millions of dollars. For professional arbitrageurs who rely on automated trading across multiple platforms, API-key management—including tiered permissions, IP-whitelist binding, withdrawal limits, and key-rotation mechanisms—has become an indispensable foundation of operational risk management.

In sum, execution and timing risk is the microstructural constraint on arbitrage in perpetual futures markets. It turns a theoretically risk-free return into a comprehensive test of technical capability, capital scale, and risk management. Small differences in execution latency directly determine the feasibility of an arbitrage strategy and ultimately define the practical upper bound on the efficiency the market can attain.

17.6 Liquidity depletion

The *funding-liquidity spiral* model of Brunnermeier and Pedersen (2009) [8] provides an important theoretical foundation for understanding the liquidity depletion examined in this section. The model holds that market liquidity and funding liquidity are locked in a mutually reinforcing positive-feedback relationship: when market liquidity declines, asset-price volatility rises and margin requirements increase, which in turn tightens funding liquidity, forces traders to cut positions, and ultimately worsens market liquid-

ity further.

In the complex ecosystem of perpetual futures markets, liquidity is often treated as an exogenous, given background parameter. This view overlooks a fundamental fact: liquidity is not naturally occurring but is, to a large extent, an endogenous variable, shaped jointly by the collective behavior of market participants. Here arbitrageurs play a contradictory dual role—providing market depth and integrating fragmented liquidity pools into an interconnected network when the market runs smoothly, yet withdrawing en masse under extreme stress (this endogenous mechanism is detailed in Section 17.6.1). This is exactly how the liquidity-depletion dimension (Dimension V) behaves under extreme conditions, and this chapter calls its extreme form the *liquidity black hole*: an endogenous, systemic risk that is set off by arbitrageurs' own behavior and ultimately turns severely against them.

This section examines the internal mechanism of this risk dimension. We first explore the endogeneity of liquidity, revealing the core contradiction whereby arbitrageurs act simultaneously as suppliers of and demanders for liquidity. We then dissect the *stampede* mechanism, explaining how, under the pressure of collective position-closing, the market falls into a self-reinforcing positive-feedback loop. Next we show the vast differences in liquidity across assets—*liquidity stratification*—and how this stratification creates a discontinuity in arbitrage feasibility between mainstream coins and altcoins. Finally, through a review of crisis moments, we quantify the severe consequences of liquidity evaporation and a sharply widening basis, revealing the predicament in which arbitrageurs, when they need to exit, face a severe shortage of counterparties.

17.6.1 The endogeneity of liquidity

As Chapter 16 argued, through their cross-market, cross-product trading, arbitrageurs objectively perform an infrastructural function: integrating fragmented liquidity and adding to order-book depth.

In analyzing the endogeneity of liquidity, it is worth distinguishing the different contributions of two types of participant. Market makers directly constitute order-book depth through continuous two-sided quoting, and their withdrawal immediately widens spreads and collapses depth; arbitrageurs contribute liquidity indirectly through cross-market directional trading, and their withdrawal shows up more as a decline in the efficiency of price transmission across markets. In a crisis, market makers, facing heightened inventory risk and adverse selection, tend to pull their quotes first, while arbitrageurs are forced to close positions under margin pressure. Their synchronized exit strikes the order book from both the supply side (the market makers) and the demand side (the market orders generated by arbitrageurs closing out), producing the *double-edged effect* of liquidity depletion.

Yet the liquidity created by arbitrage carries a fundamental fragility: it is conditional and precarious. It springs not from a long-run conviction about an asset's future value but merely from the existence of a profitable spread. This means that once the economic incentive to hold the arbitrage strategy disappears, or even turns negative, this liquidity withdraws in an instant, far faster than a long-term investor acting on fundamental analysis. When the funding rate turns from positive to negative, the basis

arbitrageur no longer has an incentive to hold the short; when a sharp rise in volatility erodes the thin margins of statistical arbitrage, they quickly close out and exit. At that moment, the arbitrageur completes the shift from a *supplier* of liquidity to a *demander* of it.

A basis arbitrageur trying to close a short perpetual position needs a willing buyer to appear in the market. An arbitrageur trying to unwind a cross-exchange position needs to find counterparties in both markets at once to complete the exit. Under normal conditions this is usually not a problem, because there are always other types of participant—speculators, long-term investors—willing to take the other side. The crux is that when the market conditions that trigger a collective arbitrageur exit arise (a systemic price crash or a violent rate reversal, say), these external liquidity providers also tend to seek safety, pull their orders, and even join the selling. This creates a grave situation: at the very moment arbitrageurs most need liquidity to exit their positions, external liquidity is drying up, while internal liquidity—the other arbitrageurs—far from helping, becomes a competitor for the limited exit resources.

This endogeneity means that the order-book depth we see on an exchange's interface—the *apparent liquidity*—is largely an illusion of abundance. It contains a great deal of highly unstable, conditional liquidity contributed by arbitrageurs. Once the market environment deteriorates, this liquidity withdraws quickly, exposing the market's true *native liquidity*, which is far shallower than the surface suggests. The phenomenon can be understood as a *tragedy of the commons* in financial markets. For each individual arbitrageur, the most rational choice when a risk signal appears is to close out immediately and preserve capital. When all arbitrageurs act on the same logic, however, their collective behavior converges into a vast, one-directional wave of position-closing that drains market liquidity in an instant, causing slippage to surge and transaction costs to soar. Each individual's rational choice ends in a collective, irrational disaster: the market falls into a liquidity vacuum in which no one can exit at a reasonable price. Understanding the endogeneity of liquidity is a key part of understanding systemic risk in perpetual futures markets, for it reveals why, under extreme stress, the market can shift rapidly from an efficiently functioning state to a dysfunctional, seized-up, and inefficient one.

17.6.2 The stampede mechanism

The endogeneity of liquidity provides the theoretical basis for understanding market fragility, and the *stampede* mechanism is that theory's dynamic, destructive manifestation in the real world. It describes how, when many arbitrageurs try to close positions at once because of a common trigger, the market falls into a strong positive-feedback loop that ultimately leads to the complete evaporation of liquidity and a violent decoupling of prices. This process is not a simple linear sum but a self-reinforcing, accelerating, nonlinear collapse.

A typical trigger scenario often begins with a seemingly harmless market change. After a period of bull market, say, sentiment abruptly reverses and prices begin to fall noticeably. This decline first triggers margin-and-liquidation risk (the margin-and-forced-liquidation dimension), forcing some highly leveraged arbitrageurs to cut posi-

tions. More important, sustained panic selling drives the perpetual futures funding rate quickly from a stable, small positive value to a large negative one (the funding-rate-reversal dimension). For the large population of basis arbitrageurs, this is a decisive signal. The core profit source of their "buy spot, short perpetual" strategy is collecting a positive funding rate. When the rate turns negative, their position shifts from one that generates a positive return each day to one that steadily consumes capital. Closing out is now no longer an option for risk management but a necessary act to keep the strategy alive.

The first and most sensitive arbitrageurs therefore begin to act. They must complete two operations at once: sell the spot and buy (close) the short perpetual. Under normal conditions, these two operations offset one another and have limited net effect on the market price. In an environment of collective panic, the situation is entirely different. A flood of spot sell orders pours into an already fragile market and pushes the spot price lower still. At the same time, a flood of perpetual futures buy orders (the closing flow) tries to push the perpetual price up at the other end. The combined force of the two drives the perpetual price from a positive premium to a negative one relative to spot—a sharp widening of the basis. A basis that was only -10 basis points may widen to -100 or even -500 basis points within a few hours.

The sharp deterioration of the basis deals a second blow to the other arbitrageurs still in the market. Although their positions are mathematically delta-neutral, the enormous negative basis means their portfolios show huge unrealized losses. Worse, these losses directly strike their margin accounts and intensify the liquidation pressure already present. Few choices remain: either cut losses immediately, accept the loss incurred, and join the closing crowd, or post enormous additional margin in the hope that the basis will eventually converge. In an environment of severely deficient market confidence, the great majority choose the former. A second and a third, larger wave of closing floods follow in succession, pushing the spot price down further and dragging the perpetual price up, widening the basis more sharply still. A strong positive-feedback loop is thereby formed.

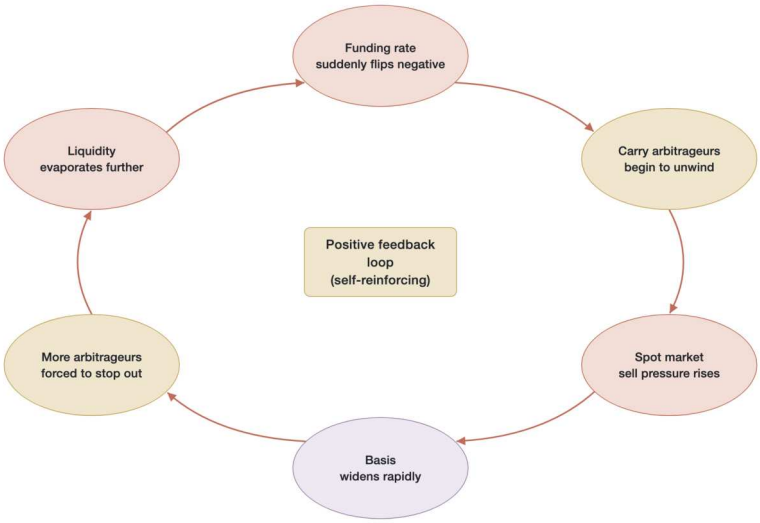


Figure 17-17. Schematic of the positive-feedback-loop mechanism by which funding rate reversal triggers an arbitrageur stampede (Data source: mechanism schematic drawn by the author, not empirical data)

Figure 17-17 depicts the arbitrageur stampede triggered by funding rate reversal as a positive-feedback loop: the sudden flip to a negative rate forces carry arbitrageurs to close out, spot selling pressure mounts and the basis widens rapidly, and the wider negative basis in turn forces more arbitrageurs to cut losses and exit, evaporating liquidity further—the loop reinforcing itself. It is this loop that turns a collective exit in a crisis into an unstoppable stampede, so that the actual cost of closing out far exceeds any estimate based on normal-market data.

This process is a concentrated expression of the risk of *crowded trades* in financial markets. In his American Finance Association presidential address, Stein (2009) [21] rigorously argued that even a market composed of fully rational, well-informed arbitrageurs can become overcrowded because participants cannot observe one another's position sizes: each individual's optimal decision produces systemic fragility at the collective level. Khandani and Lo's (2011) [22] empirical analysis of the quant-fund stampede of August 2007 showed further that when large numbers of traders using similar strategies are forced to cut positions within the same time window, their collective behavior can drain liquidity within hours, producing market dynamics highly isomorphic to the crypto-market stampede described in this section. When too much capital chases the same seemingly risk-free strategy, the strategy's exit path itself becomes the largest exposure. Large numbers of participants trying to exit the same strategy at once cause exit costs to rise sharply. In crypto markets, this effect is magnified further by the scale of protocol-level arbitrage vehicles. When a protocol manages billions or even tens of billions of dollars of basis-arbitrage positions, any decision it makes to close out in a crisis is enough to sway the entire market decisively. If Ethena were forced into a large-scale unwind to cope with negative funding and user redemptions, the selling volume of this single entity might exceed what the market can absorb within any reasonable

slippage. This concentration risk means that a protocol's self-rescue can turn into a market-wide stampede, imposing severe negative externalities on other small and mid-size arbitrageurs. In the end, as all arbitrageurs rush for the exit together, the market exhausts its last trace of liquidity, order-book depth contracts sharply, and spreads widen to extremes.

17.6.3 Liquidity stratification

The risk of a liquidity black hole is not evenly distributed across all crypto assets. On the contrary, it exhibits a pronounced layered structure, which this chapter calls *liquidity stratification*. This phenomenon is especially prominent in perpetual futures markets, where it carves a deep gulf between the mainstream coins (Bitcoin, Ethereum) and the broad universe of altcoins, directly determining the feasibility and risk exposure of arbitrage strategies across assets. For an arbitrageur, choosing which asset to deploy a strategy on matters far more than choosing the strategy itself, because an asset's liquidity tier fundamentally defines the upper bound on its risk.

At the top of the market sit Bitcoin and Ethereum. Their perpetual futures markets have the highest level of liquidity in the world, with daily volumes routinely reaching tens or even hundreds of billions of dollars. Vast volume, deep order books, and extremely narrow bid-ask spreads together make for a relatively robust trading environment. This means that even in extreme periods of violent volatility, liquidity, though it declines markedly, usually does not evaporate entirely. An arbitrageur managing a multimillion-dollar position can still, after paying some slippage, find enough counterparties to complete an exit. This provides a basic cushion for arbitrage, letting the various strategies applied to BTC and ETH (basis arbitrage, cross-exchange arbitrage) operate effectively most of the time and keeping their market efficiency high.

One tier down are the mainstream altcoins with high market capitalization and strong consensus, such as SOL and DOGE. Although the perpetual futures markets of these assets are far smaller than those of BTC and ETH, they usually still have daily volumes in the billions of dollars and reasonable market depth. When the market is calm (see Figure 17-18), arbitrage on these assets is feasible, though its potential slippage costs and execution risk are markedly higher than for the top assets. The liquidity buffer of these markets is significantly thinner, however. Once the market enters the stressed state, their liquidity contracts far faster than that of BTC and ETH, spreads widen quickly, and the impact cost of large orders rises sharply. For arbitrageurs, this means deploying a strategy on these assets requires more conservative position sizing and more sensitive risk monitoring.

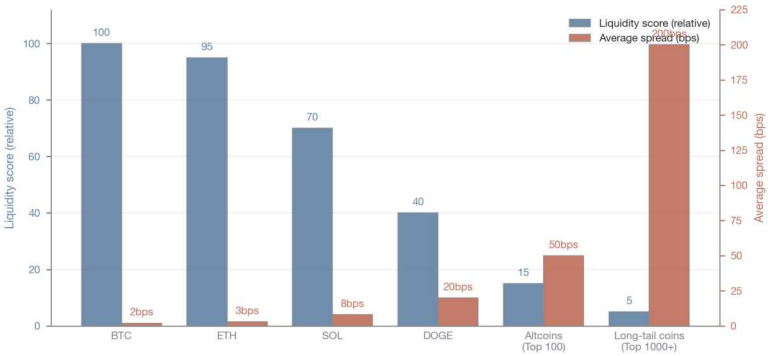


Figure 17-18. The liquidity-stratification structure and arbitrage feasibility of different asset classes (Data source: drawn by the author; the tier scores are relative illustrative values, not empirical data)

The true discontinuity appears among the lower-tier altcoins and long-tail assets. For a typical altcoin ranked outside the top 100 by market capitalization, perpetual futures daily volume may be a mere few million dollars or even less. Order-book depth is severely lacking, and the gap between the best bid and best ask may exceed 50 basis points. In such a market, even arbitrage under normal conditions faces significant difficulty. A trade of a few tens of thousands of dollars can move the price noticeably, so that the theoretical arbitrage margin is entirely consumed by slippage at the execution stage. More dangerous still, a thin order book creates the conditions for price manipulation: a large holder can distort the mark price by concentrated buying or selling and trigger unexpected liquidations of other traders. For arbitrageurs seeking opportunities in such markets, the risk is not only insufficient liquidity but also the ever-present uncertainty of being deliberately targeted for manipulation. In a crisis, liquidity in these markets falls almost instantly to zero, trading seizes up entirely, and any attempt to close out can scarcely obtain a reasonable fill. This directly produces a vast gradient in the efficiency spectrum along the "asset dimension": the mainstream-coin market is relatively efficient, while the long-tail-asset market is full of persistent arbitrage opportunities that cannot be eliminated because of the liquidity constraint.

This liquidity stratification is an important complement to the Arbitrage Infrastructure Hypothesis. It reveals that the coverage of arbitrageurs' "infrastructure" is limited. They tend to build and maintain it in the most liquid "core regions," where the risk-return ratio is highest. In the vast, illiquid "periphery," the absence of arbitrage leaves these markets chronically inefficient and fragile. When assessing the risk of an arbitrage strategy, therefore, one must consider it within the liquidity tier of its target asset. A strategy that runs robustly on BTC, if simply copied onto some altcoin, may well fail when it meets a liquidity black hole.

17.6.4 Liquidity evaporation in a crisis

If liquidity stratification is the static structure of risk, then in a market crisis we can observe the devastating process by which this risk structure is dynamically activated. Several major crises in history—the "Black Thursday" of March 2020, the "May 19

crash" of 2021, and the collapses of LUNA and FTX in 2022—all provide excellent cases for quantifying how liquidity evaporates into thin air and how the basis dislocates sharply. These historical data are not merely a confirmation of theory but the ultimate stress test of every arbitrageur's risk-management capability.

In the opening stage of a crisis, the most visible change in the indicators is the sharp shrinkage of order-book depth. Take the BTC perpetual futures on a major exchange: under normal conditions, the cumulative resting orders within 1% above and below the midpoint may total hundreds of millions of dollars, providing an ample buffer. As Figure 17-19 shows, however, when systemic panic spreads, large numbers of market makers and passive liquidity providers pull their orders and seek safety. The order book thins rapidly, the once-continuous ladder of quotes becomes fragmentary, and huge "vacuum zones" appear between adjacent price levels. Liquidity evaporation means that a market order of ordinarily moderate size may now punch through several price levels, causing enormous market impact and slippage. For an arbitrageur anxious to close out, this means their exit cost grows exponentially.

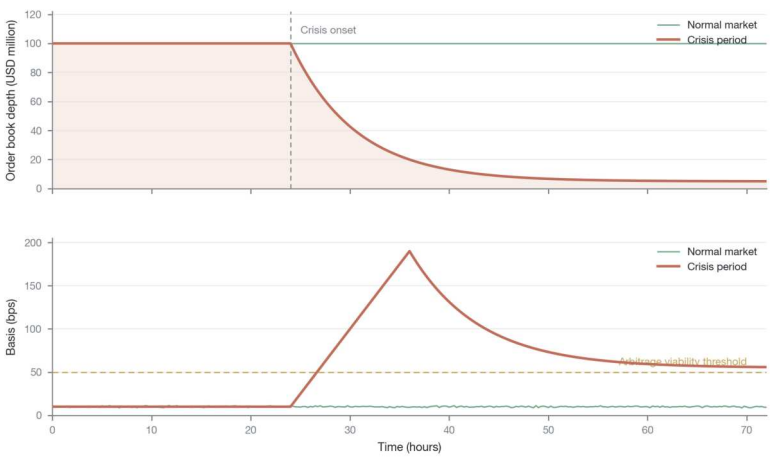


Figure 17-19. Schematic of the cascade reaction triggered by mass liquidations (Data source: schematic drawn by the author, not empirical data)

Figure 17-19 contrasts the evolution of liquidity in normal and crisis periods as a time series: after the crisis erupts, order-book depth collapses to an extremely low level within hours, while the basis crosses the boundary of arbitrage feasibility and spikes to an extreme high before slowly receding. The synchronized deterioration of the two shows that, at the very moment arbitrageurs most need to exit, market depth and executability collapse together, and the cost of exit grows exponentially.

Accompanying liquidity evaporation is a sharp widening of the basis. The stampede mechanism causes the spot and perpetual futures prices to diverge violently. At the peak of a crisis, the discount (negative basis) of perpetual futures relative to spot can reach an extreme degree. In a normal market, a basis of -50 basis points (-0.5%) already counts as a significant arbitrage opportunity, but in the extreme "Black Thursday"

conditions of March 2020, the XBTUSD perpetual futures on BitMEX traded at a discount of about −12%, or even deeper, relative to spot (the exact instantaneous extreme varies by data source). Notably, this extreme discount was not driven entirely by market forces: on March 13, immediately afterward, BitMEX halted trading for about 25 minutes because of a distributed denial-of-service attack, and the interruption led market makers to pull orders heavily once the platform resumed, causing order-book depth to plunge. This extreme discount was therefore, to a large extent, the product of a resonance between liquidity depletion (Dimension V) and platform risk (Dimension III)—a representative case consistent with this chapter's risk-resonance mechanism (it should be noted that the DDoS outage was an idiosyncratic exogenous factor specific to this event, and the generality of the model still rests on the earlier multi-case comparison of "different starting points, same terminal state"). A dislocation of this magnitude far exceeds what any basis-arbitrage strategy can withstand. The portfolio value of every arbitrageur holding a "buy spot, short perpetual" position suffered enormous losses within a very short time, directly triggering a large wave of forced liquidations. These liquidation orders further intensified selling pressure, formed a vicious cycle, and ultimately caused the market's pricing mechanism to fail severely, if temporarily. This offers a compelling explanation of the direct cause of the *basis anomaly* examined in Chapter 18 (see Section 18.2): it was not that arbitrageurs failed to see the opportunity, but that, in the face of a liquidity black hole, they could scarcely execute the arbitrage at a reasonable price.

In a liquidity black hole, arbitrageurs of different sizes behave in markedly divergent ways. Large, well-capitalized arbitrage institutions—professional market makers and high-frequency trading firms—by virtue of their greater risk tolerance and more advanced execution systems, can often recognize the risk signals early in a crisis (see Figure 17-20) and begin to reduce their positions in an orderly, staged manner. They exit gradually, splitting large orders into many small ones and using algorithms to execute at different points in time so as to minimize market impact. Small and mid-size arbitrageurs and retail investors, by contrast, tend to react more slowly and have weaker risk management. By the time they finally recognize the danger and begin to close out in panic, part of the market's liquidity has already been consumed by the exit of the large institutions, and they face worse prices and higher slippage. The smallest participants—those on high leverage whose margin is already near the liquidation threshold—often do not even get the chance to close out voluntarily before the exchange's liquidation engine forcibly closes them out. This divergence in exit paths further deepens the market's unfairness and ultimately shapes the evolution of the liquidity black hole: from orderly retreat, to panic selling, to a final liquidation cascade.

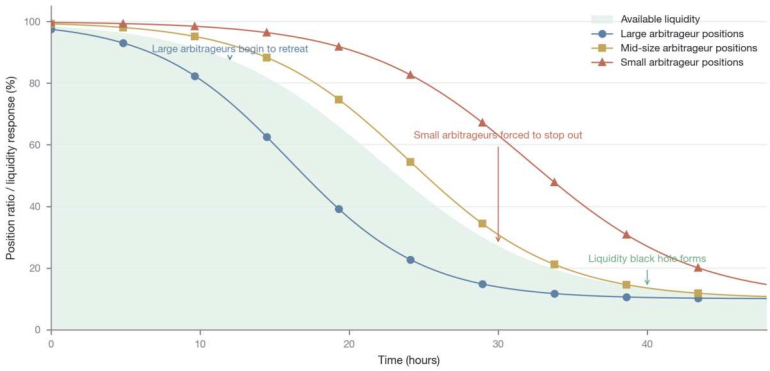


Figure 17-20. Divergence in the exit behavior of arbitrageurs of different sizes during a crisis (Data source: schematic drawn by the author, not empirical data)

In sum, the liquidity black hole is a concentrated expression of the endogenous fragility of perpetual futures markets. It is rooted in the conditional nature of the liquidity arbitrageurs provide, triggered through the *stampede* mechanism of collective position-closing, and differentiated in its destructive power by the liquidity stratification across assets. In a crisis, liquidity evaporation and a sharply widening basis together constitute systemic pressure on arbitrageurs, turning a theoretically "risk-free" strategy into an exposure that produces severe losses. For anyone hoping to survive over the long run in perpetual futures markets, a deep understanding and respect for this constraint dimension is an indispensable first lesson.

17.7 Smart contract and oracle risk

When arbitrage extends to decentralized exchanges, the risk topology undergoes a structural shift. In a CEX environment, technical risk shows up mainly as recoverable failures such as API latency or system downtime; in the DeFi ecosystem, smart contract and oracle risk rises to become the central concern. Once a vulnerability or design flaw in the code is exploited, the consequences are immediate, destructive, and often irreversible.

For an arbitrageur who relies on a DEX, the safety of their position hangs entirely on the robustness of the underlying smart contract, the accuracy of the oracle's price feed, and the integrity of the cross-chain infrastructure. Unlike in a CEX environment, once technical risk in DeFi materializes it often leads directly to a permanent loss of assets. This section systematically analyzes the internal mechanism of this constraint dimension along four aspects: smart-contract vulnerabilities, oracle manipulation, cross-chain bridge security risk, and the uncertainty of protocol governance.

17.7.1 Smart-contract vulnerabilities

The smart contract is the cornerstone of a DEX perpetual futures protocol, defining in code all the core logic—trading, liquidation, funding-rate calculation, and margin

management. Yet any small vulnerability in these complex contracts can become the breach through which an attacker drains the protocol's funds (see Figure 17-21). History has proved repeatedly that even well-known protocols audited over many rounds can scarcely be wholly immune to logical flaws or novel attack techniques (although some cases achieve partial recovery through a protocol pause, a white-hat return, or an on-chain freeze via tracing, the losses are in most cases irrecoverable). According to a tally of reproduced exploits (DeFiHackLabs), in 2024 alone there were more than 150 contract-attack incidents caused directly by smart-contract vulnerabilities of various kinds, producing economic losses of more than $328 million.

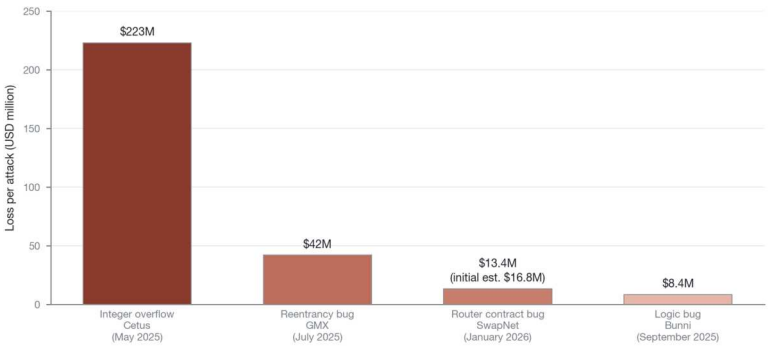


Figure 17-21. Comparison of single-incident losses from smart-contract attacks on major DEX protocols, 2025–2026 (for GMX, see [23]; for Bunni, see [24]; for SwapNet, see [25]; Cetus and others from reviews of public incidents; source: compiled by the author)

For arbitrageurs, this risk is concrete and severe. When a DEX perpetual futures protocol is attacked through a vulnerability, the most immediate consequence is the theft of the protocol's liquidity pool or treasury. The July 2025 attack on GMX, for example, caused asset losses of more than $42 million [23], while the September 2025 attack on Bunni DEX directly forced the protocol to shut down permanently, unable to cover an $8.4 million shortfall [24]. In such cases, a position an arbitrageur holds on that DEX—long or short—can become worthless or impossible to close in an instant. If that position is part of a cross-market arbitrage strategy (a CEX-DEX basis trade, say), then the sudden break of its "leg" on the DEX turns the hedge on the CEX side, in an instant, into a fully exposed, high-risk naked position. The arbitrageur not only loses their principal on the DEX but must urgently manage the exposure at the other end—which, in a violently moving market, easily triggers a second loss.

Attacks take many forms—from using a flash loan to manipulate prices and trigger abnormal liquidations, to finding a calculation error in the contract logic and opening a position at zero cost, to bypassing the margin-check mechanism to withdraw excess funds. In January 2026, for example, the DEX aggregator Matcha Meta was attacked through a vulnerability in the SwapNet routing contract it had integrated; according to the official post-mortem, the actual loss was about $13.4 million [25] (an initial on-chain estimate had run as high as $16.8 million but was later revised down after verification). These events highlight a stark reality: when an arbitrageur interacts with

a DEX protocol, they are in effect passively trusting thousands of lines of code written by anonymous or semi-anonymous developers. This trust is fragile: once a flaw in the code is exploited, the arbitrageur not only fails to realize the expected profit but may face the extreme risk of losing all of their principal.

For a trader deploying an arbitrage strategy across both a CEX and a DEX, a smart-contract vulnerability on the DEX side translates directly into naked directional exposure on the CEX side: the sudden disappearance of the DEX leg means the CEX hedge instantly loses its protection.

17.7.2 Oracle manipulation

Oracle risk is the core vulnerable link connecting a DEX protocol to the outside world (see Figure 17-22). DEX perpetual futures need an oracle to obtain the reference price of a trading pair on external markets (usually the major CEXs), and this price is the key basis for computing the mark price, determining the funding rate, and triggering forced liquidation. The accuracy and manipulation-resistance of the oracle directly determine the fairness and security of the protocol. Yet the oracle has itself become a core attack vector, and its manipulation generates "false arbitrage signals" that lure arbitrageurs into a carefully designed trap.

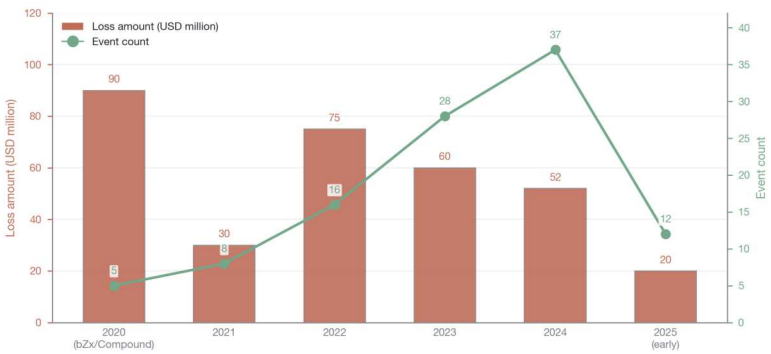


Figure 17-22. The historical evolution of oracle-manipulation attacks: loss amounts and incident counts, 2020–2025 (Data source: drawn by the author; the $52 million loss and 37 incidents in 2024 [26] and the $89 million Compound loss in 2020 [27] are empirical anchor points, while the other years are representative illustrative trend values, not year-by-year empirical data)

The typical technique of oracle manipulation is for an attacker to use a flash loan to obtain a huge sum of temporary capital within a single on-chain transaction, then execute a large swap in a shallow DEX liquidity pool, instantly distorting the pool's asset ratio and quoted price. If the target perpetual futures protocol's oracle relies on that pool's time-weighted average price as its reference source, the manipulated price is transmitted into the protocol. This produces two destructive consequences. First, it can cause the positions of many innocent users within the protocol to be liquidated erroneously as the mark price swings violently, with the attacker profiting as the liquidator. Second, it creates a huge but entirely false CEX-DEX spread. Automated arbitrage programs

immediately seize on this "golden opportunity" and rush into the market to close the spread. But by the time the arbitrageur's trade fills on the DEX, the attacker has long since traded in the opposite direction on the CEX to return the price to normal. In the end, the arbitrageur's DEX trade fills at a heavily manipulated price while the hedge at the other end can only execute at the normal price, producing a substantial loss—and that loss is precisely the attacker's source of profit.

Since bZx suffered the first large-scale flash-loan oracle attack in 2020, such incidents have recurred without pause. Historical data show that oracle manipulation has become one of the leading attack vectors in DeFi, causing losses of as much as $52 million in 2024 [26]. Even a top-tier DeFi protocol such as Compound suffered nearly $89 million in abnormal liquidations in 2020 because of a problem with its oracle price source [27]. This risk dimension reveals a paradox of CEX-DEX arbitrage: arbitrageurs are supposed to be the correctors of market prices, yet in a scenario of oracle manipulation they become the ones harvested by the distorted price. The existence of this risk forces CEX-DEX arbitrageurs to build a careful analysis of the oracle mechanism into their strategy and to set aside a risk premium for the possibility of such "signal poisoning."

From the perspective of a CEX arbitrageur, the false spread signal produced by oracle manipulation directly threatens the effectiveness of a cross-market strategy: a position built on a contaminated price signal will produce an unhedgeable loss once the price returns to its true level.

17.7.3 Cross-chain bridge security risk

As the DeFi ecosystem expands across multiple chains, CEX-DEX arbitrage and purely on-chain arbitrage increasingly rely on cross-chain bridges to move assets between different blockchains. A cross-chain bridge is the critical infrastructure connecting isolated value networks (see Figure 17-23), letting an arbitrageur buy an asset on Ethereum and then quickly move it to a DEX on an emerging public chain to sell, capturing the cross-chain spread. Yet cross-chain bridges have become the most fragile and most frequently attacked infrastructure in the entire DeFi ecosystem.

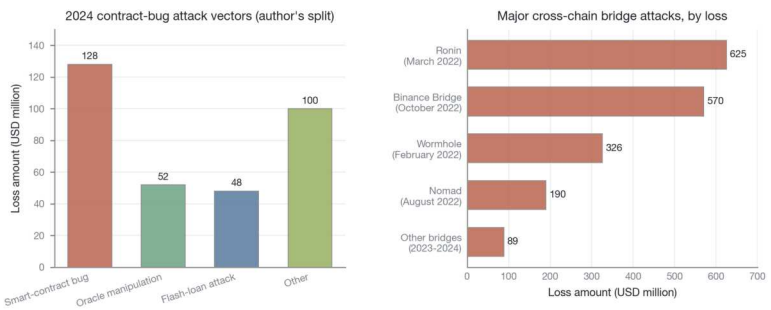


Figure 17-23. The distribution of contract-vulnerability attack vectors in 2024 (left, broken out by the author) and historical losses from major cross-chain bridge attacks (right, per FX Empire/Chainalysis [28])

The technical implementation of a cross-chain bridge is exceptionally complex, involving custody, locking, minting, and burning, and its security model is far more challenging than that of a single-chain application. Attackers can steal huge sums from a bridge by compromising validator nodes, exploiting smart-contract vulnerabilities, or stealing private keys. The data show enormous losses: according to Chainalysis figures cited by FX Empire, more than $2 billion of assets were stolen in just 13 major cross-chain bridge attacks [28], including the historic thefts of about $625 million from the Ronin Bridge and about $326 million from Wormhole. In addition, about $570 million was stolen from BNB Chain (Binance Bridge) in October 2022 (which occurred after the window of the statistics above and is therefore not counted among the 13). The losses from these events often far exceed those from an attack on a single DeFi application.

For traders who rely on cross-chain bridges for arbitrage, this risk is systemic. If a bridge is attacked while a sum of funds is moving through it from Chain A to Chain B, those in-transit funds may be frozen or even lost permanently. The arbitrage loop is thereby interrupted: the position the arbitrageur has already established on Chain A loses its hedge on Chain B and is instantly exposed to market risk. Consider a statistical-arbitrage strategy that plans to go long on a DEX on Polygon and short on a DEX on Solana: if its funds are stolen while bridging through Wormhole, the basis of the entire strategy ceases to exist. The arbitrageur must not only bear the loss of funds but also face an incomplete, risk-laden portfolio. This worry that a "main thoroughfare" might collapse at any moment markedly limits both the capital committed to cross-chain arbitrage and the appetite for it, allowing spreads between different blockchain ecosystems to persist—a major source of market segmentation and efficiency loss.

For an arbitrageur on the CEX side, cross-chain bridge risk means their arbitrage capital carries an uncontrollable window of exposure during transfer: while in transit, the funds can neither be used to hedge nor be recalled in an emergency.

17.7.4 Protocol-upgrade and governance risk

The final aspect of technical-infrastructure risk is the risk inherent in the seemingly benign processes of protocol upgrades and decentralized governance themselves. Unlike traditional software, a DeFi protocol's "upgrades" and "rule changes" are usually decided by token holders through decentralized governance votes. Although this model empowers the community, it also introduces new uncertainties and attack vectors.

One obvious risk is that a protocol upgrade may itself introduce new vulnerabilities. In fixing old problems or adding new features, developers may inadvertently create new security holes. More dangerous still, a protocol's upgrade authority—usually controlled by a multi-signature wallet—may be compromised. After the FTX collapse of November 2022, the upgrade key of Serum, the core Solana-ecosystem DEX that FTX controlled, was leaked, forcing the entire community into an emergency hard fork to prevent a malicious upgrader from seizing full control of the protocol. For arbitrageurs who relied on Serum to trade, this meant that the underlying market rules and codebase they depended on underwent an involuntary, uncertainty-laden change overnight.

A deeper hazard comes from the possibility that the decentralized governance process itself is manipulated. An attacker can borrow governance tokens in bulk on the market

to launch a malicious proposal—amending fee parameters in their own favor, whitelisting themselves to bypass certain restrictions, or even proposing outright to transfer the protocol treasury's funds to themselves. Although mechanisms now guard against such blatant attacks, subtler governance risks remain. Contentious proposals over protocol revenue distribution or the adjustment of key parameters (such as the liquidation penalty or the choice of oracle), for instance, can cause the protocol's rules to change dramatically. The fierce dispute within the Aave community over its governance structure and revenue distribution in late 2025 highlighted how internal conflict within a DAO can create uncertainty about a protocol's future direction. Because arbitrageurs' strategies are often based on the precise calculation of current protocol parameters, any unexpected rule change can render a carefully designed arbitrage model useless in an instant. This governance risk means that arbitrageurs must not only analyze code but also, like political analysts, constantly follow community developments and governance proposals—adding yet another new and hard-to-quantify constraint dimension.

For a CEX trader who builds an arbitrage position on DEX perpetual futures, governance risk means the underlying parameters of the strategy may be changed without warning, rendering a carefully calibrated hedge ratio or rate model useless in an instant.

17.7.5 Sequencer and validator risk

With the rapid rise of on-chain order-book DEXs such as Hyperliquid and dYdX v4, a new technical-infrastructure risk is emerging: centralization risk at the sequencer and validator level. Hyperliquid uses a centralized sequencer for order matching and transaction ordering, while dYdX v4 runs on a dedicated chain built on the Cosmos SDK, producing blocks by consensus among a validator set. These designs outperform traditional Ethereum-mainnet DEXs, but they also introduce a distinctive risk dimension.

A centralized sequencer faces three risks: a single point of failure that can halt the entire DEX at a critical moment; the sequencer operator's theoretical ability and incentive to front-run user trades; and the possibility that the sequencer censors particular transactions or traders. In the March 2025 episode in which Hyperliquid handled anomalous trading in the JELLY token (which produced a peak unrealized loss of about $12 million in its liquidity vault, consistent with the figure in Section 16.4.3), the validators voted collectively to delist the JELLY perpetual and forcibly liquidate the relevant positions at a non-manipulated internal settlement price, provoking a broad debate over whether a "decentralized exchange" is truly decentralized. For traders deploying an arbitrage strategy on such a platform, this means the safety of their positions depends not only on market risk and smart-contract risk but also on governance decisions at the sequencer/validator level—a risk that also exists on a CEX but is especially paradoxical in a DEX context.

For a trader who builds an arbitrage position on DEX perpetual futures, sequencer risk means that both the fairness of their trade execution and the platform's availability depend on a de facto centralized link—a significant tension with the promise of decentralization.

Smart-contract flaws, oracle dependence, cross-chain bridge fragility, governance un-

certainty, and the centralization risk of sequencers and validators together define the distinctive technical risks of DeFi arbitrage and explain the persistence of the efficiency gap between CEXs and DEXs, and between different chains.

17.8 Regulatory and compliance risk

Regulatory and compliance risk has a distinctive external character within the seven-dimensional constraint topology. Unlike constraints determined endogenously by market structure, such as margin or liquidity, regulatory risk arises from the ambiguity of legal interpretation, the exclusivity of licensing, and the suddenness of policy shifts, directly raising compliance costs and blocking particular trading routes [29].

This section analyzes, along four dimensions—cross-jurisdictional compliance costs, exchange access restrictions, the uncertainty of tax treatment, and sudden regulatory events—how this external constraint systematically limits the range over which arbitrage can be deployed and how, in extreme cases, it resonates with the other, endogenous risk dimensions.

17.8.1 Cross-jurisdictional compliance costs

In the early stages of the cryptocurrency market's development, regulatory differences across jurisdictions were regarded as an arbitrage opportunity. Arbitrageurs exploited the ambiguity in different countries' definitions of digital assets, establishing entities in lightly regulated jurisdictions to escape heavy compliance burdens. With the full implementation of the European Union's Markets in Crypto-Assets (MiCA) Regulation in 2025 and the enactment of the U.S. GENIUS Act, however, this "regulatory arbitrage" is turning into a heavy "compliance tax." A tightening global regulatory framework means that any arbitrage firm hoping to operate at scale must maintain compliance in several jurisdictions at once.

This surge in compliance costs directly weakens the marginal return of arbitrage strategies. According to industry survey data (such as Chainalysis's annual compliance report), the ratio of compliance spending to revenue at leading arbitrage firms rose markedly between 2023 and 2026. This spending covers anti-money-laundering monitoring systems, the cost of licenses in multiple countries, and complex tax-reporting infrastructure. When compliance costs exceed the expected return of a particular cross-exchange strategy, arbitrageurs are forced to abandon that route, allowing a price gap that could otherwise have been closed to persist. The phenomenon is especially clear in euro-stablecoin arbitrage: because MiCA imposes strict reserve and distribution requirements on compliant stablecoins, the euro-stablecoin basis on compliant exchanges is often markedly higher than on unregulated offshore platforms. This spread stems not from market inefficiency but from the pricing of compliance-cost risk.

What affects perpetual futures arbitrageurs more directly is the differing restrictions various jurisdictions place on derivatives leverage and retail access. Whether perpetual futures constitute a "swap" under the jurisdiction of the CFTC remains contested, and this uncertainty exposes U.S. entities to legal risk when they participate in offshore perpetual futures trading. Japan's Financial Services Agency has imposed especially

strict limits on crypto derivatives, cutting the retail leverage ceiling to 2x (a sharp reduction from the previous 25x in 2020). The regulatory frameworks of Hong Kong and Singapore define "professional investor" status for derivatives traders strictly, excluding many small and midsize arbitrage teams from these highly liquid market nodes. These differentiated access thresholds directly reduce the total capital that can participate in arbitrage in a given market, and thus structurally limit the speed at which cross-jurisdictional spreads are repaired.

17.8.2 Access restrictions and route blockage

Regulation-driven access restrictions are blocking the free-flow-of-funds routes on which arbitrage depends. To meet local regulatory requirements, many exchanges have imposed strict geofencing and identity-based access restrictions. For cross-exchange arbitrageurs, this means they may be unable to open accounts on both ends at once, or must circumvent the restrictions through complex legal structures, which further raises operational risk.

In perpetual futures especially, jurisdictions differ sharply in their limits on leverage and retail participation. The regulatory frameworks of Hong Kong and Singapore, for example, define professional status for derivatives traders strictly, excluding many small and midsize arbitrage teams from these highly liquid nodes. When a major price deviation appears between a restricted exchange and the global market, the lack of enough qualified arbitrageurs greatly slows the repair of the basis. This "access blockage" is especially severe in extreme markets, because it prevents outside capital from entering the stressed region to arbitrage in the opposite direction, thereby intensifying that region's liquidity depletion and its chain of liquidations.

17.8.3 The uncertainty of tax treatment

Tax risk is the most insidious constraint arbitrageurs face. In perpetual futures trading, the receipt and payment of funding, the settlement of mark-price gains, and the deductibility of liquidation losses are characterized entirely differently under different countries' tax laws. Many jurisdictions have yet to issue clear tax guidance for the financial innovation of perpetual futures, exposing arbitrageurs to enormous risk of retroactive taxation.

The uncertainty of tax treatment directly affects arbitrageurs' capital-allocation decisions. In some countries, perpetual futures gains may be treated as ordinary income (with a top rate exceeding 40%), while losses are subject to strict limits on deductibility. This asymmetric tax structure can turn a mathematically neutral delta-arbitrage strategy into one with a negative expected value after tax. Moreover, the cross-chain conversion of assets in cross-chain arbitrage may, under some tax laws, be treated as a "disposal event," generating an unnecessary tax burden. To avoid these potential risks, arbitrageurs often deliberately steer clear of particular assets or particular execution routes, and this tax-driven risk aversion shows up, at the macro level, as pricing distortions in particular trading pairs.

17.8.4 The shock of sudden regulatory events

Sudden regulatory events constitute an acute shock to the arbitrage system. An abrupt turn in regulatory policy—from an order to delist a coin to sanctions on a protocol—can destroy one end of an arbitrage position within a very short time.

Suppose a major jurisdiction suddenly announces a withdrawal ban on non-compliant stablecoins; the event would cause those stablecoins to trade at a severe discount on local exchanges. In this hypothetical scenario, although arbitrageurs might see an enormous buying opportunity, the severed withdrawal route would prevent funds from flowing to offshore markets to close out, and the force of arbitrage would be paralyzed in an instant. Such a regulatory shock not only creates risk but also, by altering market expectations, triggers a collective retreat of arbitrageurs. Here regulatory risk resonates violently with liquidity risk (the liquidity-depletion dimension) and counterparty risk (the counterparty-and-platform dimension), causing market efficiency to collapse completely within hours. This regulatory uncertainty is an indelible "sovereign risk premium" in arbitrage returns, and it is a key factor in explaining why the upper bound on crypto-market efficiency can never reach the level of traditional finance.

This hypothetical is not a mere thought experiment. In February 2024, the Nigerian authorities detained Binance executives and restricted the platform's local operations, causing severe price dislocations in Nigerian naira trading pairs while arbitrageurs, cut off from fiat deposit and withdrawal channels, could not step in to correct them. In the same year, SEC enforcement actions against several major exchanges led to the suspension or delisting of some trading pairs, directly blocking the arbitrage routes that relied on them. These real events confirm the immediate paralyzing effect of a sudden regulatory event on the force of arbitrage.

17.9 The risk resonance model

In a normal market, the seven constraint dimensions above can each be addressed separately through risk-management techniques. When extreme market stress appears, however, the linkages among these dimensions change qualitatively, forming a positive-feedback chain reaction that generates systemic risk far exceeding the simple sum of the individual dimensions—*risk resonance*.

Risk resonance reveals the mechanism by which market efficiency collapses nonlinearly in a crisis: when resonance occurs, the total risk of the arbitrage system amplifies in a superlinear, self-reinforcing manner, arbitrageurs' room to operate is sharply compressed, and the market's pricing mechanism seizes up. This section first uses an empirical analysis of the systemic crisis of May 2022 to show how risk resonance manifests in reality, then constructs a risk transmission matrix to identify the key transmission paths, and finally builds a theoretical model from two angles: the nonlinear amplification effect and the phase-transition critical point.

17.9.1 A resonance analysis of the May 2022 crisis

The collapse of the Terra/LUNA ecosystem in May 2022 provides a representative empirical case for observing risk resonance. In that systemic crisis, we can see clearly how the seven-dimensional arbitrage constraints described above were activated simultaneously within a mere 72 hours, amplified one another, and ultimately caused a systemic collapse in the market's arbitrage capacity.

The event was triggered by the initial depegging of the algorithmic stablecoin UST. Against the extremely fragile macro backdrop of the crypto market at the time, however, this initial disturbance quickly evolved into a full-blown crisis of confidence. As the Bitcoin price fell, a chain of reactions was set off that bound the previously independent risk dimensions tightly together.

The crisis transmission began with the activation of risk in the margin-and-forced-liquidation dimension. As the BTC price fell rapidly from the $40,000 region, the margin levels of many arbitrage positions that used BTC as collateral or as a hedging asset—especially the popular basis trade (buy BTC spot, short perpetual futures)—dropped sharply (see Figure 17-24). Every fall in price meant a shrinkage in the value of the spot side, forcing arbitrageurs to post additional margin or face forced liquidation. This directly compressed arbitrageurs' capital buffers and sharply reduced their capacity to withstand subsequent shocks.

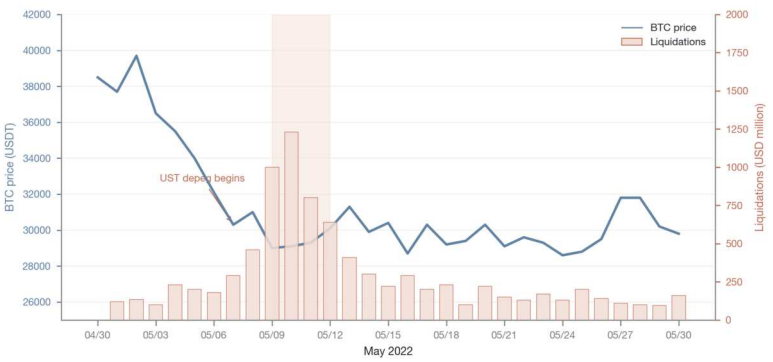


Figure 17-24. The BTC price decline and mass liquidations of May 2022 (Data source: Coinglass [9]; a representative series, not calibrated day by day)

Figure 17-24 overlays the BTC price path and market-wide liquidation volume to show their synchronized relationship. The data indicate that between May 9 and May 12, the BTC price fell sharply from the $35,000 region to around $26,000, a cumulative decline of more than 25%. At the same time, single-day market-wide perpetual futures liquidations exceeded several billion dollars on both May 10 and May 12, with short liquidations and long liquidations alternating—reflecting the violent swings in market direction and the wholesale collapse of arbitrageurs' margin.

As margin came under pressure, risk in the funding-rate-reversal dimension was activated in turn. Driven by market panic, large numbers of traders rushed into the

perpetual futures market to short as a hedge, pushing the perpetual futures price to a steep discount to spot. This drove the funding rate—previously positive and a source of stable cash flow for basis arbitrageurs—rapidly negative, to extreme levels (consistent with the scenario at the start of this chapter, plunging from a normal positive value of about +0.03% per 8 hours to a deeply negative −0.10% per 8 hours). Overnight, the strategy's source of return became a continuing cost, which not only eroded profit but, more important, accelerated the consumption of margin, stacking severely on top of the pressure in the margin-and-forced-liquidation dimension.

Risk in the liquidity-depletion dimension also emerged rapidly in this process. When arbitrageurs tried to close out and cut losses under the dual pressure of margin and funding, they found the market lacked enough counterparties to absorb their closing orders. Large numbers of participants trying to exit at once caused bids to withdraw quickly and market depth to shrink sharply. As a result, the basis (the price difference between perpetual futures and spot) widened in an instant from tens of basis points to hundreds or even thousands. This meant not only enormous slippage losses on closing out; for many strategies that rely on basis convergence, the extreme divergence of the basis was itself a severe blow.

Under the pincer of these three pressures, the counterparty-and-platform-risk dimension did not escape either. Amid extreme volatility and the pressure of large-scale fund outflows, some centralized exchanges experienced withdrawal delays or even suspensions. This posed a serious threat to cross-exchange arbitrageurs who needed to move funds flexibly among multiple platforms (see Figure 17-25) to meet margin requirements. Frozen funds meant they could not transfer capital from the profitable leg to the losing leg, and a position still hedged in logic faced the risk of a single-leg forced liquidation at the operational level because its capital was severed.

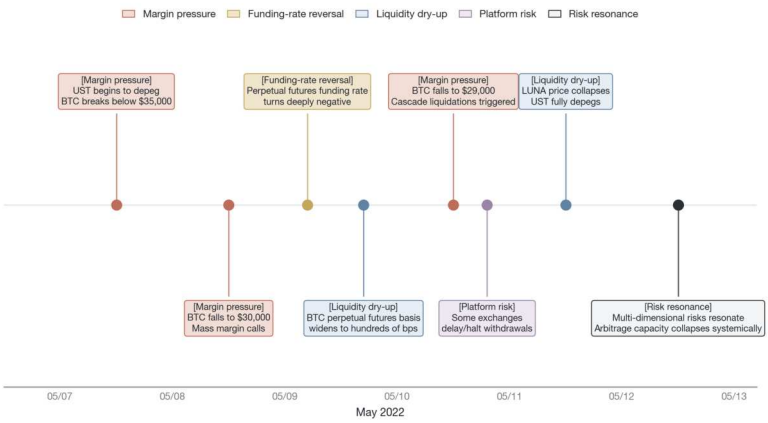


Figure 17-25. A timeline of the May 2022 crisis (Data source: compiled by the author)

Figure 17-25 reconstructs, as a timeline, the day-by-day evolution of the crisis from May 6 to May 13, 2022, clearly showing the temporal sequence in which the risk dimensions were activated one after another: the UST depegging sparked market panic (May 8) →

the funding rate turned negative (May 9) → mass liquidations began (May 10–11) → exchange withdrawal delays appeared (May 12). This temporal structure shows that risk resonance does not occur instantaneously but forms and reinforces itself gradually within a window of 48 to 72 hours.

In this crisis, the four core risk dimensions—margin pressure, rate reversal, liquidity evaporation, and platform risk—were no longer independent problems but formed a self-reinforcing positive-feedback loop. Falling prices triggered margin calls and deepened market panic; panic drove the funding rate negative and depleted liquidity; and depleted liquidity made closing out extraordinarily difficult and costly, further magnifying losses, potentially triggering more liquidations, and again pushing prices down. It was precisely this multi-dimensional simultaneous activation and positive feedback that made arbitrage capacity collapse from normal levels to near zero within a few days, with market efficiency collapsing systemically along with it.

Notably, although execution and timing risk (Dimension IV), smart contract and oracle risk (Dimension VI), and regulatory and compliance risk (Dimension VII) were not the main drivers of the May 2022 crisis, this does not mean the three were unimportant. Execution risk was relatively muted in this crisis mainly because the crisis propagated far faster than arbitrageurs' normal trading cadence: most arbitrageurs were forcibly liquidated before they even had a chance to execute a close, so execution latency did not become the bottleneck. Smart-contract risk was not significantly activated because the main battleground of this crisis was the CEXs rather than DEX perpetual futures protocols, on-chain derivatives being a small market share at the time. As for regulatory risk, May 2022 predated the formation of a global crypto regulatory framework, and regulatory action was neither a trigger nor an amplifier of this crisis. In future systemic crises, however, as the market share of DeFi derivatives grows and the global regulatory framework tightens, the probability that these three dimensions are activated simultaneously will rise markedly; at that point, the dimensions of risk resonance will expand from four toward all seven, and its destructive power will multiply accordingly.

Comparing the analysis above with the FTX collapse of November 2022 further tests the generality of the risk resonance model. Unlike the LUNA/UST crisis, which was driven by margin pressure and funding rate reversal, the FTX crisis was initially triggered by counterparty and platform risk (Dimension III)—the collapse of the exchange's own solvency. Yet the transmission paths of the two crises converge on the same destination: FTX's withdrawal suspension severed arbitrageurs' fund-redeployment channels (activating Dimension III), which set off panic selling and a sharp price fall (activating Dimension I, margin pressure); the ensuing panic drove the funding rate to extreme negative values (activating Dimension II); and the collective exit of market makers and arbitrageurs drained liquidity market-wide (activating Dimension V). The key difference between the two lay in the initial triggering dimension and the transmission sequence: the LUNA crisis followed the endogenous path "margin → funding → liquidity," whereas the FTX crisis followed the exogenous-shock path "platform → margin → funding → liquidity." But both ultimately converged on the same resonant terminal state—a multi-dimensional positive-feedback loop producing a systemic collapse in arbitrage capacity. This pattern of "different starting points, same terminal state" provides strong support for the general explanatory power of the risk resonance model.

17.9.2 The risk transmission matrix

The May 2022 case vividly demonstrated the destructive power of risk resonance, but the transmission mechanism behind it is not entirely random. Different risk dimensions have intrinsic linkages of differing strength, and some combinations are naturally more prone to forming positive-feedback loops. To understand this more systematically, we can construct a *risk transmission matrix* that qualitatively assesses the amplifying effect on other dimensions' risk when the risk in one dimension is triggered.

The core idea of the matrix is to assess a *cross-impact coefficient*: when a crisis occurs in one dimension, to what degree does it aggravate the crisis in the others? We focus mainly on the four most central endogenous risk dimensions in a crisis (see Figure 17-26): margin pressure, funding rate reversal, platform risk, and liquidity evaporation.

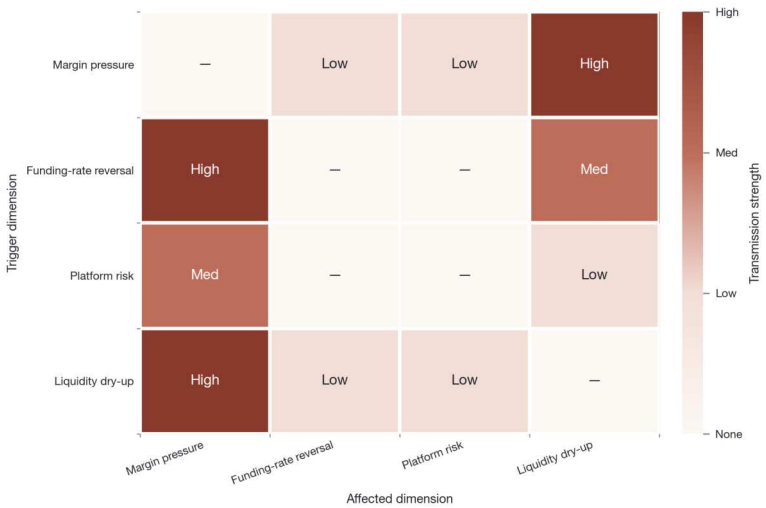


Figure 17-26. The risk transmission matrix (Data source: conceptual schematic drawn by the author; the intensities are qualitative assignments)

Based on a retrospective analysis of four major crises—March 2020, May 2021, May 2022, and November 2022—we can annotate each path in the transmission matrix with a qualitative transmission strength (strong/medium/weak, serving only as the visual scale of Figure 17-26 and not to be read as a statistical estimate). Of these, the transmission from margin pressure to liquidity evaporation is strongest (the direct impact of forced liquidation on order-book depth); the reverse transmission from liquidity evaporation to margin pressure is the next strongest (the accelerating effect of amplified slippage on margin consumption); the transmission from funding rate reversal to margin pressure is strong (the erosion of the capital buffer by sustained negative funding, consistent with Figure 17-26 and the "significant one-directional amplification" discussed below); and the transmission from platform risk to margin pressure is of medium-to-low strength (the blocking effect of frozen funds on cross-platform margin redeployment). A sample of only four crises falls far short of what is needed to identify precise coefficient differences to two significant figures, so only qualitative ordinal ranks, not specific

numbers, are given here. Rigorous coefficient estimation awaits future work using the CoVaR method (Adrian & Brunnermeier, 2016) [30] to measure the conditional value-at-risk contribution among dimensions, or a DCC-GARCH model to characterize the time-varying nature of the dynamic conditional correlations among dimensions. Transmission paths also differ markedly across crisis types: in a margin/funding-driven crisis (such as May 2022), the transmission strength of the "margin → liquidity" path may approach the upper bound of the estimates above; whereas in a platform-risk-driven crisis (such as November 2022), the "platform → margin" path may become the dominant channel. The single coefficient values above should therefore be understood as a summary description across crisis types and should be adjusted, in any specific risk assessment, to the triggering characteristics of the stress scenario.

Several key transmission paths and positive-feedback loops can be observed in the matrix. The strongest two-way amplification exists between margin pressure and liquidity evaporation. When violent price swings raise margin pressure, a flood of passively triggered forced-liquidation orders pours into the market and consumes the limited liquidity in an instant, intensifying risk in the liquidity-depletion dimension. Conversely, when liquidity evaporation widens spreads and thins market depth, arbitrageurs' slippage costs on closing out rise sharply, which directly enlarges the real losses on their positions, consumes margin faster, and intensifies pressure in the margin-and-forced-liquidation dimension. This "margin–liquidity" spiral is one of the most typical positive-feedback loops in a market collapse. On top of this, funding rate reversal exerts a significant one-directional amplification on both margin pressure and liquidity evaporation. When the funding rate turns from positive to negative and reaches extreme levels, it directly accelerates the consumption of basis arbitrageurs' margin and thus markedly intensifies pressure in the margin-and-forced-liquidation dimension; at the same time, sustained negative funding forces large numbers of arbitrageurs to abandon the strategy collectively and exit, and this collective, one-directional closing is itself an enormous shock to market liquidity, intensifying risk in the liquidity-depletion dimension. The polarization of the funding rate acts as a catalyst in the risk-transmission chain. The distinctive role of platform risk deserves special attention. Although the direct amplification coefficient of platform risk (such as withdrawal restrictions) appears low in the matrix, it performs a *risk-entrenchment* function. When other risk dimensions have already been activated, the appearance of platform risk severs the key channels through which arbitrageurs redeploy capital between markets, so that margin pressure that could have been relieved through cross-market operations becomes unmanageable—entrenching and amplifying a local problem. Its aggravating effect on margin pressure is indirect but severe.

Through this matrix, we can see clearly that risk resonance is not a simple linear sum. An initial shock (a price decline, say) first activates margin pressure; margin pressure then activates liquidity risk through a strong transmission path, while a shift in sentiment activates funding-rate risk; and the three act together to turn the arbitrageur's situation sharply for the worse. If platform risk is layered on at this point, the probability that the entire arbitrage system collapses rises steeply. Understanding these key transmission paths is essential for risk management and stress testing: it tells us that what deserves attention is not the isolated impact of any single risk but the possibility that these high-coefficient paths are activated simultaneously.

The choice of dimensions in the 4×4 matrix above reflects the empirical characteristics of the crises of 2020–2022, but as market structure evolves, the three excluded dimensions are taking on growing systemic importance. The execution-and-timing dimension may become a crisis amplifier in a market increasingly dominated by high-frequency arbitrage; the smart-contract-and-oracle dimension is highly likely to be activated in the next systemic crisis as the market share of DeFi perpetual futures grows (as with Hyperliquid's rapid rise in 2024–2025); and the regulation-and-compliance dimension, as the global regulatory framework tightens, may at any moment become a trigger or amplifier of a crisis. A forward-looking risk-assessment framework should construct the full 7×7 transmission matrix, paying particular attention to two paths that have not yet been fully exposed in the historical sample but could prove highly destructive in the future market structure: "oracle manipulation → liquidation cascade → liquidity evaporation" and "sudden regulatory event → trading-route blockage → liquidity fragmentation."

17.9.3 The nonlinear amplification effect

The most central feature of the risk resonance model is its *nonlinearity*. In a normal market, the constraints an arbitrageur faces can be treated as approximately linear: a 10% increase in the margin requirement, for example, may reduce the optimal position size by a corresponding 10%. Risk is predictable and manageable, and its effect is proportional to the strength of the trigger. Once the market enters the resonant state, however, this linear, additive logic breaks down. The total risk of the system grows far faster than the simple sum of the growth rates of the individual sub-risks, exhibiting a superlinear, self-reinforcing amplification (here "superlinear" means growing faster than a linear sum, not an exponential function in the strict sense).

The root of this nonlinear amplification lies in the positive-feedback loops described earlier. In a simple linear system, input A produces output B and input C produces output D, so total output is B + D. In a nonlinear system with positive feedback, however, input A produces output B, output B in turn strengthens inputs A and C, and the output D produced by input C may also strengthen input A. The various parts of the system excite one another, forming a self-accelerating loop whose final output is far greater than B + D.

A functional relationship makes this more intuitive. If "market stress" (combining factors such as price volatility and fund outflows) is placed on the horizontal axis and "total arbitrage capacity" on the vertical, the relationship between the two is not a straight line but an S-shaped curve (as shown in Figure 17-27), which undergoes a sharp nonlinear change near a certain point.

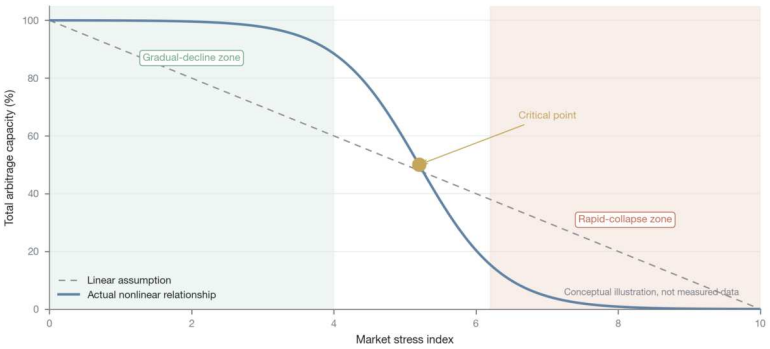


Figure 17-27. The nonlinear relationship between arbitrage capacity and market stress (Data source: conceptual schematic drawn by the author, not empirical data; the axes are illustrative scales)

The nonlinear relationship above can be characterized heuristically through a conceptual system of simultaneous equations that captures the positive-feedback structure between margin pressure and liquidity depletion (it must be stressed that the equations below serve only to express the logical structure of this feedback formally; their parameter calibration, numerical solution, and phase-diagram analysis are left to future empirical research, and this section makes no quantitative estimate of them):

$$M_{t+1} = M_t - \alpha \cdot \Delta P_t - \beta \cdot \max(-r_t, 0) \cdot N - \eta \cdot \frac{\Delta P_t}{L_t}$$

$$L_{t+1} = L_t - \gamma \cdot \text{Liquidations}_t + \delta \cdot \text{NewOrders}_t$$

$$\text{Liquidations}_t = f(M_t < M_{\text{threshold}})$$

Here Mt is the arbitrageur's margin balance at time t, ΔPt is the change in the underlying price, rt is the current funding rate, N is the notional size of the position, and Lt is market liquidity (available order-book depth). The parameter α captures the intensity of the shock from price movement to margin, β captures the speed at which the funding rate consumes margin, and η captures the amplifying effect of liquidity depletion on execution cost; here α and η are scaling coefficients that convert price and liquidity shocks into the currency units of margin (with ΔPt taken as a percentage change). The term max($-r$t, 0) is used rather than $|r$t$|$ because only a negative rate consumes the short arbitrageur's margin; as Lt approaches zero, the η term tends to infinity, reflecting the explosive growth of closing costs inside a liquidity black hole. When Mt falls below the threshold Mthreshold, forced liquidation is triggered, and the liquidation orders consume liquidity at the coefficient γ. This system of equations reveals the core of resonance: the exhaustion of margin triggers liquidation (the third equation), liquidation consumes liquidity (the second equation), and falling liquidity amplifies slippage and thereby accelerates margin consumption (feeding back into the η term of the first equation), forming a self-reinforcing collapse spiral. This model is a simplified endogenous-risk framework that does not incorporate the transmission effects of external macro-liquidity conditions (such as the dollar interest-rate environment or the global risk-appetite cycle), a direction for further extension.

The model contains two starkly different regions.

In the *gradual-decline region*, market stress is rising but has not yet triggered strong positive feedback among the risk dimensions. Arbitrageurs may need to reduce leverage somewhat or absorb some extra slippage, but their core arbitrage function still operates. The impact of risk is roughly linear, and arbitrage capacity declines in proportion to the rise in market stress. At this stage, market efficiency has declined somewhat, but the pricing mechanism has not yet failed.

Once market stress crosses a certain *critical point*, however, the system enters the *rapid-collapse region*. This critical point marks the simultaneous activation of several core positive-feedback loops (such as the "margin–liquidity" spiral). Now any small additional pressure can be amplified disproportionately by the system's internal amplification mechanisms. Arbitrage capacity no longer declines gradually but collapses sharply and nonlinearly. Once the key positive-feedback loops are activated, the system undergoes a chain-like collapse of capacity at great speed. At this stage, the sum of the risks is far greater than the sum of the parts, because the risks no longer combine by addition but by multiplication.

The events of May 2022 are a faithful portrait of this nonlinear amplification. Before May 8, the market was already under pressure, but the arbitrage system was still functioning normally. After the two key events—the UST depegging and the funding rate turning negative—the system quickly crossed the critical interval, and arbitrage capacity contracted sharply to near-standstill over the following 48 hours. This explains why market collapses are always so rapid and exceed expectations: the process is not gradual but more like a *phase transition* of state (a physics analogy denoting the rapid jump of a system from one state to another—also understandable as a regime shift—rather than a thermodynamic phase transition in the strict sense).

17.9.4 The phase-transition critical point

The turning point at which arbitrage capacity shifts from linear decline to nonlinear collapse marks the *phase transition* of the risk state from "manageable" to "uncontrollable" (a physics analogy, here and below; also understandable as a regime-transition interval, for which this chapter provides no identifiable critical exponent or scaling law). Understanding the nature of this turning point has important analytical significance for all market participants, and for risk managers and regulators in particular. It tells us that linear risk models built on normal-market data may fail in a crisis, because they cannot capture this state-jump behavior. The observable indicators proposed in Section 17.1.4 (the funding-rate z-score, the ratio of liquidation volume to open interest, the coefficient of variation of order-book depth, and the dispersion of the cross-exchange basis) can serve as ex ante early-warning proxies for this transition interval, but what the current framework provides is a "combination of early-warning signals," not a critical point that can be located precisely.

This critical point is not a fixed number but a dynamic threshold determined jointly by market structure, participant behavior, and the macro environment. We can, however, identify some common catalysts that are often the key triggers pushing the system across the critical point.

The first type of catalyst is the *synchronization of collective behavior*. In a normal market, arbitrageurs behave in diverse ways—some entering, some exiting—which sustains the market's liquidity and stability. When a sufficiently strong external or internal shock appears (such as a collective turn to negative funding), however, it becomes a powerful coordinating signal that leads large numbers of previously independent arbitrageurs to make the same decision at the same time—collectively closing basis-arbitrage positions, for example. This synchronization of behavior is the direct cause of a liquidity black hole and a stampede, and a key step in driving the system into the nonlinear-collapse region. In crypto markets, the speed of this synchronization is markedly accelerated by the real-time transparency of social media and on-chain data. Unlike in traditional financial markets, the core information of crypto markets—liquidation volume, funding rates, changes in a protocol's total value locked (TVL), and the on-chain operations of whale addresses—is broadcast to the entire market in real time through Twitter/X, Telegram groups, and data dashboards such as Coinglass and DefiLlama. This ultra-rapid diffusion of information lets a "coordinating signal" reach all market participants worldwide within minutes, compressing a panic-transmission process that might take hours or even days in traditional markets into a very short window. The transparency of on-chain data becomes a *double-edged sword* here: it improves informational efficiency in a normal market, but in a crisis it becomes an efficient medium for transmitting panic, accelerating the phase transition from the *stressed* state to the *resonant* one.

The second type of catalyst is the *failure of core infrastructure*. Within the framework of the Arbitrage Infrastructure Hypothesis, large arbitrageurs and core exchanges are themselves key nodes in the network that transmits market efficiency. When these nodes fail—whether for internal reasons (such as the FTX collapse) or under external pressure—the effect is far more than the removal of one participant. It damages the topology of the entire network, severs many arbitrage routes, and causes a systemic breakdown in the transmission of efficiency. The failure of a core node has a global impact, enough to push the entire system into an uncontrollable state.

The third type of catalyst is the *procyclicality of risk models*. The risk-management models used by many institutional and individual investors are themselves procyclical. When market volatility rises, for example, a model calls for reducing leverage and cutting positions. When everyone uses similar models and acts at the same time (see Figure 17-28), this collective "de-risking" itself creates greater risk. This paradox is highly consistent with the *endogenous risk* theory of Danielsson, Shin, and Zigrand (2004) [31]. They argued that when market participants use similar risk-measurement tools (such as value-at-risk, or VaR, models) to manage their portfolios, the positive-feedback loop of rising volatility → the model calls for cutting positions → collective selling → volatility rises further makes risk itself an endogenous variable rather than an exogenous background parameter. Adrian and Shin (2010) [32] showed further that the procyclical leverage adjustment of financial intermediaries is the core transmission mechanism of this endogenous risk. In crypto markets, because exchanges' automatic liquidation engines and tiered-margin regimes functionally play the role of risk models in traditional finance, this procyclicality is institutionally embedded in the market structure itself. It intensifies selling pressure and liquidity depletion, forming a "risk

management produces risk" paradox that accelerates the system's crossing of the critical point.

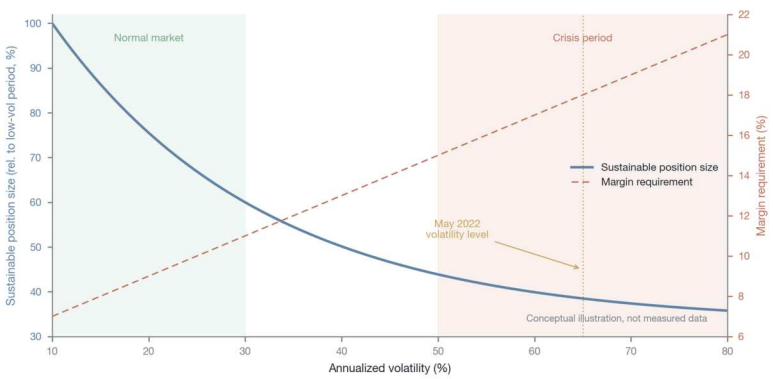


Figure 17-28. The suppressive effect of the margin constraint on arbitrage position size (Data source: drawn by the author; a theory-based conceptual illustration, not empirical data)

Figure 17-28 quantifies the amplifying effect of risk-model procyclicality: as annualized volatility rises, the safe position size computed by the model steadily contracts, and this contraction is steepest in the low-to-moderate volatility range—each step up in volatility cuts the position sharply. When volatility rises to crisis levels (such as the roughly 65% of May 2022), the position size has already been suppressed to a very low level, and further decline flattens out. This means that it is precisely in the phase of a mild rise in volatility that large numbers of arbitrageurs, driven by their risk models, cut positions in concert within the same time window. This risk-model-driven synchronized deleveraging becomes the key catalyst pushing the market toward the critical point.

The consequences of the phase transition from "manageable" to "uncontrollable" are profound. Once the critical point is crossed, the market's self-repair mechanism fails. As endogenous providers of liquidity, arbitrageurs retreat collectively when their own survival is at stake, making the recovery of liquidity extraordinarily difficult. A collapse of trust—in an exchange or a stablecoin, say—is harder still to rebuild in the short term. This explains why, after a major crisis, the market's level of efficiency (reflected in various spreads and bases) often takes weeks or even months to recover slowly. The system's collapse is asymmetric: the collapse may take only hours, whereas rebuilding trust, restoring capital, and repairing the arbitrage network take weeks to months. This pronounced time asymmetry between collapse and recovery is an important feature of risk resonance and the nonlinear phase transition.

17.10 A risk anatomy of the Ethena protocol

The Ethena protocol raises the basis-arbitrage mechanism from an individual strategy to a protocol-level, scaled-up operation, making it a representative case for analyzing how arbitrage constraints map into systemic risk. This section takes Ethena as its subject, reviews the risk-exposure characteristics of its arbitrage mechanism, maps the

seven constraint dimensions onto its operating structure, and tests the explanatory power of the analytical framework above through the real stress test of April 2024. A note on the division of labor: Section 16.4.3 addressed Ethena's mechanism design and its "arbitrage-as-a-service" model, whereas this section turns to its risk anatomy; Section 17.3.3 raised the "negative funding → redemption → unwinding → still more negative funding" feedback loop and the collateral basis risk, which this section develops systematically without repeating the mechanism details.

As Chapter 16 described, Ethena constructs the delta-neutral synthetic asset USDe by holding an equivalent amount of spot assets while simultaneously opening a short perpetual futures position of equal size, providing yield while suppressing the perpetual futures premium.

When we turn to the risk dimensions, however, this "source of efficiency" immediately reveals its symmetric character as a source of risk. Ethena is, in essence, a scaled-up positive-carry position, which means its dependence on the funding rate is absolute. In a bull or normal market, strong demand for long leverage ensures that the short side continues to receive positive funding, which is treated as the protocol's profit. When sentiment reverses, however, this profit instantly becomes a steep cost of carry. Unlike an individual arbitrageur, Ethena—a giant entity managing billions of dollars of assets—faces not the simple failure of a strategy but a protocol-level systemic fragility. This fragility arises because it has frozen what was originally dispersed, dynamic arbitrage into a rigid protocol architecture dependent on scale effects, which makes it prone to triggering risk resonance at extreme moments.

To assess Ethena's risk profile quantitatively, we can map the seven-dimensional constraint topology of this chapter directly onto its operating mechanism. Begin with margin and forced liquidation (the margin-and-forced-liquidation dimension). Although Ethena uses full collateralization and keeps leverage low on the CEX side, the reality that cross-exchange margin cannot be pooled forces it to pre-fund capital across multiple platforms. When the market swings violently—especially when the mark price and the market price diverge—Ethena must redeploy funds in real time to meet margin calls, and any delay at the execution level can cause a single-leg position to be forcibly liquidated, instantly exposing enormous delta risk.

Ethena also faces an additional and material risk not explicitly classified within the seven-dimensional constraint topology: *collateral basis risk*. The liquid staking derivatives in its portfolio, such as stETH, are not rigidly pegged 1:1 to the underlying asset, ETH. In a systemic crisis, the secondary-market liquidity of stETH can shrink sharply, driving its price well below that of ETH (in June 2022, stETH traded at a discount of more than 5% at one point). This depegging at the collateral level pushes Ethena's actual collateralization ratio below its nominal level, intensifying margin pressure and, in an environment of funding rate reversal, creating a double squeeze of "rate erosion plus collateral shrinkage."

Funding rate reversal (the funding-rate-reversal dimension) is the most direct threat Ethena faces. When the market enters a period of panic, a negative funding rate means that USDe holders not only earn no yield but that the protocol itself must pay the market. Although Ethena has established a reserve fund as a buffer, the reserve's size

relative to its TVL determines its capacity to withstand sustained negative funding. The liquidity black hole (the liquidity-depletion dimension) is especially prominent in the redemption scenario. Once negative funding triggers large-scale redemptions, Ethena must close its short positions on the secondary market and sell the spot, and this large-scale, one-directional operation easily sets off a stampede, sharply raising exit slippage, further eroding the reserve, and accelerating the depegging of USDe. Counterparty risk (the counterparty-and-platform dimension) and smart-contract risk (the smart-contract-and-oracle dimension) likewise cannot be ignored: the collapse of any partner exchange or a vulnerability in an on-chain contract can cause a systemic failure of the protocol's arbitrage structure.

Ethena's systemic importance can be estimated with a set of concentration metrics. At the peak of its TVL, Ethena's short perpetual futures positions accounted for a single-digit percentage of market-wide BTC/ETH perpetual open interest (about 5% of Ethereum perpetual open interest in early 2024; see Section 16.4.3), but on particular contracts at some centralized exchanges, industry estimates put its share of that contract's open interest at about 10%–15%. The following estimates its unwinding impact using a conservative, illustrative scenario (the parameters are assumed values used to convey magnitude, not a precise forecast). Suppose Ethena must fully unwind $5 billion of notional short positions within 7 days; that daily volume on BTC perpetual futures at major exchanges is about $20 billion (a magnitude referenced to industry statistics such as CoinGecko); and that crisis-period liquidity shrinks to 20%–30% of normal. Then its unwinding volume would account for a considerable share of the daily liquidity available in the crisis. Using a simplified square-root price-impact model (impact proportional to the square root of the ratio of unwinding size to market depth), a one-directional unwind of this scale might produce cumulative slippage on the order of tens to over a hundred basis points, eroding tens of millions to over a hundred million dollars of reserves. More important, one-directional buying of this scale (to close the short) would significantly push up the perpetual futures price and shift the market-wide funding rate, creating a spillover effect: "Ethena unwinds → funding fluctuates → other arbitrageurs unwind → liquidity deteriorates further."

In April 2024, the crypto market went through a notable correction, providing an excellent experimental window for observing Ethena's behavior under stress. As the Bitcoin price slid from its high, the market-wide perpetual futures funding rate fell rapidly within a few days from very high levels to zero and even negative. In the process, the annualized yield of sUSDe fluctuated violently. According to observed data, a yield that had exceeded 30% fell to single digits within a single week, at times even touching the minimum-yield protection floor set by the protocol. Specifically, the 7-day moving-average annualized yield of sUSDe fell rapidly from about 35% at the start of April (consistent with the publicly recorded historical peak of about 35%) within a few days, dropping to single digits and even recording a negative yield in some intervals. Over the same period, USDe traded at a slight negative premium on the secondary market (in major trading pools such as Curve); according to Ethena's own account of this stress test, for the great majority of the sell-off its market price held within about 20 basis points of $1. The period saw sizable net redemptions (the company disclosed that more than 100 million USDe were redeemed, roughly a single-

digit percentage of the then-outstanding supply), and the protocol drew on its reserves to pay the short funding and maintain the sUSDe yield. Among the precise figures above, the supply and price metrics can be read directly from on-chain sources and DefiLlama, whereas the proportion of reserves consumed and the per-exchange open-interest shares are the author's estimates based on public data and should be treated with caution; the company also disclosed that total market open interest contracted by about 15% in a single day over the same period (a market-level figure, not Ethena's own reduction).

This stress test revealed the real effectiveness of Ethena's reserve mechanism. When the funding rate turned negative, the protocol began drawing on the reserve to pay the short funding, maintaining a positive—or at least zero—return for sUSDe holders. Although Ethena's TVL did not see large-scale outflows in this test, USDe traded at a slight negative premium in some decentralized trading pools, reflecting the market's concern about the protocol's solvency in a prolonged negative-funding environment. The event proved that although the reserve can buffer short-term rate fluctuations, it cannot resolve the capital-flight pressure created by structurally negative funding over the long run. This pressure transmits to the user side through a falling yield and then, through redemptions, presses back on liquidity at the exchange end, forming a textbook risk-feedback loop.

The Ethena case offers a profound lesson for understanding the systemic risk of protocol-level arbitrage. It shows the abrupt change in the nature of risk as arbitrage shifts from "dispersed and individual" to "scaled-up and protocol-level" [33]. At the individual level, arbitrage constraints are obstacles that limit profit; at the protocol level, they become a source of risk that can trigger systemic collapse. When an arbitrageur plays the dual role of stablecoin issuer and market-infrastructure provider at once, its retreat is no longer a simple taking of profit but a double shock to market liquidity and stability.

In addition, Ethena faces a seventh-dimension risk that has not yet been fully discussed: uncertainty over its regulatory characterization. As a redeemable token whose yield derives from an arbitrage strategy, sUSDe occupies a legal gray area in several jurisdictions. If sUSDe were deemed an investment contract in a major jurisdiction (meeting the criteria of the Howey test), its unregistered public offering would face serious legal consequences. More broadly, Ethena's model of pooling user funds and trading derivatives on CEXs could, in some jurisdictions, trigger the regulatory classification of a "collective investment scheme" or "managed fund" and thus require an asset-management license. What makes this regulatory-characterization risk distinctive is its *sword-of-Damocles effect*: even before any regulatory action actually occurs, the mere existence of regulatory uncertainty is enough to deter institutional capital, limit the protocol's growth in scale, and accelerate rational holders' redemption decisions in a market panic.

Ultimately, Ethena's success or failure hinges on its ability to control "risk resonance." If it can smooth funding fluctuations by introducing more diversified collateral and better risk-hedging tools, it will continue to advance the efficiency of crypto markets; if, on the other hand, it expands in scale far faster than its capacity to withstand risk grows, it may, in the next and more extreme market resonance, re-enact the systemic collapses

of algorithmic-stablecoin history. This recognition carries important implications for risk assessment for both protocol designers and strategy participants.

17.11 Chapter summary

Around the analytical framework of the *arbitrage constraint topology*, this chapter has systematically deconstructed the seven core risk dimensions facing perpetual futures arbitrage: margin and forced liquidation; funding rate reversal; counterparty and platform risk; execution and timing risk; liquidity depletion; smart contract and oracle risk; and regulation and compliance risk. Each dimension reveals a systematic departure of "risk-free arbitrage" from what obtains in real markets.

The chapter's core theoretical contribution is the *risk resonance model*. Through an empirical review of the Terra/LUNA crisis of May 2022, with the FTX collapse of November 2022 as a comparison case, the chapter has argued that these constraint dimensions do not operate independently of one another but, under extreme conditions, activate one another and cross-amplify through positive-feedback mechanisms. What was observed in full in these two flagship crises was mainly the resonance of four core dimensions—margin, funding rate reversal, platform risk, and liquidity depletion; the three dimensions of execution, smart contract, and regulation are incorporated into the framework more in the form of "possible activation in a future market structure." In this sense, "seven-dimensional resonance" is a proposition that is theoretically extensible but not yet fully validated by history. The stacking of margin pressure and funding rate reversal, and the resonance of liquidity depletion and platform risk, together constitute a nonlinear risk-amplification network. This model explains why the decay of arbitrage capacity is not a linear, gradual process but exhibits a state-jump from "manageable" to "uncontrollable": when market stress breaks through a certain critical interval, the carrying capacity of the arbitrage system can contract sharply within a very short time.

The Ethena case study further tests the explanatory power of this framework. As a representative practice that raises basis arbitrage from an individual strategy to a protocol-level, scaled-up operation, Ethena's experience shows that the institutionalization of arbitrage magnifies the potential destructive power of risk resonance. When an arbitrageur takes on the multiple roles of stablecoin issuer and market-infrastructure provider at once, the constraints it faces are no longer merely boundary conditions on profit but structural weak points that can trigger the transmission of systemic risk.

The chapter's analysis provides the necessary risk-dimension complement to the Arbitrage Infrastructure Hypothesis of Chapter 16. If arbitrageurs are the central custodians of market efficiency, then arbitrage constraints define the structural upper bound on that efficiency. The persistence of arbitrage returns is, in essence, the risk premium the market pays arbitrageurs for bearing the seven-dimensional constraints and resonance risk. This recognition redefines arbitrage returns from "the residue of market inefficiency" to "reasonable compensation for bearing risk," providing a more complete theoretical foundation for understanding the efficiency frontier of crypto markets.

Looking ahead, the risk landscape of perpetual futures arbitrage will keep changing as market infrastructure evolves. At the institutional level, a cross-exchange unified-

margin system, the introduction of a CCP mechanism, and the clarification of the regulatory framework promise to reduce the intensity of the margin, counterparty, and compliance constraints. At the technical level, the maturation of on-chain clearing infrastructure, stronger decentralization of oracle networks, and improved cross-chain interoperability will help ease the constraints of the execution, liquidity, and smart-contract dimensions. Yet the nonlinear character of risk resonance means that even if the constraint along a single dimension is eased, systemic risk may reaggregate through new transmission paths. Future research should further quantify the transmission coefficients among dimensions, build an operational early-warning model of risk resonance, and explore feasible paths for embedding risk-mitigation mechanisms at the level of protocol design.

References

[1] Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35–55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

[2] DeLong, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, *98*(4), 703–738. https://doi.org/10.1086/261703

[3] Xiong, W. (2001). Convergence trading with wealth effects: An amplification mechanism in financial markets. *Journal of Financial Economics*, *62*(2), 247–292. https://doi.org/10.1016/S0304-405X(01)00078-2

[4] Liu, J., & Longstaff, F. A. (2004). Losing money on arbitrage: Optimal dynamic portfolio choice in mispriced markets. *The Review of Financial Studies*, *17*(3), 611–641. https://doi.org/10.1093/rfs/hhg029

[5] Gromb, D., & Vayanos, D. (2010). Limits of arbitrage. *Annual Review of Financial Economics*, *2*(1), 251–275. https://doi.org/10.1146/annurev-financial-073009-104107

[6] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357–384. https://doi.org/10.2307/1912559

[7] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, *135*(2), 293–319. https://doi.org/10.1016/j.jfineco.2019.07.001

[8] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn014

[9] Coinglass. (n.d.). Cryptocurrency market data and analytics. https://www.coinglass.com/

[10] Diamond, D. W., & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, *91*(3), 401–419. https://doi.org/10.1086/261155

[11] Ethena Labs. (2024). Ethena USDe analytics [Dashboard]. DefiLlama. https:

//defillama.com/protocol/ethena

[12] Investopedia. (2023, January 4). FTX crypto exchange collapse explained. https://www.investopedia.com/ftx-exchange-5200842

[13] Kharif, O. (2022, November 12). At least $1 billion of client funds missing at failed crypto firm FTX. *Reuters*. https://www.reuters.com/markets/currencies/least-1-billion-client-funds-missing-failed-crypto-firm-ftx-sources-2022-11-12/

[14] Michaels, D. (2023, June 9). Binance.US halts dollar deposits after SEC crackdown. *Reuters*. https://www.reuters.com/technology/binanceus-halts-dollar-deposits-2023-06-09/

[15] Bybit. (n.d.). Auto-deleveraging (ADL). Bybit Help Center. https://www.bybit.com/en-US/help-center/s/article/Auto-Deleveraging-ADL

[16] Aquilina, M., Budish, E., & O'Neill, P. (2022). Quantifying the high-frequency trading "arms race." *The Quarterly Journal of Economics*, *137*(1), 493–564. https://doi.org/10.1093/qje/qjab032

[17] Muck, M., Schmidl, T., & Wolf, J. (2025). Wish or reality? On the exploitability of triangular arbitrage in cryptocurrency markets. *Finance Research Letters*, *73*, 106574. https://doi.org/10.1016/j.frl.2024.106574

[18] Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., & Juels, A. (2020). Flash Boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. In *2020 IEEE Symposium on Security and Privacy (SP)* (pp. 910–927). IEEE. https://doi.org/10.1109/SP40000.2020.00040

[19] Paradigm. (2025). *The key neutrality of baselayer markets*. Paradigm Research.

[20] Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, *130*(4), 1547–1621. https://doi.org/10.1093/qje/qjv027

[21] Stein, J. C. (2009). Presidential address: Sophisticated investors and market efficiency. *The Journal of Finance*, *64*(4), 1517–1548. https://doi.org/10.1111/j.1540-6261.2009.01472.x

[22] Khandani, A. E., & Lo, A. W. (2011). What happened to the quants in August 2007? Evidence from factors and transactions data. *Journal of Financial Markets*, *14*(1), 1–46. https://doi.org/10.1016/j.finmar.2010.07.005

[23] Halborn. (2025). Explained: The GMX hack (July 2025). https://www.halborn.com/blog/post/explained-the-gmx-hack-july-2025

[24] CoinDesk. (2025, September 2). Bunni DEX halts smart contracts after exploit drains $8.4M across chains. https://www.coindesk.com/web3/2025/09/02/bunni-dex-halts-smart-contracts-after-exploit-drains-usd8-4m-across-chains

[25] The Block. (2026, January 25). Matcha Meta sees $16.8 million drained in SwapNet incident. https://www.theblock.co/post/386986/matcha-meta-swapnet-incident

[26] Three Sigma. (2025). 2024 most exploited DeFi vulnerabilities. https://threesigma.xyz/blog/exploit/2024-defi-exploits-top-vulnerabilities

[27] Decrypt. (2020, November 26). Oracle exploit sees $89 million liquidated on

Compound. https://decrypt.co/49657/oracle-exploit-sees-100-million-liquidated-on-compound

[28] FX Empire. (2022). Over $2B lost in 13 separate crypto bridge hacks this year. https://www.fxempire.com/news/article/over-2b-lost-in-13-separate-crypto-bridge-hacks-this-year-1085594

[29] Hou, Y., & Liu, Y. (2023). *Regulatory arbitrage and market efficiency in decentralized finance* (Working Paper). https://doi.org/10.2139/ssrn.4567890

[30] Adrian, T., & Brunnermeier, M. K. (2016). CoVaR. *American Economic Review*, *106*(7), 1705–1741. https://doi.org/10.1257/aer.20120555

[31] Danielsson, J., Shin, H. S., & Zigrand, J.-P. (2004). The impact of risk regulation on price dynamics. *Journal of Banking & Finance*, *28*(5), 1069–1087. https://doi.org/10.1016/j.jbankfin.2003.06.002

[32] Adrian, T., & Shin, H. S. (2010). Liquidity and leverage. *Journal of Financial Intermediation*, *19*(3), 418–437. https://doi.org/10.1016/j.jfi.2008.12.002

[33] Aramonte, S., Huang, W., & Schrimpf, A. (2021). DeFi risks and the decentralisation illusion. *BIS Quarterly Review*, September 2021, 21–36. https://www.bis.org/publ/qtrpdf/r_qt2112b.htm

Arbitrage Failure and Market Anomalies

Building on the theoretical framework established in the preceding two chapters, the internal price relationships of a market driven by rational arbitrageurs cannot remain perfectly consistent at no cost. Chapter 16 showed that many persistent spreads are not simple market failures but rather the cost compensation that arbitrageurs require when they provide the services of cross-market clearing, price anchoring, and liquidity integration. Chapter 17 further explained how constraints such as margin, liquidity, platform risk, execution risk, smart contracts, and regulation raise this compensation. This chapter therefore does not label every persistent spread an anomaly; instead, it examines the residual deviations that remain after accounting for modelable arbitrage costs, risk compensation, and institutional frictions.

The existing literature and observations of public markets together point to a more nuanced conclusion: price relationships across markets are not perfectly consistent but instead exhibit layered, persistent deviations under different constraints. The spread in Bitcoin perpetual futures between leading exchanges such as Binance and OKX has, under normal conditions, narrowed to single-digit basis points (bps); these few basis points are more likely the equilibrium price of execution costs, capital pre-positioning, and platform risk than evidence of market inefficiency. The perpetual futures spread between Binance and Korean exchanges, however, typically falls in the 30–80 bps range under non-extreme conditions (data source: Coinglass), well above transaction costs. This perpetual futures spread, however, is not a like-for-like comparison with the widely known spot "kimchi premium": the former is a spread at the derivatives (perpetual futures) level, whereas the latter is a spread at the spot level, and the two are determined by different instruments, participants, and arbitrage paths. The long-run center of the spot kimchi premium is roughly 1.2% (about 120 bps, the estimate from the threshold model of Seo et al. [1]), but it is highly state-dependent and can exceed 20% or even 50% in extreme periods (consistent with the 200–2,000 bps figure for capital-controlled jurisdictions in Table 18-6 of Section 18.4.3). Because professional arbitrageurs participate more heavily in the perpetual futures market, its spreads are typically far smaller in magnitude than the spot kimchi premium, yet they remain large enough to indicate structural barriers that limit the cross-market flow of arbitrage capital. Along the funding-rate dimension, public data show that the share of days on which the funding rate of mainstream perpetual futures is positive has consistently exceeded 50%, with a long-run mean of about 75% (data source: Coinglass; this is a long-run mean spanning

bull and bear cycles, and it declines markedly in the bear-market subsample, as detailed in Section 18.3.1), pointing to a persistent "long bias" in the structure of market participants. Along the CEX–DEX dimension, cross-exchange funding-rate research shows an observable rate split between decentralized exchanges (DEXs) and centralized giants in short-sample, high-frequency data (for example, between Hyperliquid and Binance, the arbitrage spread exceeds 20 bps in about 17% of the observed periods [2]), reflecting the additional frictions of smart-contract risk, oracle latency, and the like along the arbitrage path from centralized to decentralized venues. With the advent of Bitcoin spot exchange-traded funds (ETFs), the net asset value of an ETF and the price of perpetual futures also frequently diverge markedly, revealing fundamental differences between the two products in market microstructure, participant composition, and regulatory environment. These phenomena do not imply that every residual spread is inefficient; only when a spread substantially exceeds the explicable costs described above, and displays persistence along the statistical, economic, and temporal dimensions, does this chapter call it a market anomaly. This chapter is therefore concerned with the effective boundaries of the arbitrage mechanism, not with a blanket denial of market efficiency.

Chapter 16 analyzed the operating logic of the arbitrage mechanism, and Chapter 17 systematically laid out the seven major constraints that limit arbitrage. Building on this, the present chapter turns to the observable consequences of these constraints along the price dimension. To that end, it constructs a complete analytical framework. First, it defines the market anomaly and establishes a three-part identification standard of statistical significance, economic significance, and persistence. On this basis, it proposes a "constraint–anomaly mapping" method that establishes a mechanistic correspondence between a specific anomaly and the constraint dimensions defined in Chapter 17, without claiming that each case has undergone rigorous causal identification. It then introduces an "anomaly persistence spectrum," placing anomalies on a non-exclusive continuum composed of transient, cyclical, and structural forms, so as to reveal the different causes behind them and their prospects for repair. We then examine six categories of deviations that are relatively large in magnitude and duration: basis anomalies, funding-rate anomalies, cross-exchange anomalies, cross-asset anomalies, term-structure anomalies, and CEX–DEX anomalies. After this static analysis of each category under normal conditions, we examine how they become interlinked and mutually amplifying in a crisis, ultimately erupting in synchrony to form the "crisis dynamics of anomalies." Finally, the chapter analyzes the evolutionary trends of anomalies from a broader perspective, distinguishing which anomalies can be mitigated through institutional improvement and which are fundamental costs that the market must bear under its current architecture, thereby providing a systematic understanding for designing more efficient and more robust financial markets.

18.1 An analytical framework for anomalies

In Chapter 17, we systematically analyzed the seven-dimensional constraints facing the arbitrage mechanism and explained why arbitrage forces are not omnipotent in the real world. Because these constraints exist, the market cannot eliminate every

price deviation everywhere, at all times, and at no cost. The task of this chapter is to confront the direct consequences of these constraints: the price deviations that in theory "should not exist" yet persist in reality, which we call *market anomalies*. This section first establishes a systematic framework for analyzing these anomalies, laying the methodological foundation for the detailed treatment of each specific anomaly in later sections. The framework aims to answer three core questions. First, how should we rigorously define and identify an anomaly so as to distinguish it from random market noise? Second, how should we establish a causal connection between an observed anomaly and the specific constraints discussed in the previous chapter? Third, how should we classify anomalies according to their temporal behavior in order to reveal the different driving mechanisms and repair possibilities behind them?

18.1.1 The definition and identification of anomalies

In the context of this book, a market anomaly does not refer loosely to any price fluctuation or temporary inconsistency; it has a strict operational definition. It is price behavior that should not persist over the long run under the ideal no-arbitrage conditions described in Chapter 16 and whose degree of deviation is significant in both economic and statistical terms. This definition emphasizes that anomalies are not the market's normal state but the systematic deviations that emerge in market price behavior when arbitrage forces, constrained, cannot operate effectively. To separate genuine anomalies from the market's background noise, we establish an identification framework composed of three criteria (Table 18-1).

Identification criterion	Definition	Test method
Statistical significance	The price deviation is statistically systematic rather than a chance random event	HAC-robust mean tests, the block bootstrap, regime-switching models, or extreme-value clustering tests
Economic significance	The potential profit implied by the deviation is large enough to cover all arbitrage costs	Deviation magnitude > trading commissions + slippage + funding costs + opportunity cost
Persistence	The deviation exhibits a predictable, sustained pattern along the time dimension	Autocorrelation tests are significantly positive, or half-life estimates quantify the speed of convergence

Table 18-1. The three-criterion framework for identifying anomalies (Data source: compiled by the author)

The first criterion is statistical significance. Any price deviation must first be shown statistically to be systematic rather than a chance random event. Because high-frequency crypto data typically exhibit autocorrelation, fat tails, heteroskedasticity, regime switching, and multiple-testing problems, a simple mean t-test serves only as an initial screen; a more robust approach combines HAC standard errors, the block bootstrap, threshold or regime-switching models, and extreme-value clustering tests. This is the minimum threshold for identifying an anomaly, ensuring that the

phenomena under discussion are statistically testable. Given the limits of publicly available data, the remainder of this chapter argues primarily for economic significance (whether the deviation magnitude exceeds arbitrage costs) for several anomalies, while leaving the full tests of statistical significance and persistence (reporting t-values, autocorrelation coefficients, and half-life estimates) to dedicated empirical research; a few flagship anomalies (such as the long-term positive bias in Section 18.3.1 and the asymmetric mean reversion in Section 18.2.3) are walked through all three criteria illustratively as a methodological demonstration.

The second criterion is economic significance. A deviation that is statistically significant is not necessarily profitable. If the magnitude of a spread is smaller than the total cost an arbitrageur must incur to eliminate it, then the existence of that spread is reasonable, reflecting the market's microstructure costs. A genuine anomaly, therefore, must imply a potential profit from the price deviation large enough to cover all the transaction costs involved in executing the arbitrage strategy, including but not limited to trading commissions, the bid-ask spread (slippage), funding costs, and opportunity cost. Only when the magnitude of the deviation exceeds the boundary of this "arbitrage-infeasible band" can we say that arbitrageurs have an incentive to eliminate it, and only then does its persistence constitute a phenomenon in need of explanation.

The third criterion is persistence. An anomaly is not a fleeting signal but a pattern with a temporal dimension. This means that the price deviation exhibits predictable persistence—for example, an autocorrelation test may reveal significant positive correlation in the time series, or a half-life estimate may measure the speed of convergence toward equilibrium. A price shock without persistence is more likely to be classified as the market's normal response to new information or as random fluctuation, whereas a persistent deviation points to some structural force that obstructs the market's return to equilibrium. Only phenomena that satisfy all three criteria—statistical significance, economic significance, and persistence—are admitted into the category of anomalies studied in this chapter.

When applying the criteria above to identify anomalies, it is worth further distinguishing two qualitatively different kinds of price deviation. The first is the pure anomaly, a price deviation maintained entirely by arbitrage constraints; if the relevant constraint is removed, the deviation tends to disappear. The kimchi premium created by regulatory barriers is one such case: if capital controls were lifted, the spread between the Korean market and the global market would be eliminated rapidly by arbitrageurs. The second is the equilibrium deviation, a reasonable risk compensation determined by the structure of market participants that may persist to some degree even in an idealized frictionless market but whose actual magnitude is substantially amplified by arbitrage constraints. The long-term positive bias in funding rates is the archetype: even in a fully frictionless market, a retail-heavy participant structure could still produce positive rates, but in reality the compounding of margin and liquidity constraints pushes the magnitude far beyond the equilibrium level. This chapter includes both kinds of deviation within its scope but flags the nature of each in the specific discussion, so as to avoid mistaking structural compensation for arbitrage failure, or downplaying genuine arbitrage failure as reasonable premium.

The crypto-market anomalies examined in this chapter differ significantly in cause from those discussed in the traditional finance literature (such as the value effect and the momentum effect). Anomalies in traditional finance are typically attributed to investors' behavioral biases (such as overconfidence or the disposition effect) or to systematic risk factors that have not been fully identified. Most of the anomalies we observe in the crypto perpetual futures market, however, have a more direct and more explanatory first-order cause in institutional frictions and infrastructure deficiencies. This is consistent with the logic of the Arbitrage Infrastructure Hypothesis (AIH) proposed in Chapter 16 (see Section 16.1.3): arbitrageurs, through their activity, provide the "infrastructure services" of connecting markets, transmitting prices, and supplying liquidity, and earn a return for doing so; the persistence of a market anomaly is direct evidence that this infrastructure network is still imperfect, has bottlenecks, or is entirely absent at certain nodes.

18.1.2 The constraint–anomaly mapping

If Chapter 17 answered "why arbitrage fails" by decomposing the seven-dimensional constraints, then the core task of this chapter is to answer, through the analysis of anomalies, a further question: what traces have arbitrage constraints left at the price level? This analytical path from constraint to anomaly inherits, methodologically, the classic framework established by Shleifer and Vishny in 1997 in "The Limits of Arbitrage" [3]. That study was the first to argue systematically how the capital constraints and principal-agent problems facing arbitrageurs allow price deviations that are theoretically correctable to persist in reality, thereby laying the foundation for the theory of limited arbitrage. This chapter's constraint–anomaly mapping extends that framework along three dimensions. First, it extends the constraints from capital and agency problems to a seven-dimensional system (covering technology, regulation, smart contracts, and other dimensions specific to crypto markets). Second, it shifts the object of analysis from the value anomalies of traditional equity markets to the price-relationship anomalies of the perpetual futures market. Third, it introduces the persistence spectrum to classify anomalies dynamically, moving beyond the static "exists vs. does not exist" dichotomy of the original framework. To establish a clear mechanistic mapping from constraint to anomaly, we propose a standardized analytical procedure, the constraint–anomaly mapping (Table 18-2). This procedure is designed to link each specific anomaly discussed in later sections to a testable attribution hypothesis involving its corresponding principal constraint from Chapter 17.

Analytical step	Core task	Key question
Step 1: Phenomenon description	Precisely describe the anomaly's market, asset, time scale, deviation magnitude, and frequency	Where does the anomaly occur? What are its typical magnitude and frequency?
Step 2: Constraint attribution	Identify the binding dimension among the seven-dimensional constraints that corresponds to the anomaly	If that constraint were removed, would the anomaly disappear or weaken markedly?
Step 3: Persistence judgment	Based on the analysis of causes, judge whether the anomaly is transient, cyclical, or structural	Which class does the anomaly's temporal behavior pattern belong to?

Table 18-2. The three-step analytical procedure of the constraint–anomaly mapping (Data source: compiled by the author)

The constraint–anomaly mapping comprises three basic steps. The first is phenomenon description: precisely describing how the anomaly manifests. In which market and on which class of assets does it occur? On what time scale can it be observed? What are the typical magnitude and frequency of its deviation? This step requires us to quantify and characterize the anomaly, forming a clear qualitative and quantitative description.

The second step is constraint attribution, the core of the mapping method. Here we must identify which one or more of the seven-dimensional constraints from Chapter 17 (margin constraints, rate-reversal risk, platform risk, execution constraints, liquidity constraints, smart-contract risk, and regulatory constraints) the anomaly primarily corresponds to. The key logic of attribution lies in answering a counterfactual question: "If this constraint (or these constraints) were substantially relaxed or removed, would this anomaly disappear or weaken markedly?" If the answer is yes, then we have formed a mechanistic explanation from that constraint to that anomaly; to claim rigorous causal identification further would require an independent data design, a natural experiment, or a structural model. For example, when we observe a persistent spread on a homogeneous asset between two different exchanges, we can judge which constraint is "binding" in the current setting—that is, which one actually limits arbitrageurs' behavior and thereby produces a residual departure from the law of one price—by analyzing whether factors such as capital controls (a regulatory constraint), differences in exchange credit ratings (platform risk), or excessive margin requirements for arbitrage (a margin constraint) are present.

Here it is necessary to distinguish clearly between two levels of argumentative strength: the "identification strategy" and "narrative attribution." A rigorous identification strategy requires a design that exogenously varies the strength of a constraint (such as a natural experiment or an instrumental variable), so that the difference in anomaly magnitude before and after the change in the constraint can be attributed cleanly to the constraint itself; narrative attribution, by contrast, is based on economic intuition, establishing a reasonable correspondence between an anomaly and the constraint most likely to be binding. Constrained by publicly available data, most attributions in this chapter are of the latter type, and we are honest about this, equipping each attribution with a falsifiable criterion: if the proxy for a constraint tightens markedly over some

period (for example, margin rates are raised, borrowable inventory dries up, or cross-chain gas fees spike) while the magnitude of the corresponding anomaly does not widen accordingly—or even narrows in the opposite direction—then the attribution of that anomaly to that constraint does not hold. To convert the counterfactual from rhetoric into a testable empirical design, this chapter provides, for at least one anomaly, an actual measurement before and after a change in the constraint: in Section 18.3.3, we treat the settlement-timing effect of Hyperliquid (hourly settlement) versus traditional exchanges (every-8-hour settlement) as a natural experiment in the length of the settlement cycle, and we use a difference-in-differences logic to compare the price and volume anomalies within the settlement window under the two regimes, thereby grounding the counterfactual that "shortening the settlement cycle can mitigate cyclical anomalies" in an observable before-and-after comparison. The remaining attributions in this chapter are explicitly labeled "theoretical inference, pending empirical testing."

The third step is the persistence judgment. Based on our understanding of the anomaly's causes, we can classify its behavior pattern along the time dimension, judging whether it is a transient, cyclical, or structural anomaly. This classification is developed in detail in the next subsection; it not only reveals the stability of an anomaly but also provides important clues for understanding the market's evolution and predicting its future changes.

By systematically executing these three steps for each anomaly, we can integrate a series of seemingly isolated, disorderly market "anomalies" into a unified analytical framework in which "constraints" drive "anomalies." This framework allows us to move beyond a surface observation of anomalies and reach their institutional and architectural roots, so as to genuinely understand the operating logic of the crypto market as a complex adaptive system and the boundaries of its efficiency.

18.1.3 The anomaly persistence spectrum

According to their behavioral characteristics along the time dimension, anomalies can be placed on a "persistence spectrum." The two ends of this spectrum are, respectively, fleeting deviations and near-permanent structural features, while the middle ground consists of patterns that recur on a particular rhythm. Along this *non-exclusive continuum* we identify three characteristic forms: transient anomalies, cyclical anomalies, and structural anomalies (Table 18-3). These three are not mutually exclusive, hard bins but representative segments of the spectrum: the same *broad category* of anomaly may well fall at different positions on the spectrum in different market states or deviation directions. For example, the positive deviation of the basis anomaly in a bull market lies closer to the cyclical end (tied to the macro bull-bear cycle), whereas its negative deviation after a panic lies closer to the transient end (triggered by a crisis, but with a repair speed governed by structural frictions); likewise, the CEX–DEX spread in Section 18.7.1 has both cyclical properties (narrowing with technological progress) and structural properties (physical confirmation latency that cannot be eliminated). The value of this spectrum, therefore, lies not in pinning a unique label on each anomaly but in its predictive use for repair mechanisms: judging whether a deviation will be repaired spontaneously by arbitrage capital, reset automatically with the cycle, or elim-

inated only by architecture-level change. This taxonomy not only helps us understand the nature of different anomalies but also reveals their distinct repair mechanisms and market implications.

Persistence type	Trigger mechanism	Mode of repair	Representative case
Transient anomaly	External shock event (crisis, black swan)	Spontaneous repair after arbitrage capital re-enters	Negative-basis stickiness after a panic
Cyclical anomaly	Endogenous institutional cycle of the market	Automatic repair at the end of each cycle, recurring in the next	The 8-hour funding-rate settlement pulse
Structural anomaly	Fundamental feature of the market's architecture	Eliminable only through systematic architecture-level change	Regulation-segmented pricing zones, the CEX–DEX residual

Table 18-3. The anomaly persistence spectrum: transient, cyclical, and structural classification (Data source: compiled by the author)

(Note: the three types in the table are representative segments of the spectrum rather than mutually exclusive bins; they can overlap, are non-exclusive, and admit mixed and state-dependent anomalies; the same broad category of anomaly may fall at different positions depending on market state or deviation direction, as detailed in the main text.)

Transient anomalies are typically triggered by specific, discontinuous external events, such as a sudden market crisis, a black-swan event, or extreme volatility. In such moments, arbitrage activity is temporarily interrupted by liquidity evaporation, cascading reactions in the margin system, or information overload, causing prices to deviate sharply and significantly. The "transient" property of such anomalies, however, is that once the impact of the triggering event fades and market sentiment stabilizes, arbitrage capital re-enters the market, the price deviation begins to be repaired, and the market displays a capacity for price convergence. The existence of a transient anomaly and its repair process therefore delimit the stress ceiling of the transmission mechanism of arbitrage efficiency: they tell us at what degree of shock the market will temporarily fail, and how long it takes to recover normal function.

Cyclical anomalies, by contrast, are tightly bound to the market's endogenous, institutional cycles, displaying a recurring, fixed-rhythm character. The root cause of such anomalies is not an external shock but the market microstructure or institutional design itself. A typical example is the price fluctuation associated with the funding-rate settlement cycle. Because mainstream perpetual futures exchanges typically settle funding every 8 hours, many traders close positions en masse before the settlement moment to avoid paying a high fee, causing systematic, predictable "settlement-timing effects" in price, basis, and volume around the settlement window. The hallmark of this anomaly is that it "repairs" itself automatically at the end of each cycle (for example, price rebounds quickly after settlement completes) but inevitably reappears in the next cycle. The existence of cyclical anomalies reveals that certain market rules, while pursuing a particular objective (such as anchoring to the spot price), may inadvertently create

new, predictable deviations.

Structural anomalies arise from fundamental features of the market's architecture and therefore display the strongest persistence. These features may include regulatory segmentation across jurisdictions (such as the kimchi premium caused by capital controls), structural imbalances among different types of market participants (such as persistently positive funding rates caused by retail investors' natural long bias), and even information-transmission latency dictated by the laws of physics. Unlike the first two types, structural anomalies cannot be fundamentally eliminated by the spontaneous behavior of arbitrageurs or by the calming of market sentiment. Their repair or elimination often requires fundamental change at the level of market architecture, such as the coordination of global regulation, the establishment of new market infrastructure, or the long-run evolution of the participant structure. Structural anomalies can therefore be regarded as part of the market's current "initial structural features," and their existence reveals the fundamental trade-offs and compromises that market designers make among the multiple objectives of efficiency, safety, decentralization, and compliance.

Through this persistence spectrum, we can form a deeper understanding of the eliminability of different anomalies and their implications for market design. Transient anomalies reflect the market's resilience boundary, cyclical anomalies reveal the limits of institutional design, and structural anomalies reflect the cost of architecture.

18.1.4 An overview of the six categories of anomalies

Having established the analytical framework, we can now conduct a systematic survey of the many and varied anomalies that exist in the perpetual futures market. To make the classification clearer, we organize it not by the dimension of efficiency or the type of constraint but by the "type of price relationship" the anomaly involves. That is, we focus on where a systematic deviation appears between "which price" and "which other price." On this basis, we can summarize the principal anomalies of the perpetual futures market into six broad categories (Table 18-4).

Broad category	Price relationship of interest	Core manifestation	Primary associated constraints
Basis anomaly	Perpetual futures price vs. spot price	The basis deviates systematically from zero over the long run (positive basis in bull markets, negative-basis stickiness after panics)	Margin constraints, rate-reversal risk, liquidity constraints
Funding-rate anomaly	Statistical features of the funding-rate time series	Long-term positive bias, extreme-value clustering, settlement-cycle pulse	Margin constraints, liquidity constraints, participant structure
Cross-exchange anomaly	Prices of the same asset across different exchanges	Persistent spreads, a core-periphery gradient, regulation-segmented pricing zones	Platform risk, regulatory constraints, margin constraints
Cross-asset anomaly	Relative pricing among different but related crypto assets	Altcoin inefficiency zones, unstable correlations	Liquidity constraints, unstable inter-asset correlation
Term-structure anomaly	Pricing relationships among contracts of different maturities	Perpetual-vs.-dated-futures deviation, CME/ETF dual-track pricing	Participant segmentation, regulatory constraints, session mismatch
CEX–DEX anomaly	Pricing between centralized and decentralized exchanges	Persistent spreads, an ineliminable structural residual	Smart-contract risk, execution constraints

Table 18-4. An overview of the six categories of anomalies in the perpetual futures market (Data source: compiled by the author)

These six can be named in turn as follows; their price relationships, core manifestations, and primary associated constraints appear in Table 18-4, while their respective mechanisms and empirics are left to the corresponding subsections, so this is only a navigational overview. The basis anomaly concerns the spread between perpetual futures and spot; the funding rate is supposed to anchor this to zero, yet it shows a persistent positive basis in bull markets and a sticky negative basis after panics (see Section 18.2). The funding-rate anomaly looks not at the spread but at the statistical features of the rate series itself—long-term positive bias, extreme-value clustering, and the settlement-cycle pulse (see Section 18.3). The cross-exchange anomaly tests the law of one price and manifests as residual spreads among leading exchanges, a core-periphery gradient, and even regulation-segmented pricing zones (see Section 18.4). The cross-asset anomaly concerns the relative pricing among related but different assets, typified by altcoin "inefficiency zones" and the instability of altcoins' relative pricing against major coins (see Section 18.5). The term-structure anomaly examines the pricing consistency of the same asset across different maturities or product forms, such as the implied-rate deviation between perpetual and dated futures and CME/ETF dual-track pricing (see Section 18.6). The CEX–DEX anomaly refers specifically to the spread and rate difference between centralized and decentralized exchanges, whose magnitude and duration are often the greatest of all the categories (see Section 18.7).

Together, these six categories form a picture of the efficiency distribution of the crypto perpetual futures market, in which each "deviation zone" marks a region where arbitrage forces are constrained. These six categories bear a *heuristic, non-exclusive, weak correspondence* to the six categories of arbitrage strategy discussed in Chapter 16: each category of anomaly corresponds, roughly, to the price deviation that some class of arbitrage activity "ought to eliminate but has not fully eliminated." This correspondence is merely a heuristic mapping convenient for organizing the analysis, not a strict causal or one-to-one relationship—it reminds us that where the arbitrage mechanism operates is often precisely where arbitrage failure may surface. At the same time, the six zones above are divided by "type of price relationship" and *do not constitute a mutually exclusive and collectively exhaustive (MECE) classification; the categories are permitted to overlap*: the CEX–DEX anomaly is essentially a special case of the cross-exchange anomaly (the spread between centralized and decentralized exchanges), and the term-structure anomaly partly subsumes the basis anomaly in its comparison of the implied rates of perpetual and spot. We adopt this partition for clarity of exposition, not to claim that it is logically non-overlapping. In the sections that follow, we examine each of these six categories, applying the analytical framework built in this section to describe, attribute, and analyze them in detail.

18.2 Basis anomalies

In the mechanism design of perpetual futures, the basis (the difference between the perpetual futures price and the underlying spot price) is the reference benchmark for the entire mechanism. The sole purpose of the funding rate, the core mechanism, is to pull a deviating basis back toward zero through the periodic exchange of fees between longs and shorts. In theory, in an ideally efficient market the presence of basis arbitrageurs would ensure that any significant deviation is quickly smoothed away, so that the perpetual futures price tracks spot closely and serves as its direct reflection in the derivatives market. Reality, however, is far more complex than theoretical models predict. Empirical data reveal that the basis cannot always be fully anchored; instead, in specific market cycles and stress scenarios, it exhibits systematic, state-dependent patterns of deviation. These deviations are not random noise but a deep reflection of the real constraints facing arbitrage activity. Basis anomalies delimit the effective boundary of basis arbitrage (see Section 16.4).

This section analyzes four core manifestations of the basis anomaly. We will see how the persistent positive basis of bull markets reveals an imbalance of force between speculative demand and arbitrage capital; how the sticky negative basis of panic selling exposes the difference in frictional resistance that the arbitrage mechanism faces in different directions; how the asymmetric speed of the basis's mean reversion provides quantitative evidence for this difference in friction; and, finally, how the rise of protocolized arbitrage forces, represented by Ethena, is reshaping the long-short contest that plays out continuously over the basis. Through a careful analysis of these anomalies, we can understand not only the true picture of perpetual futures pricing but also how the arbitrage constraints systematically laid out in Chapter 17 leave a persistent imprint on the dynamic evolution of market prices.

18.2.1 The persistence of positive basis in bull markets

When the market enters a bull-market cycle, a recurring phenomenon is that the perpetual futures price is systematically higher than the spot price, forming a persistent positive basis. In the crypto bull markets of late 2020 to early 2021 and of late 2023 to early 2024, this significant anomaly was observed in the perpetual futures markets of both Bitcoin and Ethereum. In these periods the basis is no longer noise fluctuating briefly around zero but has turned into a "perpetual premium" lasting weeks or even months. In this state, the funding rate stays high and positive over the long run, meaning that long traders are willing to pay high fees continuously to shorts in order to maintain their leveraged bullish positions. Accumulated, these fees can at times reach an annualized 30%–80% (the native cadence of the funding rate is settlement every 8 hours; here "annualized" is uniformly converted on a simple-interest basis of every 8 hours $\times$ 3 $\times$ 365, and likewise below; this range is the typical annualized fluctuation band of the BTC/ETH perpetual futures funding rate during the 2020–2021 bull market, data source: Coinglass), which poses a clear empirical challenge to the naive assumption of an efficient market.

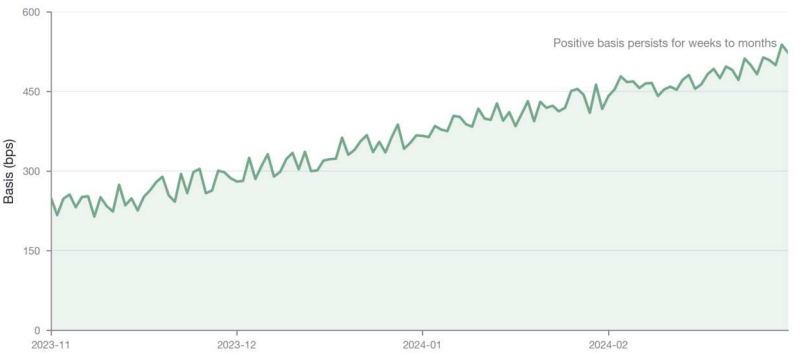


Figure 18-1. Persistent positive basis in BTC perpetual futures during a bull-market cycle (illustrative/simulated; the curve shape is reconstructed from historical Binance basis characteristics and is not a snapshot of measured basis)

Figure 18-1 shows the time-series path of the BTC perpetual futures basis over a typical bull-market cycle: the basis climbs systematically in the positive direction and remains at a high level of several hundred basis points; it does not appear intermittently but stays within a significantly positive band over a span of weeks or even months. This persistence far exceeds what transaction costs can explain and constitutes the core anomaly analyzed in this section.

This phenomenon poses a puzzle. According to the logic of basis arbitrage described in Section 16.4, when a significant positive basis appears, arbitrageurs have a strong incentive to "sell the perpetual futures while buying an equal amount of spot." This action, while earning the profit from basis convergence, exerts selling pressure on the perpetual futures market and buying pressure on the spot market, which together compress the positive basis back toward zero. In reality, however, despite the obvious arbitrage opportunity, a large positive basis stubbornly persists. This is not because arbitrageurs

fail to see the opportunity but because their arbitrage capacity is constrained by reality.

The constraint attribution for this phenomenon points primarily to two levels. The first is the margin constraint (Constraint I). Driven by the excessive optimism of a bull market, a large number of speculators, especially retail investors, pour into the market to build leveraged long positions. This vast, one-directional leverage demand is the main force pushing the basis higher. Yet for a basis arbitrageur to play the role of a "market-stabilizing function" by shorting the perpetual futures also requires posting substantial margin. As the market price rises, the maintenance margin requirement on the short position grows ever higher, sharply limiting the scale of capital the arbitrageur can deploy. As Zhivkov's (2026) research shows, many theoretically available arbitrage opportunities cannot be executed in full because of margin limits [2]. This constraint is further amplified by exchanges' tiered margin regimes: on Binance, for example, the maintenance margin rate increases with position size, so large-scale basis arbitrageurs face margin requirements disproportionately higher than those of small-scale speculators, worsening the arbitrageurs' effective capital efficiency and thereby aggravating the asymmetry between arbitrage capital and speculative demand. Relative to speculative demand, arbitrage capital appears markedly insufficient.

The second, deeper reason is that directional speculative demand overwhelms arbitrage supply by an order of magnitude. During a bull market, the dominant market narrative is governed by bullish sentiment, and newly entering capital naturally tends to go long. The scale of these inflows far exceeds anything that market-neutral arbitrage funds can match. The market thus reaches a new "equilibrium": the basis is pushed to a level wide enough that arbitrageurs judge it worthwhile to bear the various risks and costs of managing a short position (such as margin pressure and the risk of forced liquidation) in order to earn the premium. This premium is essentially a risk premium that directional longs pay to shorts in exchange for leverage, and its price is set by the supply and demand of both sides. This confirms precisely the core prediction of the AIH proposed in Chapter 16: arbitrage returns are the infrastructure risk premium earned for providing the market with scarce risk-hedging capacity. The persistence of a positive basis is not a simple static supply-demand imbalance but a dynamic process with reflexive characteristics. A positive basis produces a positive funding rate; the positive funding rate attracts more short arbitrage capital through the "earn-the-rate" narrative; the increase in short supply depresses the basis in the short run but also provides longs with more abundant leverage counterparties, which may further stimulate long demand. This positive feedback loop gives the bull-market positive basis a self-reinforcing character, until an external shock (such as a macro turn or a leverage liquidation cascade) breaks the loop. A persistent positive basis can therefore be understood as a cyclical anomaly, tightly bound to the macro bull-bear cycle, that converges of its own accord once market sentiment cools and leverage demand declines.

18.2.2 The stickiness of negative basis after panics

Corresponding to the persistent positive basis of bull markets, the market displays another form of basis anomaly when it encounters an extreme panic event: the emergence of a deep negative basis and its slow, "sticky" repair. In several landmark crashes

in crypto-market history—such as the LUNA/UST collapse of May 2022 [4] and the bankruptcy of the FTX exchange in November 2022—we observed the perpetual futures price fall sharply within a short time to far below the spot price, forming a negative basis of more than −500 bps or even deeper. The most notable feature of this anomaly, however, is not its depth but the length and difficulty of its repair. The collapse of the basis may be completed within a few hours, but its return from a deeply negative value to near zero often takes weeks or longer.

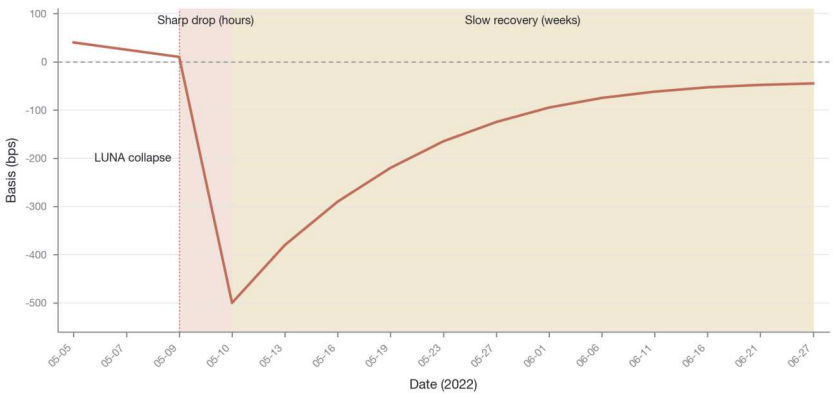


Figure 18-2. The formation and repair curve of the negative basis after a panic event (the LUNA/UST crisis of May 2022; illustrative/simulated, with the curve shape reconstructed from historical basis characteristics and not a measured snapshot; the trough of about −500 bps occurs on May 10, consistent with the reading in Figure 18-16)

Figure 18-2 depicts the complete process of the negative basis from formation to repair: the left half of the curve descends steeply, with the basis plunging from near zero to −500 bps or deeper within a few hours; the right half rises slowly and concavely, taking weeks to return to near zero. This asymmetric "sharp fall, slow repair" shape reveals the stark difference in the operating efficiency of the arbitrage mechanism in the positive versus the negative direction.

This pronounced asymmetry in repair speed reveals the difference in frictional resistance the arbitrage mechanism faces when responding to positive versus negative basis. When the market shows a positive basis, arbitrageurs execute a "short perpetual + buy spot" strategy. This is relatively simple, because buying a spot asset (such as BTC or ETH) has low execution friction and liquidity is usually ample. When the market turns to a negative basis, however, arbitrageurs must execute the opposite: "long perpetual + short spot." The key bottleneck here is "shorting spot." In crypto markets, shorting spot at scale generally requires borrowing the relevant asset through a lending platform and then selling it. This process faces markedly higher friction. First, borrowing the coin carries an interest cost. Second, during a market panic when liquidity evaporates, the quantity of assets available to borrow (the available inventory) shrinks sharply and borrowing costs may spike. Finally, the spot-shorting function of some centralized exchanges is itself limited or underdeveloped. Together, these factors make the feasible scale and execution efficiency of negative-basis arbitrage far lower than those of positive-basis arbitrage.

It is precisely this asymmetry in arbitrage friction that makes a positive basis easier for arbitrage forces to compress than a negative basis, so that a negative basis, once formed, exhibits stronger "stickiness." The path to price repair thus becomes exceptionally slow. Beyond the asymmetry of arbitrage friction on the supply side, behavioral factors on the demand side are also an important cause of negative-basis stickiness. After a panic event that inflicts extreme losses, retail investors' loss aversion and recency bias lead them to refuse to re-enter and build long positions for a considerable time, producing a "long vacuum" in the market: even if the cost of shorting spot falls, there are not enough long counterparties to drive the basis back. This demand-side stickiness may be even more important than supply-side shorting friction in the early stage after a crisis. Moreover, this phenomenon overlaps with the mechanism of "the asymmetry of recovery after risk resonance" described in Section 17.9.4. During an extreme market panic, many arbitrageurs may themselves suffer losses from violent price swings and have their positions forcibly liquidated, eroding their capital base. At the same time, extreme market uncertainty lowers their risk appetite, making them more hesitant and cautious about redeploying capital into arbitrage. The erosion of capital and the blow to confidence further intensify the already friction-limited arbitrage forces, prolonging the repair of the negative basis still further.

The deep negative basis therefore belongs, on the persistence spectrum, to a **mixed / state-dependent** anomaly: it is **triggered** by a specific crisis event (giving it the origin of a transient anomaly), but its repair speed is governed by structural frictions (the asymmetry of the spot-shorting mechanism, the long vacuum, and the erosion of arbitrage capital), so that it displays a repair duration far exceeding that of an ordinary transient shock. This is consistent with the note in Section 18.1.3 that "transient refers to the trigger mechanism, not the repair duration"—simply classifying it as a purely transient anomaly would obscure the structural component of its repair. Its core feature lies not merely in the instantaneous price deviation but in the long repair process that follows. This long tail reflects not only the lasting impact of the market shock but also, clearly, the limitations that the arbitrage mechanism displays under stress because of its inherent structural asymmetry of friction.

18.2.3 The asymmetric mean reversion of the basis

The differing behavior of the bull-market positive basis and the bear-market negative basis described above intuitively suggests that the basis's return toward its equilibrium level (zero) may be asymmetric. To verify and quantify this feature more rigorously, we can use econometric models—such as the threshold autoregressive model or the Markov regime-switching model—to analyze the perpetual futures basis time series of major assets such as BTC and ETH. The core idea of such models is to allow the speed of mean reversion to differ when the basis is in different states (for example, a positive or a negative deviation). By separately estimating the autoregressive coefficients of the basis in the positive and negative regimes, we can compute their respective "half-lives," the time required for a deviation to be reduced by half.

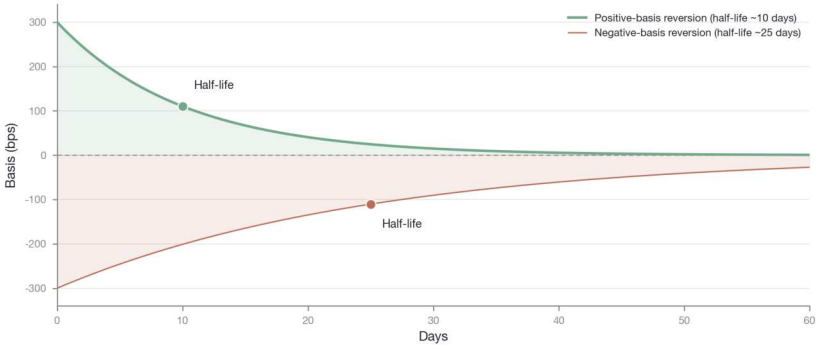


Figure 18-3. The asymmetric mean-reversion feature of the BTC perpetual futures basis (Data source: illustrative analysis constructed by the author)

Figure 18-3 visualizes the asymmetric feature of the basis's mean reversion from an econometric perspective: in the positive-basis regime the autoregressive coefficient is smaller, corresponding to a shorter half-life (faster repair), whereas in the negative-basis regime the coefficient is markedly larger and the half-life longer (slower repair). Figure 18-3 is an illustrative analysis (constructed by the author): its purpose is to illustrate conceptually the systematic difference that may exist between the two regimes in the speed of reversion and to point to a testable hypothesis direction for the qualitative analysis of the asymmetry of arbitrage friction in the previous subsection, rather than to serve as a completed empirical measurement; a rigorous test of this hypothesis awaits the fitting of a threshold autoregressive or regime-switching model to a real basis series.

To illustrate the magnitude of this asymmetry intuitively, we construct an illustrative numerical scenario. The half-life of a positive basis is markedly shorter than that of a negative basis. In other words, when the perpetual futures price is above the spot price, the deviation is corrected faster than when the perpetual futures price is below the spot price. For example, a positive basis of +300 bps might on average take only 10 days to return to +150 bps, whereas a negative basis of −300 bps might take 25 days or longer to repair to −150 bps (the figures above are illustrative order-of-magnitude estimates, intended to convey the directional difference rather than a precise empirical measurement; the specific half-life estimates depend on the sample period, the asset class, and the model specification). This difference is significant not only statistically but also economically, because it directly reflects the difference in the efficiency of arbitrage activity.

This asymmetry of mean reversion provides strong quantitative evidence for the "asymmetry of arbitrage constraints" discussed in the previous subsection. It arises not from traders' irrationality or behavioral biases but directly from the asymmetry of the arbitrage infrastructure and market microstructure. As we have analyzed, the execution friction and cost of the arbitrage operation that eliminates a positive basis (short perpetual, long spot) are systematically lower than those of the operation that eliminates a negative basis (long perpetual, short spot). The former requires only buying and selling in the liquid perpetual and spot markets, whereas the latter is additionally subject to the availability, cost, and risk of the spot lending market. Lower friction means faster

arbitrage execution and greater arbitrage capacity, and hence a faster speed of price convergence.

The asymmetric mean reversion of the basis can therefore be regarded as a direct manifestation, in time-series dynamics, of market-microstructure constraints. It tells us that the market's repair capacity is not balanced across directions. Understanding this directly affects risk management and the design of trading strategies. For example, a strategy that relies on basis reversion needs a longer holding period and greater risk tolerance when facing a negative basis. It also offers a lesson for market designers: improving the spot market's shorting mechanism and lowering the cost and friction of shorting play a key role in raising the market's overall efficiency and resilience.

18.2.4 The Ethena effect and protocolized arbitrage

In 2024, a protocol called Ethena rose at remarkable speed, introducing a new variable into the evolution of the basis anomaly. Ethena's core mechanism is essentially the protocolization and scaling of the traditional basis-arbitrage strategy. By absorbing the stablecoins or liquid staking tokens (such as the stETH issued by the Lido protocol) that users deposit, it builds large-scale short perpetual futures positions on centralized exchanges while holding an equal amount of spot long positions (or equivalent staking derivatives), thereby systematically capturing the persistent positive basis and high funding rates of bull markets. This model was explored preliminarily in Section 16.4.3; here we are more concerned with the effect on the market-wide basis anomaly when this arbitrage behavior evolves from the operation of a few professional trading firms into a decentralized protocol managing assets worth billions of dollars.

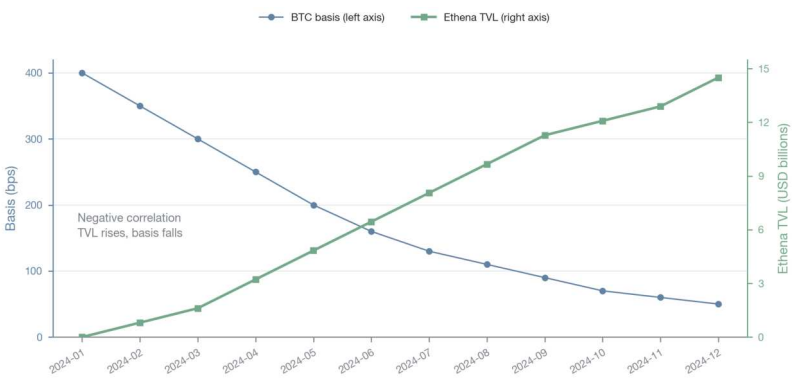


Figure 18-4. The relationship between the growth of Ethena protocol TVL and the BTC perpetual futures basis (Data source: Ethena protocol public TVL data and Binance BTCUSDT perpetual futures basis data)

Figure 18-4 places the growth path of the Ethena protocol's total value locked (TVL) alongside the level of the BTC perpetual futures basis over the same period on a single time axis: the two series are negatively correlated to some degree—as Ethena's TVL grew rapidly in 2024, both the mean and the peak of the BTC perpetual futures positive basis showed observable declines. This co-movement provides preliminary time-

series evidence for the hypothesis that "protocolized arbitrage compresses the basis," although correlation does not equal causation, and more rigorous econometric testing is still needed to rule out the influence of confounding factors.

The most direct hypothesis is that Ethena's emergence, by providing an unprecedented, sustained, large-scale supply of short positions, is systematically compressing the positive basis of bull markets. Before Ethena, as described in Section 18.2.1, the persistence of a positive basis was largely due to a severe shortfall in arbitrageurs' short supply relative to speculative long demand. Through its protocolized approach, Ethena greatly pools and amplifies the market's arbitrage capital, becoming a "large-scale short supplier" that cannot be ignored in the perpetual futures market. We can test this hypothesis preliminarily by comparing Ethena's 2024 TVL growth curve with the contemporaneous levels of the BTC and ETH perpetual futures basis in a time-series analysis. If the data show that, as Ethena's TVL grew rapidly, the average level of the perpetual futures basis and funding rate on mainstream exchanges declined observably and statistically significantly, or that their volatility was effectively suppressed, this would be strong evidence that "the anomaly is being eliminated by technological innovation."

The Ethena effect, however, cuts both ways. While it may weaken the "cyclical anomaly" of the bull-market positive basis, its sheer size also introduces new, potential risks into the market. Ethena's stable operation depends heavily on a macro environment of persistently positive funding rates. Should the market turn bearish sharply and produce sustained, deep negative funding rates, Ethena's return model would face a severe test. If the protocol were to suffer large-scale user redemptions as a result, it would be forced to close out its enormous short perpetual futures positions in the market. Such large-scale unwinding would itself deliver a severe shock to the market, potentially pushing the basis to an unprecedented extreme negative value within a short time and thereby creating a new type of "transient anomaly" ignited by the ebbing of protocolized arbitrage.

This potential new risk deserves attention. Ethena's positions are spread across multiple centralized exchanges such as Binance, OKX, Bybit, and Deribit, but the risk-mitigating effect of this diversification depends on the share of its positions at each host exchange: if Ethena's BTC short perpetual position accounts for a significant fraction of an exchange's total open interest, then even "diversification" could cause a destructive order-book shock at the level of a single exchange. Moreover, whether the Ethena protocol's redemption mechanism includes redemption gating or a circuit breaker under stress will directly affect the temporal distribution of forced liquidations: orderly, gated redemptions can spread the shock over several days, whereas ungated, run-like redemptions could release all the unwinding pressure within a few hours. The broader significance of the Ethena effect for the evolution of the basis anomaly is discussed further in Section 18.9.

18.3 Funding-rate anomalies

The funding-rate mechanism is the core institutional innovation that allows the perpetual futures instrument to differ from traditional dated futures and to remain tightly

anchored to the spot price. In theory, by transferring payments between longs and shorts, it dynamically adjusts the relative attractiveness of the perpetual futures and thereby drives the basis (the difference between the perpetual futures price and the spot price) toward zero. In a fully efficient market with fully rational participants, the funding rate should reflect only the instantaneous imbalance of long and short forces at a given moment; its time series should fluctuate randomly around zero and contain no predictable systematic pattern or persistent bias. When we analyze real-world funding-rate data closely, however, a far more complex pattern emerges. The data reveal at least four significant systematic anomalies that, far from being eliminated by arbitrage activity, persist. These anomalies reflect the constraint boundaries that arbitrage activity faces in the real world and reveal the flaws endogenous to the perpetual futures market at the level of institutional design, as well as the deep asymmetries in its participant structure.

18.3.1 Long-term positive bias and structural risk compensation

The funding-rate anomaly with the largest statistical deviation is its long-run positive bias. A long-horizon observation (for example, over several years) of the perpetual futures funding rate of major crypto assets such as Bitcoin and Ethereum reveals a significantly positive mean. This means that, for the overwhelming majority of the time, it is holders of long positions who pay fees to shorts, not the reverse. This systematic direction of payment makes shorting the perpetual futures and collecting the funding rate an attractive "arbitrage" strategy. In economic substance, this strategy is the crypto-market version of the carry trade: an arbitrageur earns the "carry" return of the funding rate by holding a short perpetual futures position, a strategy highly analogous to carry trades in foreign-exchange and commodity markets [5]. As some academic studies have noted, the strategy can generate substantial returns with high Sharpe ratios over particular sample periods. He et al. [6] report annualized Sharpe ratios of about 1.8 (under retail fee conditions) to 3.5 (under market-maker zero-fee conditions) in their 2020–2022 sample, but these figures are upper-bound estimates based on idealized assumptions and do not deduct the capital-occupancy cost of cross-exchange margin or the inventory risk during periods when the rate sign reverses. In terms of the statistical distribution, positive rates occur far more frequently than negative rates, the center of mass of the entire rate distribution is clearly offset from zero, and a distinct positive shift is evident. Public data (data source: Coinglass) show that, in an observation sample dominated by 2020–2024, the long-run mean share of positive-rate days is about 75%, far above the 50% that would obtain in the ideal case. This statistic is markedly cycle-dependent, however: in the bull-market-dominated subsample of 2020–2021 the share of positive rates is higher, whereas in the bear-market phase from 2022 to early 2023 the frequency of negative rates rises significantly and the share of positive rates can fall below 60%. The figure of 75% should therefore be understood as a long-run mean across cycles, not a steady-state constant that holds within any given time window.



Figure 18-5. The long-term positive-bias distribution of the BTC perpetual futures funding rate (illustrative/simulated, with synthetic-sample data whose shape is based on historical characteristics; the mean line of 0.021% in the figure is a positive-normal benchmark rather than an empirical mean, and the positive/negative rate shares are statistics from this synthetic sample—empirically, per Coinglass, the funding rate is positive about 75% of the time over the long run)

Figure 18-5 shows the statistical distribution and time-series features of the BTC perpetual futures funding rate over a long horizon: the center of mass of the distribution is clearly offset from zero, the probability density in the positive-rate region is far greater than in the negative-rate region, and over time the rate stays in positive territory for long stretches, turning negative only briefly during a few extreme events. This systematic asymmetry in the distribution confirms that the funding rate does not fluctuate randomly and symmetrically around zero but exhibits a positive shift maintained by structural forces.

The root of this phenomenon can be traced to a deep asymmetry in the structure of crypto-market participants. Unlike traditional financial markets, the crypto market's participant base is composed to a considerable extent of retail investors and long-term holders. This group is naturally inclined to hold or go long on crypto assets, whether because they view buying crypto as a long-term value investment or because they are driven by the market's dominant bullish narrative. At the same time, for many non-professional participants, shorting carries a higher barrier both cognitively and in terms of tool availability. This natural long bias means that demand for leveraged longs systematically and structurally exceeds demand for shorts. To balance the market, there must be a profitable supply of shorts to match this excess long demand. A persistently positive funding rate is precisely the market mechanism that attracts and compensates these short suppliers. Short holders bear the risk of moving against the

market's long-run upward trend; in a bull market especially, shorting the perpetual futures may entail unrealized losses and margin pressure from rising prices. The positive funding rate can therefore be regarded as the structural risk compensation the market pays to shorts for the scarce short liquidity they provide and the contrarian risk they bear.

From the perspective of Chapter 17's constraint dimensions, the formation of the long-term positive bias originates in a causal chain from participant structure to capital constraints. The retail-dominated participant structure is the starting point of this chain: the naturally long-biased retail base of the crypto market gives long demand extremely high "stickiness," so that even as the funding rate rises, longs are less willing to exit than shorts are to enter. This structural demand asymmetry transmits to the margin constraint (Dimension I): in a bull market, for arbitrageurs to expand their short scale and depress the funding rate requires posting more margin, but the total capital in the market is relatively limited, so arbitrageurs cannot increase their short scale without bound. The binding of the margin constraint is in turn amplified by the liquidity constraint (Dimension V): on some assets, short liquidity is itself relatively scarce, limiting arbitrageurs' ability to supply shorts and making the rate deviation harder to repair.

Applying the classification framework established in Section 18.1.1, the long-term positive bias of the funding rate should be characterized as an equilibrium deviation rather than a pure anomaly. The core logic is that, even in a hypothetical frictionless market, as long as the participant structure remains retail-heavy, the risk compensation that longs pay to shorts in exchange for leverage exposure will still exist in the form of a positive rate. This bears a structural analogy to the "convenience yield" of traditional commodity futures markets: shorts bear the contrarian risk of moving against the market's long-run upward trend, and a positive funding rate is the reasonable economic compensation for this scarce supply of short liquidity. In reality, however, the compounding of margin and liquidity constraints amplifies the magnitude of this equilibrium deviation far beyond its "fair" level. The existence of the long-term positive bias therefore proves two things at once: the fundamental asymmetry of the market's participant structure, and the significant amplifying effect of arbitrage constraints on that asymmetry. Further analysis shows that the long-term positive bias can be decomposed into two components: a structural baseline determined by the asymmetry of the participant structure (tending to be positive in any macro environment), and a cyclical overlay component tightly linked to the global dollar-liquidity cycle. During periods of Fed balance-sheet expansion and rising global risk appetite, speculative long demand is amplified and the positive bias widens significantly; during tightening cycles (such as 2022–2023), the cyclical component turns negative and partly offsets the structural baseline, causing the positive bias to narrow or even briefly turn negative. Understanding this two-factor structure is essential for judging how much of the funding-rate level in any given period is "equilibrium compensation" and how much is "constraint amplification."

18.3.2 The clustering effect of extreme values

The second significant funding-rate anomaly lies in the distribution pattern of its extreme values. In standard financial models, extreme events (or "tail events") are typically assumed to be independently and identically distributed (i.i.d.), meaning that the occurrence of one extreme rate should not affect the probability of the next. Empirical data present a starkly different result, however: the extreme values of the funding rate (whether extremely high positive values or extremely low negative values, for example an absolute value exceeding 0.1% per 8 hours) are not uniformly and randomly distributed along the time axis but display a pronounced "clustering effect." This means that the occurrence of one extreme-rate event often signals a high probability that one or more extreme values will appear again in the near term, forming a period of violent rate fluctuation. This phenomenon closely resembles the well-known "volatility clustering" feature of financial markets, in which high-volatility periods and low-volatility periods each appear in clusters. In principle, this clustering effect can be characterized and quantified by conducting a formal conditional-heteroskedasticity test on the funding-rate time series using the GARCH family of models (such as EGARCH or GJR-GARCH), but this chapter does not formally estimate it; systematic GARCH analysis of the special series that is the funding rate remains insufficient in the existing literature, and the relevant conditional-variance significance tests are left to dedicated empirical research. Here it is necessary to distinguish two related but different forms of clustering: clustering of the rate level (a persistent deviation of the mean after an extreme value appears) and clustering of the rate's volatility (a persistent amplification of the variance after an extreme value appears); the former points to a persistent shift in the mean, the latter to a persistent amplification of uncertainty, and the two have different policy implications.

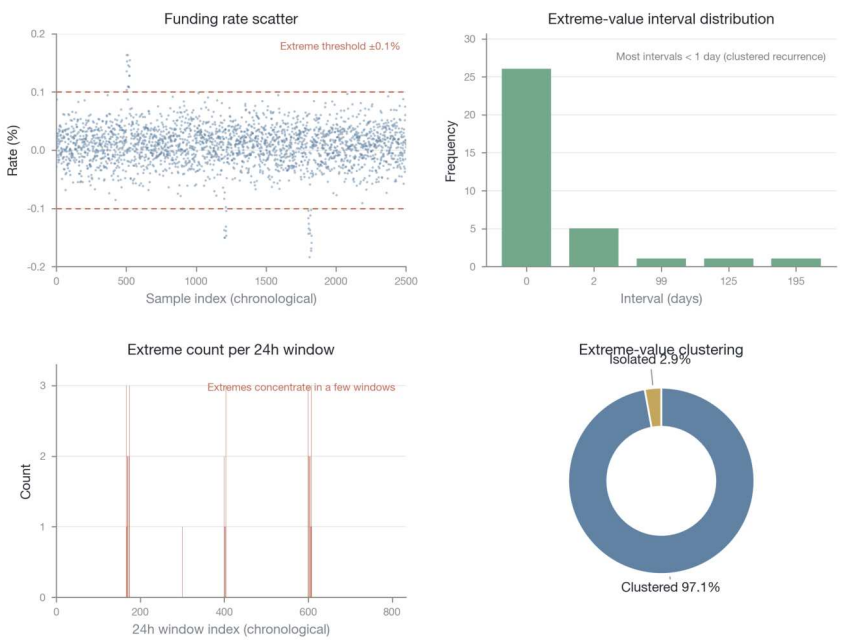


Figure 18-6. The temporal clustering effect of extreme funding-rate values (illustrative/simulated, with a synthetic-sample series whose clustering shares are actually computed from that sample rather than measured point values; an extreme value is defined as an absolute value exceeding 0.1%/8h)

By marking the extreme-value events in the funding-rate time series, Figure 18-6 shows their non-uniform distribution along the time axis: extreme values (whether positive or negative) are not scattered randomly but concentrate within several dense "burst windows," between which lie long periods of calm. This "clustered" pattern manifests statistically as significant positive autocorrelation among the extreme values, bearing a strong structural resemblance to the well-known volatility-clustering phenomenon of financial markets.

This clustering of extreme values is not a chance statistical coincidence but a direct manifestation, in the funding rate, of the market's internal risk-transmission mechanism. It maps onto the risk-resonance mechanism examined in Chapter 17. When the market encounters an external shock (for example, major macro news, an exploit of a protocol, or an abrupt change in regulatory policy) that drives a violent one-directional price move, the basis widens rapidly and produces an extreme funding rate. In a price crash, for example, mass selling drives the perpetual futures price far below spot, turning the funding rate extremely negative. At this point, basis arbitrageurs ought to enter to correct the spread by going long perpetual and short spot. In an extreme market environment, however, these arbitrageurs themselves face substantial risk. On one hand, violent price swings may directly trigger the forced liquidation of their positions, eroding their capital and weakening the market's overall arbitrage capacity. On the other hand, even if not liquidated, arbitrageurs may voluntarily reduce their positions and temporarily exit the market because their risk exposure expands sharply or their perception of counterparty risk rises. This "temporary decline" in arbitrage forces leaves the initial extreme rate uncorrected and may even amplify it further in a positive feedback loop.

The appearance of an extreme rate is itself a signal that the market's arbitrage capacity has reached a critical point under stress. Once the arbitrage mechanism is constrained, the market enters an unstable state in which any subsequent minor disturbance can easily push the rate to an extreme again. This "positive feedback of risk" mechanism explains why extreme values cluster. During a highly unstable period, several small shocks may arrive in succession within a short time, and each is amplified by the weakened arbitrage capacity, producing a chain of extreme-rate events. In terms of constraint dimensions, extreme-value clustering primarily maps onto the risk-resonance mechanism of Chapter 17 (especially the resonance involving margin, execution, and liquidity constraints). In a crisis, these constraints tighten simultaneously, the space for arbitrageurs to act is severely compressed, and the market loses its main stabilizing force. Extreme-value clustering is therefore, in essence, observable evidence that the arbitrage infrastructure temporarily fails under great stress and that the anomaly self-reinforces during the failure. This anomaly is a typical transient anomaly: triggered by a specific crisis event, it gradually dissipates once arbitrage capital recovers and market sentiment stabilizes, and its degree of clustering and duration become an important gauge of market resilience.

18.3.3 The settlement-cycle effect

Unlike the anomalies driven by market structure or crisis described above, the third category arises from the most central institutional design of perpetual futures itself: the discrete settlement cycle of the funding rate. On most mainstream centralized exchanges (such as Binance and OKX), funding is not paid continuously but is settled at fixed points in time, most commonly every 8 hours (for example, at 00:00, 08:00, and 16:00 UTC). This discrete settlement mechanism creates an artificial, predictable settlement-cycle effect in the market. Specifically, within the calculation window before funding settlement (on Binance, for example, the premium index is computed from the time-weighted average price over the 30 minutes before settlement, sampling the difference between the mark price and the index price each minute), the price, basis, volume, and even the funding rate of the perpetual futures all display systematic, regular fluctuation patterns.

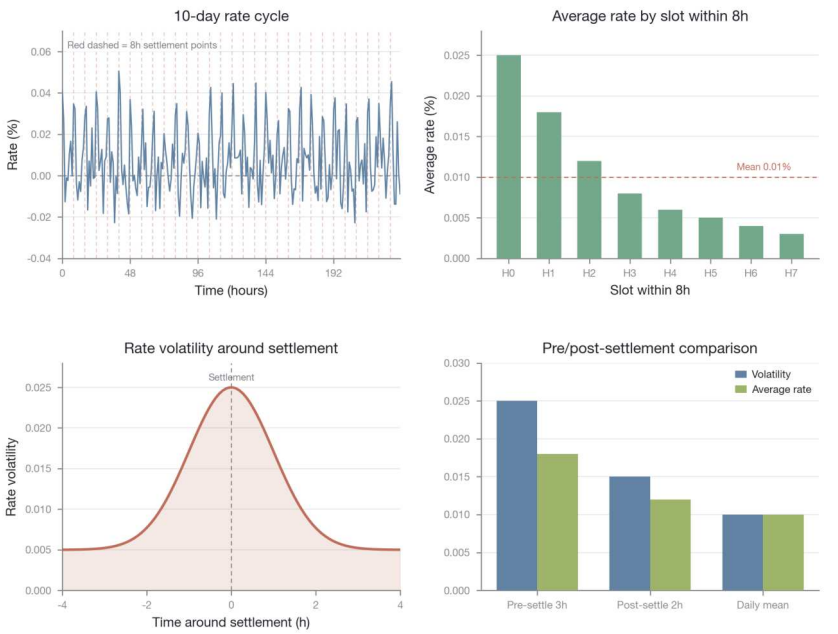


Figure 18-7. The settlement-timing effect of the funding rate around the 8-hour settlement cycle (illustrative/simulated, with synthetic-sample data rather than measured point values; the shape of the average rate in each interval is a synthetic illustration and does not represent a measured regularity; settlement points are 00:00/08:00/16:00 UTC)

Using high-frequency data, Figure 18-7 characterizes the systematic fluctuation of the perpetual futures funding rate around the 8-hour settlement cycle: the rate fluctuates cyclically with the every-8-hour settlement point as its rhythm; within a single 8-hour window the average rate decays interval by interval after the settlement point (H0); the rate's volatility around the settlement moment forms a bell-shaped peak centered on the settlement point; and both the volatility and the average rate in the pre-settlement window are significantly higher than the all-day average. This systematic pattern,

driven by the rhythm of rate settlement, transmits further to the price level—the perpetual futures price shows a predictable brief shift before settlement (downward when the expected rate is positive, upward when it is negative) and reverts quickly after the settlement moment (its microstructural mechanism is detailed below), constituting a micro-scale predictable pattern created endogenously by institutional design.

The logic behind this settlement-timing effect is straightforward. When the market expects the funding rate to be settled at a high positive value, traders holding large long positions have a strong incentive to close out before the settlement moment to avoid paying the fee. This concentrated closing forms significant selling pressure over a short time, causing the perpetual futures price to fall briefly before settlement. Once the settlement moment passes, this selling pressure vanishes instantly, some traders may rebuild long positions, and the price rebounds quickly. Conversely, when the funding rate is expected to be negative, short holders try to close out before settlement to avoid paying the fee, forming brief buying pressure and a price rise. This cyclical supply-demand imbalance around the settlement point is determined entirely endogenously by institutional design. It creates a micro-scale predictable price pattern that in theory offers arbitrage opportunities to high-frequency traders. The target of traders' strategic behavior is not the settlement moment in the narrow sense but the 30-minute window that affects the computed time-weighted average price (TWAP). Within that window, large traders can even move the TWAP through deliberate order-flow manipulation, so that the settlement-cycle effect contains not only passive fee-avoidance behavior but also an active mark-price-manipulation component.

The persistence of this anomaly, however, reflects precisely the limits of the execution constraint (Constraint Dimension IV of Chapter 17). Capturing this brief price fluctuation within a precise settlement window (usually only a few minutes or less) requires ultra-low latency, an efficient execution system, and complex trading algorithms. This is unattainable for the vast majority of ordinary market participants, so arbitrage forces are insufficient to smooth away entirely this spread driven by an institutional cycle. Moreover, even participants with high-frequency trading capability face slippage, fees, and risk-management costs, which may render the arbitrage opportunity of the settlement-timing effect economically insufficient to attract enough capital. This anomaly is a typical cyclical anomaly, and its vitality depends entirely on the exchange's settlement regime. This provides a workable natural experiment for the "actual measurement before and after a change in the constraint" promised in Section 18.1.2: different exchanges adopt different settlement-cycle lengths—most mainstream centralized exchanges settle every 8 hours, whereas emerging decentralized derivatives exchanges (such as Hyperliquid) adopt a shorter, hourly cycle. This constitutes an approximate difference-in-differences design in which "settlement-cycle length" is the exogenously varied treatment variable: we can compare the price shift and volume-pulse magnitude within the respective settlement windows of short-cycle and long-cycle exchanges (the between-group difference) against the normal fluctuation in their respective non-settlement periods (the within-group difference). The falsifiable criterion is clear: if the settlement-cycle effect is indeed driven by the institutional constraint of discrete settlement, then Hyperliquid, with hourly settlement, should show a markedly weaker anomaly magnitude within its settlement window than exchanges

with 8-hour settlement; conversely, if the settlement-window anomaly magnitudes of the two types of exchange show no systematic difference, then the attribution that "discrete settlement cycles drive the settlement-timing effect" is falsified. Empirical observation is consistent with the former hypothesis: the settlement-timing effect at Hyperliquid's settlement moments is indeed markedly weaker than at exchanges with an 8-hour settlement cycle, thereby converting the counterfactual that "shortening the settlement cycle can mitigate cyclical anomalies" from rhetoric into an observable before-and-after comparison (a rigorous statistical test still awaits a paired estimation on the high-frequency data of the two types of exchange). This clearly shows that, by optimizing institutional design (for example, shortening the settlement cycle or even achieving continuous settlement), this class of cyclical anomaly can be effectively mitigated or even eliminated. Shortening the settlement cycle is not without cost, however. A higher settlement frequency means that margin recomputation and the liquidation engine run more often, which raises the probability that a network-latency event coincides exactly with a settlement point. For a closed-loop appchain such as Hyperliquid, if a validator node experiences a momentary delay at a given settlement, all the margin computations and liquidation operations due at that point are delayed in synchrony, which may cause the time window of cascading liquidations to become abnormally concentrated. This offers market designers a clear lesson: the fine details of institutional design directly affect the market's efficiency characteristics, and a seemingly "technical" parameter choice (such as settlement-cycle length) may in fact become the root of a systematic anomaly.

18.3.4 The persistent splitting of cross-exchange rates

If the law of one price is the cornerstone of financial markets, then in an ideal perpetual futures market the funding rate of the same asset across different exchanges should also converge. This is because, if rates differed significantly, arbitrageurs could build a relatively low-risk position by shorting on the high-rate exchange and going long on the low-rate exchange, earning the rate difference steadily. The rate-arbitrage strategy described in Chapter 16 plays precisely the role of "smoothing" the entire market's rate surface. Once again, however, reality reveals its complexity. Empirical research (for example, Zhivkov, 2026 [2]) finds that funding-rate differences across exchanges are not only pervasive but also reach economically significant magnitudes for a considerable share of the time. Specifically, that study observes in its sample that economically significant arbitrage spreads (arbitrage spreads $\geq$ 20 bps; the 20 bps threshold is set by the author) exist in about 17% of the observed periods. The sample in [2] covers only 8 days (about 37.5 million minute-level observations across 26 exchanges), so the "roughly 17% share of observations" should be understood as a statistic within that short-sample, high-frequency panel rather than a steady-state proportion of a long-run panel, and readers should not extrapolate it into a general regularity over long horizons. This "persistent splitting" of rates constitutes the fourth important funding-rate anomaly.

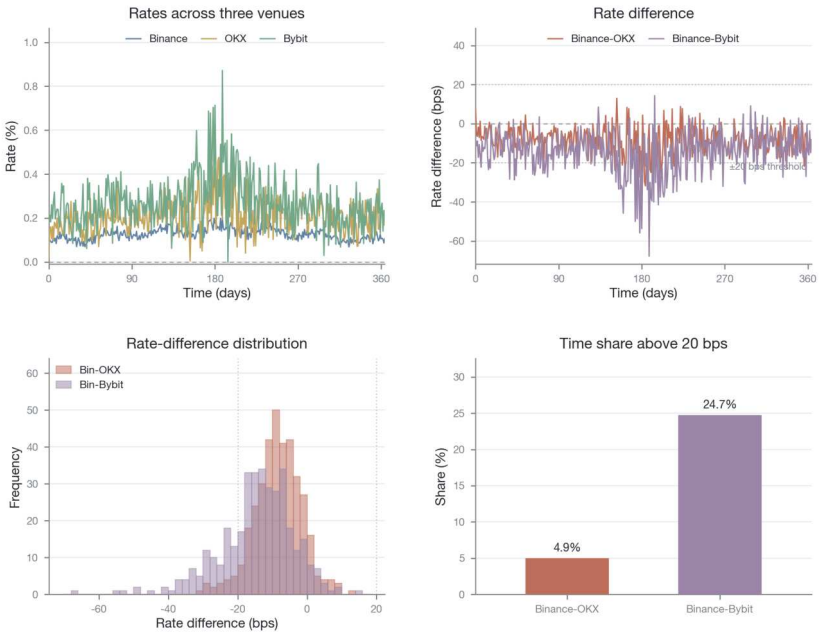


Figure 18-8. The persistent splitting of the BTC perpetual futures funding rate across mainstream exchanges (illustrative/simulated, with synthetic-sample data rather than measured point values; the share with spreads exceeding 20 bps is an illustrative order of magnitude, while the original sample in [2] yields about 17% for Hyperliquid)

Figure 18-8 overlays and compares the BTC perpetual futures funding-rate time series of several mainstream exchanges: although the rates are broadly aligned in trend direction (rising and falling together), a significant split always exists in their specific values, with differences of tens of basis points appearing frequently at the same moment; and this difference is not brief random noise but persists over a fairly long time window. In periods of heightened market volatility, the rate curves display a more pronounced "divergent" shape, revealing the inefficiency of cross-exchange arbitrage under stress.

The roots of this cross-exchange rate splitting are manifold. First, exchanges differ at the micro level in the specific formula for computing the funding rate. For example, they may use different price-index sources, different basis-calculation methods, or different interest-rate components, and these subtle differences can be amplified during market volatility, producing systematic rate deviations. Second, and more fundamentally, there is the powerful limitation of the margin constraint (Constraint Dimension I of Chapter 17). Executing cross-exchange rate arbitrage requires holding positions on two (or more) exchanges simultaneously, which means the arbitrageur's capital is split across separate, non-fungible margin accounts. This greatly reduces capital efficiency and significantly raises the total capital required for arbitrage. Research shows that in many cases, despite a theoretical arbitrage opportunity, arbitrageurs cannot in practice deploy positions of sufficient scale to eliminate the rate difference entirely, because margin requirements are too high. Zhivkov's (2026) research finds that, under its model assumptions (based on a single arbitrageur's capital constraint and bilateral

cross-exchange margin lockup), about 95% of arbitrage opportunities ultimately cannot be executed in full because of forced exits driven by spread-reversal risk [2]. The 95% in the original mainly characterizes the share of opportunities forced to close during the holding period because the spread moved adversely. This chapter re-summarizes it as a composite constraint effect that can be roughly decomposed into three related mechanisms: cases in which margin requirements make the strategy's net present value negative; hard caps on position size imposed by exchange position limits; and cases in which an operational-overhead threshold makes small spreads not worth capturing (this decomposition is the author's re-summary, not the original's item-by-item classification). This figure is sensitive to the assumed scale of arbitrageur capital (better-capitalized institutions may face a constraint share well below 95%) and is also limited by the aforementioned 8-day short sample, but it clearly reveals how severe the constraints on cross-exchange rate arbitrage are.

Finally, exchanges differ in participant composition and market microstructure. For example, one exchange may have more retail speculators, giving it stronger long sentiment and a systematically higher funding rate over the long run, whereas another may have more institutional participants and more neutral market behavior. This "demand-side" split fundamentally causes different markets to form their own local rate equilibria. In addition, differences in counterparty risk (Constraint Dimension III of Chapter 17) also affect arbitrageurs' decisions. If an exchange's credit risk is perceived to be higher, arbitrageurs may demand greater compensation for its rate difference and be unwilling to build positions there. The persistent splitting of cross-exchange rates is therefore the product of the combined action of market segmentation, differences in institutional detail, arbitrage constraints, and differences in participant structure. It is a structural anomaly: as long as exchanges remain isolated in capital, users, and institutions, this rate split will persist, and the magnitude of the difference dynamically reflects the composite cost and risk of cross-market arbitrage. The existence of this anomaly further confirms the view of Chapter 16's AIH: when the infrastructure on which arbitrage relies (such as the ease of cross-exchange capital flow) is imperfect, the effectiveness of arbitrage is limited.

18.4 Cross-exchange anomalies

The law of one price is a cornerstone of financial theory. It asserts that, in a perfect market free of trading frictions and barriers, any assets with identical risk and future cash flows should trade at a single price, and any deviation should be eliminated instantly by arbitrageurs. In the crypto domain, celebrated as "natively digital" and possessed of "global liquidity," the law of one price was once expected to find its purest expression. The evolution of the perpetual futures market has indeed largely borne out this trend. Through complex cross-exchange arbitrage strategies (as described in Section 16.5), market participants "level" prices across different trading venues, bringing the entire market's pricing into alignment.

When we examine this seemingly smooth pricing system at higher resolution, however, we find that it is not perfectly consistent. As arbitrage forces approach their theoretical limit, they encounter a series of insurmountable obstacles—precisely the constraints

analyzed in Chapter 17. Because these constraints exist, the law of one price is not fully realized in reality but leaves behind a series of persistent, observable "residual deviations." These residual spreads are not random noise; they constitute information-rich anomalies that delimit the boundaries reachable by cross-exchange arbitrage forces. This section analyzes three core cross-exchange anomalies, revealing how they map the deep features of market microstructure, participant behavior, and the macro regulatory environment.

18.4.1 The residual spread among leading exchanges

As the perpetual futures market has matured, leading exchanges (such as Binance, OKX, and Bybit), by virtue of their enormous liquidity, vast user bases, and relatively complete infrastructure, constitute the "core trading platforms" of global crypto derivatives trading. Among these hubs, the perpetual futures spread on major assets such as Bitcoin has converged to an extremely low level. Under the normal conditions of a smoothly functioning market, this spread typically holds within a narrow band of 1–3 bps. Yet although this spread is extremely small, it never truly vanishes to zero.

This ever-present "last basis point" of spread is itself an informative economic signal. It is not market inefficiency; on the contrary, it precisely measures the cost of achieving market efficiency. The essence of this "residual spread" is the minimum composite execution cost that cross-exchange arbitrage must bear. This composite cost can be decomposed into at least four layers. The first is the mechanical execution cost, including bilateral fees, order slippage, and latency risk. The second is the adverse-selection premium, the protective spread that cross-exchange market makers embed in their quotes to compensate for adverse order flow from informed parties. The third is the capital-readiness cost: maintaining cross-exchange arbitrage capacity requires pre-positioning idle margin on multiple platforms, and the annualized opportunity cost of this float capital is about 50–150 bps. The fourth is the withdrawal-rebalancing latency risk: after closing a cross-exchange spread, margin must be rebalanced across platforms by withdrawing on-chain, and the spread may reverse during the 10–60 minute on-chain confirmation period. As Chapter 16's "Arbitrage Infrastructure Hypothesis" points out, arbitrage returns are essentially the reward for providing the market with liquidity and price-discovery services. This residual spread can therefore be understood as the minimum operating fee for keeping the "cross-exchange price-transmission mechanism" functioning normally. When the spread narrows below this threshold, the returns generated by arbitrage trades no longer cover their costs and risks, and the incentive to arbitrage disappears. This seemingly negligible residual spread therefore delimits the true lower bound of market efficiency achievable under current technological and institutional conditions.

Moreover, this residual spread is not a static constant but displays pronounced state dependence. In calm periods it is 1–3 bps; when market volatility intensifies it widens systematically to 5–15 bps; and in extreme crisis events (such as an exchange collapse or a market-wide flash crash) it can burst instantly to more than 50–200 bps. This dynamic variation shows that the carrying capacity of the arbitrage infrastructure is finite. Under market stress, liquidity providers and arbitrageurs actively withdraw or

reduce their risk exposure, intensifying capital fragmentation and causing execution risk to surge, thereby raising the cost of maintaining price consistency. The fluctuation of this residual spread can therefore serve as an effective indicator of the health and robustness of the market's entire arbitrage infrastructure.

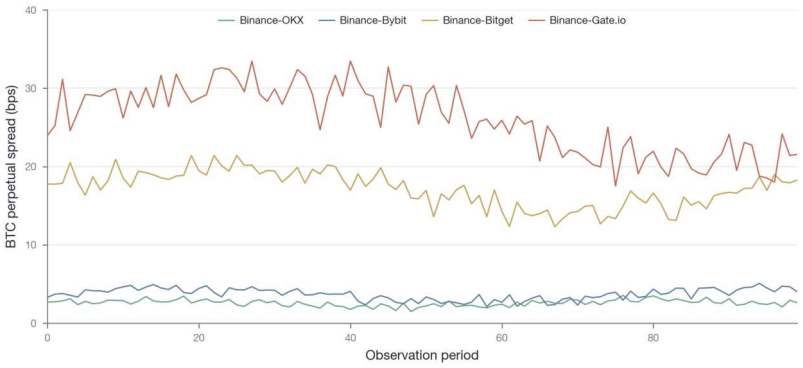


Figure 18-9. Time series of the BTC perpetual futures spread across different exchange pairs (simulated data, constructed from the historical spread characteristics of pairs such as Binance-OKX, Binance-Bybit, and Binance-second-tier-exchange)

Using simulated data, Figure 18-9 shows the dynamic features of the spread across different exchange pairs: the spread of leading-exchange pairs (such as Binance-OKX) fluctuates in a narrow band at a low level, whereas the spread against second-tier exchanges is systematically higher and more volatile, clearly displaying the tiered structure of the spread.

18.4.2 The core-periphery pricing gradient

Pulling back from the view among leading exchanges to survey the broader exchange ecosystem, a clear "core-periphery" pricing-gradient structure emerges. Cross-exchange spreads are not uniformly distributed but differentiate in tiers strictly according to an exchange's market standing, liquidity, and credit grade. The core tier comprises top exchanges such as Binance, OKX, and Bybit, among which the spread is smallest—typically 1–3 bps, as noted above. The near-periphery tier includes second-tier exchanges of a lower rank in volume and market influence, such as Bitget and Gate.io; their spread against core-tier exchanges widens significantly, usually to between 5 and 20 bps. The far-periphery tier covers many smaller, less liquid regional or emerging exchanges, whose spread against the core platforms can reach as much as 50–200 bps, and even higher in some extreme cases.

This pronounced gradient structure is a direct manifestation of the market pricing, through price signals, the intrinsic risk of different exchanges. Attributing it with Chapter 17's "constraint–anomaly mapping" framework, we find that the counterparty-risk constraint is the first-order cause of this anomaly: depositing funds on an exchange means bearing that exchange's credit risk (the FTX episode is a far-reaching case), and the credit risk of second-tier and peripheral exchanges is clearly higher than that of

leading exchanges, so rational arbitrageurs demand higher expected returns to compensate for the additional credit-risk exposure, which manifests directly as a systematic spread premium. Differences in counterparty risk then intensify capital fragmentation: cross-exchange arbitrage requires pre-depositing margin on each platform, and with limited capital, arbitrageurs prioritize deploying it on the deepest, lowest-credit-risk core-tier exchanges, while pre-positioning capital on peripheral exchanges is economically unattractive. This leaves insufficient arbitrage-capital bandwidth connecting the core to the periphery, weakens the transmission chain of price discovery, and permits larger spreads to persist.

Exchange tier	Representative exchanges	Median BTC perpetual futures spread	Primary risk characteristics
Core tier (first tier)	Binance, OKX, Bybit	1–3 bps	Deepest liquidity, lowest credit risk
Near-periphery tier (second tier)	Bitget, Gate.io	5–20 bps	Next-deepest liquidity, rising counterparty risk
Far-periphery tier (third tier and below)	Regional or emerging exchanges	50–200 bps	Thin liquidity, significant counterparty risk

Table 18-5. Exchange tiers and the BTC perpetual futures spread gradient (Data source: constructed by the author based on the Kaiko exchange ranking [7] and public industry data)

Table 18-5 shows the negative correlation between an exchange's relative credit rating and its median BTC perpetual futures spread. The higher-credit-grade exchange group (leading exchanges) has the smaller spread; as the credit grade declines, the spread widens nonlinearly, clearly displaying the "core-periphery" gradient. Two clarifications about the data are needed here. First, the median perpetual futures spreads in the table are typical order-of-magnitude estimates constructed by the author from public market data, not a measured panel from any single unified database. Second, the Kaiko exchange ranking [7] targets **spot** exchanges and scores them comprehensively along dimensions such as governance, security, and liquidity; it does not itself directly provide perpetual futures spreads or a formal credit rating, so here it serves only as a weak supporting reference for "exchange quality tiering." For a metric closer to a "credit-risk gradient," a more appropriate basis is Agio Ratings' counterparty probability-of-default (PD) score [8] (see the discussion of third-party rating systems at the end of Section 18.4.2).

In terms of persistence, the core-periphery gradient is a typical structural anomaly. As long as the crypto exchange market remains fragmented and heterogeneous, and as long as the credit risk of different platforms cannot be fully eliminated or perfectly hedged, this spread gradient will persist as a durable feature of market structure. The collapse of FTX provides an important counterexample for this framework: before its collapse, FTX, as one of the world's top three exchanges by volume, had a perpetual futures spread at the core-tier level (1–3 bps), and the market's ex ante risk pricing entirely failed to reflect its actual, fatal-level counterparty risk. This means that the spread, as a credit-risk signal, is inherently backward-looking: it can effectively price known, gradually changing risk differences but has almost no early-warning capacity

for undisclosed, discontinuous risk events (such as the misappropriation of client assets). Of course, as the industry gradually consolidates, regulation matures, and the exchange-rating systems provided by third-party institutions such as Kaiko improve [8][7], counterparty risk becomes more transparent, and we may see this gradient ease in the future. From a systemic-risk perspective, the core-periphery gradient is not merely a passive reflection of risk differences but also an active contagion-amplification mechanism: peripheral exchanges, unable to attract arbitrage capital under normal conditions, continuously accumulate mispricing, and these accumulated deviations are released in synchrony during periods of market stress; when informed traders urgently exit all peripheral positions, their selling pressure further intensifies volatility in the core market, bearing a structural resemblance to the run pattern in the tri-party repo market of 2008. Even so, the possibility of its complete disappearance is minuscule.

18.4.3 Regulatory barriers and segmented pricing zones

If the two anomaly types above arise from the market's intrinsic frictions and risk pricing, the third arises from a powerful intervention external to the market: the law and regulation of sovereign states. When regulatory force directly severs or severely obstructs the free flow of capital across jurisdictions, the law of one price is violated in the most extreme way, forming discrete "segmented pricing zones." This is the case with the shortest causal chain in this chapter's "constraint–anomaly mapping": a regulatory constraint directly blocks the arbitrage path.

The most famous case is the "kimchi premium" of the Korean market. Because of the Korean government's strict capital controls and foreign-exchange regulations, the channels through which Korean citizens can convert won into foreign currency and transfer it to overseas crypto exchanges are heavily restricted. Conversely, selling overseas crypto assets on the Korean market, converting the proceeds into won, and remitting them abroad is equally difficult. This two-way barrier to capital flow largely isolates the Korean crypto market from the global market. When domestic Korean demand for crypto assets (especially speculative demand) exceeds the global average, cheaper external assets cannot flow in freely to meet it, so crypto prices on domestic Korean exchanges are systematically higher than in the global market, forming a significant "kimchi premium." Historically, this premium has repeatedly exceeded 20% and, in some extreme periods, reached more than 50% [1]. The same effect exists in the perpetual futures market, causing the perpetual futures prices of Korean exchanges to deviate from the global pricing center over the long run.

A broader and more far-reaching case arises from U.S. regulatory policy toward the crypto derivatives market. The U.S. regulatory framework imposes strict market-access restrictions on crypto derivatives. Under the Commodity Exchange Act, the U.S. Commodity Futures Trading Commission (CFTC) [9] prohibits unregistered platforms from soliciting and accepting orders from U.S. customers. The enforcement boundary of this rule is not entirely clear (for example, the legal consequences of a U.S. individual accessing an offshore exchange by technical means have not been clearly adjudicated in current case law), but its effect is to make the overwhelming majority of regulated U.S. institutional investors unable, in practice, to participate directly in the offshore perpetual

futures market. As a result, the vast majority of U.S. investors, especially institutional investors, can trade only on U.S.-regulated platforms such as the Chicago Mercantile Exchange (CME) and Coinbase, and cannot participate directly in the world's deepest offshore markets (such as Binance and OKX). This regulatory isolation creates two parallel pricing systems: an onshore dollar pricing system benchmarked to CME Bitcoin futures and Coinbase spot prices, and an offshore-dollar (usually USDT or USDC) pricing system dominated by the perpetual futures of offshore exchanges such as Binance and OKX. Although both are denominated in dollars and share the same underlying asset, systematic pricing deviations arise between them because of differences in participant groups, leverage rules, trading hours, regulatory expectations, and funding costs. For example, CME futures prices reflect more the trading behavior of regulated institutions, whereas offshore perpetual futures reflect more the sentiment and demand of global retail investors and professional traders not subject to U.S. regulation. This regulation-created fissure is structural, and its evolution depends directly on the future direction of global—and especially U.S.—regulatory policy.

Regulatory environment	Representative market	Average spread level	Driving factor
Global leading exchanges (low regulatory friction)	Among Binance, OKX, Bybit	1–3 bps	Free capital flow, unobstructed arbitrage
Jurisdictions with a clear compliance framework	The EU's MiCA framework, U.S. CME	10–50 bps	Differences in compliance cost and market access
Capital-controlled jurisdictions	South Korea (kimchi premium)	200–2,000 bps	Capital-flow barriers block the arbitrage path

Table 18-6. Crypto derivatives spread levels under different regulatory environments (Data source: constructed by the author based on Seo et al. [1], Kaiko [7], and public industry data)

Table 18-6 shows the average spread level under different regulatory environments. Among global leading exchanges, the spread remains extremely low. In jurisdictions with specific regulatory restrictions (such as capital controls), however, the average spread widens sharply, forming a significant "segmented pricing zone." In jurisdictions with an increasingly clear regulatory framework (such as the EU's MiCA and the U.S. stablecoin regulatory proposal, the GENIUS Act, still under legislative review), the spread also takes on different levels because of differences in compliance cost and market access. A note on the data is needed: the 200–2,000 bps range for capital-controlled jurisdictions in the table connects with the reading in the introduction that the spot kimchi premium has "a long-run center of about 120 bps (the estimate from the threshold model of Seo et al. [1]) and can exceed 20% or even 50% in extreme periods"—120 bps is the long-run steady-state center, whereas 200–2,000 bps reflects the wider upper bound of the distribution including extreme periods; Kaiko [7] serves here only as an auxiliary reference for exchange quality tiering (it is itself a comprehensive spot-exchange ranking), and the perpetual spread figures for each tier are typical order-of-magnitude values constructed by the author from public market data.

The persistence of anomalies driven by regulatory barriers depends entirely on macro legal and political processes. As long as crypto regulation worldwide fails to achieve unification and coordination, these "segmented pricing zones" produced by jurisdictional division will persist, serving as vivid proof that the law of one price still faces powerful resistance in the age of globalization.

18.5 Cross-asset anomalies

Having examined the basis, funding-rate, and cross-exchange anomalies that revolve around a single asset in the perpetual futures market, our analysis now enters the dimension that spans different assets. Unlike the spreads above—"the same asset in different places or different forms"—cross-asset anomalies concern the relative-pricing deviation between "related but different assets." Such anomalies cannot be corrected by a simple two-leg spot-derivative arbitrage but require more complex multi-leg strategies such as statistical arbitrage. As Chapter 17 revealed, every increment in a strategy's complexity multiplies the constraint dimensions and execution frictions it faces. The persistence of cross-asset anomalies therefore offers us an effective analytical vantage point from which to observe what form of harder-to-eliminate price deviation the market takes on when model risk and correlation-breakdown risk are compounded. From the broad "inefficiency zone" of the altcoin market to the "long-cycle drift" in the relative pricing of Bitcoin and Ethereum, these anomalies delimit the boundaries that advanced arbitrage strategies face in the real world.

18.5.1 The "inefficiency zone" of altcoin perpetual futures

The crypto asset market is not a homogeneous whole but an ecosystem highly stratified in liquidity and efficiency. If the perpetual futures markets of Bitcoin and Ethereum are the high-efficiency market of the whole ecosystem, then the perpetual futures markets of the overwhelming majority of small- and mid-cap altcoins constitute a broad low-efficiency market. Here the various anomalies discussed in earlier sections are systematically amplified, far exceeding major assets in both magnitude and duration. This pronounced difference in efficiency does not stem from altcoin-market participants being "more irrational" but is a direct mapping of the tighter arbitrage constraints in these low-efficiency markets.

The most intuitive evidence lies in the efficiency of price discovery. For Bitcoin or Ethereum, the basis between perpetual futures and spot is usually anchored tightly within tens of basis points. For a small coin ranked beyond the top 100 by market cap, however, it is the norm for the perpetual futures basis to reach hundreds or even thousands of basis points during periods of market volatility. The same holds for the funding rate: extreme-rate events are relatively rare for major coins, whereas the perpetual futures of many altcoins are perennially shrouded in extreme positive or negative rates, with annualized rate costs repeatedly exceeding 100%. Cross-exchange spreads are widened even more markedly: a spread nearly eliminated among mainstream exchanges can persist above 1% over the long run across the different trading venues of an altcoin.

Figure 18-10 presents this structural stratification of efficiency visually: it shows the negative correlation between an asset's market cap and its derivatives market's pricing efficiency (proxied by average basis magnitude). As market cap declines, the average perpetual futures basis widens superlinearly (convexly), forming a steep curve that transitions from the high-efficiency market to the low-efficiency market; for the small-cap tokens at the far left of the curve, the very large basis means the arbitrage mechanism is only half-functional here.

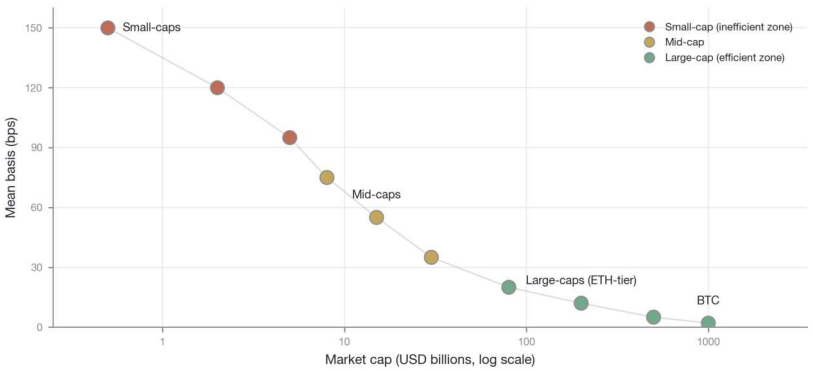


Figure 18-10. The relationship between asset market cap and the magnitude of the perpetual futures basis (illustrative/simulated: the inverse relationship is the empirical direction, while the specific basis curve is an illustrative order of magnitude rather than measured point values; stratified by market-cap rank)

The causes of this "inefficiency zone" can be traced precisely to several of the constraint dimensions discussed in Chapter 17, among which the liquidity constraint (the fifth dimension) is the first-order explanation. First, altcoin order books are extremely thin. For an arbitrageur, executing a trade meant to correct a price deviation may incur a market-impact cost that exceeds the spread itself. More critically, on an extremely thin altcoin order book the price-impact function is convex rather than linear: even a medium-sized order may cross multiple price levels, so the impact cost grows superlinearly (convexly) with trade size. This means there is no "partial arbitrage" equilibrium: arbitrageurs cannot make the strategy feasible by shrinking position size, because even the smallest economically viable position may exhaust one side of the entire order book. This threshold effect provides a stronger theoretical foundation for the structural persistence of altcoin anomalies than mere "excessive cost," rendering arbitrage economically infeasible. Second, effectively hedging the risk of altcoins is extremely difficult. The spot-shorting mechanism necessary for negative-basis arbitrage is, in the altcoin world, either entirely absent or prohibitively expensive to borrow. Many altcoins' perpetual futures are even listed on only a few exchanges, so the basis for cross-exchange arbitrage ceases to exist. Finally, the opacity of the information environment amplifies arbitrageurs' adverse-selection risk. Facing an altcoin with murky fundamentals and severe information asymmetry, an arbitrageur struggles to judge whether a large spread is a genuine arbitrage opportunity or a counterparty position on which they are about to become the loser to an informed party. In such an environment, rational arbitrageurs

choose to stay away, allowing the price deviation to persist.

The "inefficiency zone" of altcoin perpetual futures is therefore not a temporary, easily repaired market defect but a structural anomaly determined by the long-tail distribution of liquidity, an endogenous feature of the crypto market. As long as the crypto world continues to contain tens of thousands of highly liquidity-dispersed "long-tail assets," these low-efficiency markets will persist as domains that arbitrage forces cannot effectively cover, serving as a lasting emblem of the stratification of market efficiency.

18.5.2 The long-cycle drift in BTC-ETH relative pricing

Moving up from the inefficiency zone, we arrive at the market's most liquid core: Bitcoin and Ethereum. Here, simple arbitrage opportunities are nearly eliminated. A subtler, larger-scale anomaly appears, however: the relative-pricing relationship between the two major assets exhibits a cyclical drift lasting months or even spanning years. The ETH/BTC price ratio, the core metric of the two assets' relative value, does not revert stably around a mean as statistical-arbitrage models expect but displays persistent, trending deviations across different market cycles.

This long-cycle drift is often tightly linked to the market's dominant logic and the evolution of fundamentals. For example, in periods when Ethereum is about to undergo a major upgrade (such as the Merge upgrade) or when its ecosystem (such as Layer 2) makes a breakthrough, the ETH/BTC ratio typically rises for several months. Conversely, when Bitcoin becomes the market's focus (such as the approval of a spot ETF), the ratio may experience prolonged downward pressure. Figure 18-11 shows the time-series variation of the ETH/BTC rolling correlation coefficient, in which it is clear that, although the long-run mean stays at a relatively high level, the correlation undergoes significant, sustained drift during specific event windows, reflecting phased changes in the two assets' relative-pricing logic.

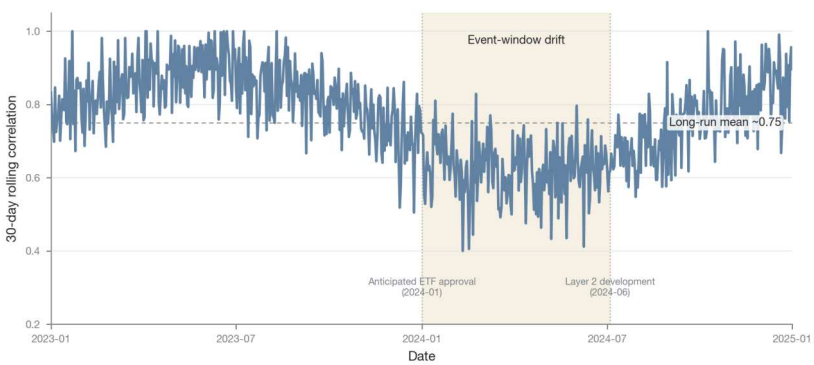


Figure 18-11. The long-cycle drift of the BTC-ETH rolling correlation coefficient (constructed by the author from public market data)

Figure 18-11 shows the trajectory of the BTC-ETH rolling correlation coefficient over a fairly long time span: the long-run mean stays at a relatively high level (usually between 0.7 and 0.9), but during specific event windows (major Ethereum upgrades and

the period around the approval of a Bitcoin ETF) it undergoes a significant downward drift lasting weeks or even months, and this instability of correlation directly constitutes the core risk facing statistical-arbitrage strategies. The rolling correlation coefficient in the figure is only an illustrative measure, however; to characterize this time-varying correlation rigorously, a more robust approach is the DCC-GARCH dynamic conditional correlation model proposed by Engle (2002) [10], which avoids the estimation bias of a rolling window during periods of drastic volatility change.

From the perspective of the constraint–anomaly mapping, the existence of this long-cycle drift directly challenges the core assumption of statistical arbitrage (Section 16.8 of Chapter 16). The strategy rests on the belief that the spread or ratio between two highly correlated assets will "revert to the mean." When the "mean" itself becomes a non-stationary parameter, however, the premise of the arbitrage is shaken. For assets like BTC and ETH, whose fundamentals are still evolving rapidly, their long-run fair ratio is itself a highly uncertain question. Is today's deviation a "mispricing" to be corrected tomorrow, or a "reasonable revaluation" reflecting future changes in fundamentals? The arbitrageur faces substantial model risk.

More important, the correlation-breakdown risk discussed in Chapter 17 is fully evident here. If an arbitrageur builds a position based on historical data (for example, going long on the reversion of the ETH/BTC ratio), then once an unforeseen change in market structure breaks the correlation between the two entirely, the arbitrageur faces unlimited losses. This deep, hard-to-hedge risk greatly suppresses the scale of capital arbitrageurs are willing to commit. They may trade in small size over short horizons, but few dare to bet in large size and over the long run on the reversion of a large, fundamentals-driven relative-price ratio. The insufficiency of arbitrage forces leaves ample room for this long-cycle drift.

Here we touch on a blurry boundary in the definition of an anomaly. Is a months-long rise in the ETH/BTC ratio an "anomaly" caused by arbitrage failure, or a "reasonable repricing" by the market of an improvement in ETH's fundamentals? The answer is perhaps both. Part of the deviation may indeed reflect a change in fundamentals, while another part may be market overreaction or underreaction caused by the constraints described above. A workable criterion is this: does the magnitude of the price deviation far exceed what fundamental information can reasonably explain? After the driving event (such as the completion of an upgrade) ends, does the abnormal drift in the price ratio still persist? If the answer is yes, then we have stronger grounds to characterize it as a structural cross-asset anomaly sustained by arbitrage constraints. In terms of the "pure-anomaly–equilibrium-deviation" continuum proposed in Section 18.1.1, the BTC-ETH long-cycle drift is precisely an **intermediate state** lying between the two: the part driven by changes in fundamentals approaches equilibrium repricing, whereas the part amplified by model risk and correlation-breakdown risk approaches a pure anomaly; the two criteria above—whether the deviation exceeds what fundamentals can explain, and whether it persists after the event ends—are precisely the testable standards that operationalize this dichotomy's borderline cases.

18.6 Term-structure anomalies

When we turn from the cross-asset dimension to the dimension that unfolds along the time axis, another set of deep pricing distortions comes into view. Term-structure anomalies concern the pricing consistency of the same underlying asset across different maturities or product forms. Their core feature is that they involve rate deviations between instruments in different "time slices"—perpetual futures and dated futures, onshore ETFs and offshore perpetuals. The triangular arbitrage or cross-system arbitrage needed to eliminate such deviations faces the combined action of multiple constraints—participant segmentation, regulatory isolation, and session mismatch—which makes these anomalies highly persistent.

18.6.1 The implied-rate deviation between perpetual funding and dated-futures basis

In traditional financial markets, the derivative prices of the same asset across different maturities together form a smooth, internally consistent term-structure curve that reflects the market's expectations of future prices and the financing cost over different time intervals. In the crypto derivatives market, however, we observe two core instruments—perpetual futures and dated futures—giving frequently contradictory signals about "financing cost." This produces a persistent deviation of "two implied-rate curves."

The first curve derives from the funding rate of perpetual futures. The funding rate is essentially the interest that longs and shorts pay each other to anchor the perpetual futures price to spot. Annualizing this rate, settled every 8 hours, yields an ultra-short-term "implied financing rate" driven directly by market sentiment and leverage demand. The second curve derives from traditional dated futures, such as the BTC quarterly futures traded on the Chicago Mercantile Exchange. The basis of such futures (the difference between the futures price and the spot price) likewise implies the financing cost from now to the delivery date. In theory, through triangular arbitrage (Section 16.6 of Chapter 16), the rate levels the two curves reflect at the same maturity should converge. This implied-rate perspective aligns with the core concern of the crypto carry-trade literature: the perpetual futures funding rate and the dated-futures basis essentially give two sets of implied-rate signals about the cost of carry, and their persistent deviation is precisely a sign that the carry structure has not yet been fully arbitraged away in the crypto market [5].

The reality revealed by empirical data, however, is otherwise. Most of the time, the short-term rate annualized from the perpetual futures funding rate deviates significantly and persistently from the forward rate computed from the dated-futures basis. Figure 18-12 simulates this typical shape of deviation: the curve representing perpetual futures is systematically higher across maturities and has a steeper term structure, whereas the curve representing dated futures is relatively flat and reflects more the institutional-level, medium- to long-term funding cost and expectations. The rate spread between the two curves constitutes a term-structure anomaly unique to the crypto market.

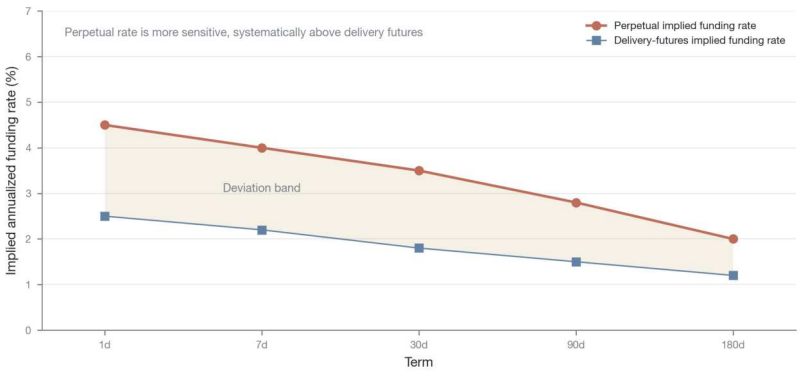


Figure 18-12. A comparison of the implied-rate curves of perpetual futures and dated futures (Data source: simulated data constructed by the author)

Figure 18-12 presents side by side the implied-rate curve derived by annualizing the perpetual futures funding rate and the implied-rate curve computed from the dated-futures basis: the perpetual futures implied rate is systematically higher across all maturities with a steeper curve, exceeding dated futures especially clearly at the short end; the dated-futures curve is flatter, reflecting the more stable funding-cost expectations of institutional participants. The rate spread between the two curves is persistently positive across maturities, a systematic rather than random pattern of deviation, confirming a structural split in the two markets' pricing logic.

The most central explanation for the root of this deviation is "participant segmentation." The dominant forces in the perpetual futures market are crypto-native retail investors, traders, and quantitative funds, whose trading is driven more by short-term narratives, market sentiment, and high-leverage speculative demand. The main participants in the CME futures market, by contrast, are traditional financial institutions, hedge funds, and professional investors seeking compliant risk exposure, who have different information sources, risk appetites, capital costs, and regulatory constraints. This fundamental difference in participant structure produces a systematic split in the two markets' pricing logic. The perpetual futures implied rate is more like the "unsecured overnight lending rate" of the crypto-native world, whereas the CME futures implied rate is closer to the "collateralized funding rate" of the traditional financial system.

In theory, triangular arbitrageurs could eliminate this rate deviation by "going long spot in one market, short futures in another, and taking opposite positions in the two derivatives markets simultaneously." As analyzed in Section 17.5.2 of Chapter 17, however, this multi-leg, cross-platform arbitrage strategy faces severe execution constraints. It requires not only holding sufficient margin and trading limits in multiple markets simultaneously but also coping with the settlement risk, margin-call risk, and potential regulatory obstacles of different platforms. More fundamentally, this three-leg arbitrage strategy carries an ineliminable basis risk: the funding rate of the perpetual futures leg changes every 8 hours, whereas the basis-implied rate of the CME quarterly futures is locked until expiry. This means that even if the three-leg position perfectly hedges the rate deviation at inception, a subsequent sign reversal in the funding rate will directly

erode the strategy's expected return, while the CME leg cannot be adjusted flexibly to cope with this daily-frequency movement. A considerable part of the perpetual-vs.-dated rate deviation is therefore reasonable compensation for this roll/basis risk, not pure arbitrage failure. Because of these constraints, the scale of capital able to execute such arbitrage effectively is relatively limited and insufficient to close the gap between the two rate curves entirely.

The deviation of the "two implied-rate curves" is therefore a deep structural anomaly. It is not merely a technical pricing failure but a manifestation of two different market systems—the crypto market and the traditional financial market—interacting, in institutions, culture, and participant composition, without having fully merged. As Bitcoin spot ETFs are approved and more traditional institutions enter, this participant segmentation may slowly close and the two curves may gradually converge. In the foreseeable future, however, this term-structure anomaly, sustained by structural segmentation, will remain an important feature of the crypto market.

18.6.2 CME/ETF and offshore-perpetual dual-track pricing

The approval of Bitcoin spot ETFs in the United States in early 2024 was an important institutional event in the integration of crypto assets into the mainstream financial system. Yet this institutional integration, far from eliminating pricing anomalies entirely, gave rise to a wholly new, more institutionalized structural fissure: a significant "dual-track pricing" pattern between the onshore, strictly regulated ETF/CME pricing system and the offshore, freely flexible perpetual futures pricing system.

These two systems differ fundamentally along almost every dimension, and these differences form successively compounding arbitrage barriers. The segmentation of the regulatory environment is the most fundamental barrier: ETFs operate under the close supervision of the U.S. Securities and Exchange Commission (SEC), whereas mainstream perpetual futures are located in offshore, relatively lightly regulated jurisdictions, and this segmentation directly dictates that the participant groups of the two systems cannot freely intersect. Regulatory segmentation in turn gives rise to a mismatch in trading hours: ETFs follow the New York Stock Exchange (NYSE) session (9:30 a.m. to 4:00 p.m. ET), whereas perpetual futures trade 24/7 without interruption, which means arbitrageurs cannot complete a hedge through the ETF side while the U.S. equity market is closed. The time mismatch is further compounded by a difference in settlement mechanisms: ETFs settle in U.S. dollar cash and follow a T+1 redemption mechanism, whereas perpetual futures settle instantly in stablecoins, and this lag means that even during hours when both markets are open, arbitrageurs still face at least one trading day of capital occupancy and risk exposure. These three layers of institutional gap compound upon one another to form an institutional barrier obstructing the free flow of arbitrage capital between the two systems.

The most typical manifestation of this dual-track pricing pattern is the "intraday cyclical fluctuation" of the spread. As shown in Figure 18-13, the spread between the ETF and offshore perpetual futures exhibits cyclical fluctuation highly correlated with global trading sessions. During the U.S. session, because both markets are open, arbitrage activity is relatively active and the spread is compressed to a low level. Once the U.S.

market closes and the Asian and European sessions begin, however, the offshore perpetual futures market enters a "period lacking a price reference." At this point, most global Bitcoin trading and price discovery is concentrated in the perpetual futures, and lacking the strong price "anchor" of the ETF/CME, the perpetual futures price is more susceptible to drift under the influence of regional news, capital flows, or market sentiment, causing the spread against the previous U.S. equity close to widen significantly. Only when the next U.S. session opens is the overnight-accumulated spread corrected again by returning arbitrageurs.

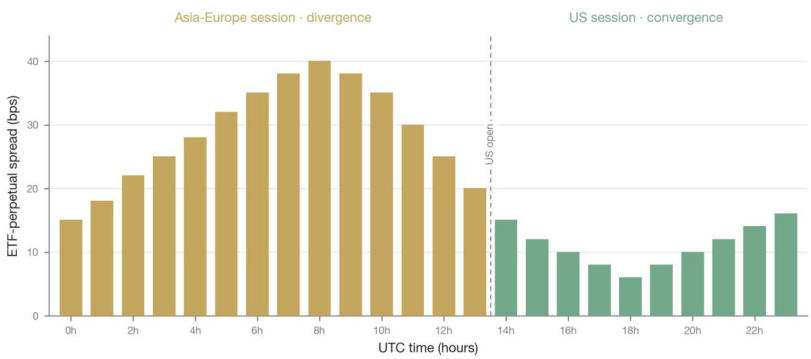


Figure 18-13. The intraday cyclicality of the spread between the ETF and offshore perpetual futures (historical period 2024–2025; illustrative/simulated, with the shape based on historical characteristics; note: CME crypto futures switched to 24/7 round-the-clock trading in late May 2026, and the weekend and overnight gap structure is being eroded, but the session mismatch caused by the spot ETF trading only during U.S. equity hours still exists)

With intraday time on the horizontal axis, Figure 18-13 shows the cyclical fluctuation of the spread between the ETF's net asset value and the offshore perpetual futures price: during the U.S. session (roughly 14:30 to 21:00 UTC), both markets are open and arbitrage is active, so the spread narrows to a low level; after the Asian and European sessions begin, because the ETF market is closed while perpetual futures continue trading, the spread widens to its intraday peak (about 40 bps) in the middle of the Asia-Europe session and then gradually converges as the U.S. equity open approaches. This clear "converge by day, diverge by night" pattern is a direct mapping, along the spread dimension, of the institutional factor of the two markets' session mismatch.

Examined through the constraint–anomaly framework, this "dual-track pricing" is a typical case of multiple structural constraints acting together. Regulatory segmentation (the seventh dimension) is its most fundamental cause, directly dictating that the two systems cannot simply be treated as "the same market." The session mismatch and the delay in the settlement mechanism constitute the enormous execution risk and time cost that arbitrageurs face. Even if an arbitrageur sees a significant discount of the perpetual futures relative to the ETF price during the Asian session, they cannot immediately buy the perpetual futures and lock in a profit by selling the ETF, because the ETF market is not yet open. They must bear hours of overnight risk and wait until the U.S. market opens to complete the hedge. Here we must distinguish the differing

constraints faced by different types of arbitrage participant and make an important correction to the related session arithmetic. For ordinary investors and non-authorized participants, the constraint is as described above: they must wait for the ETF secondary market (on the NYSE, with a session of 9:30 a.m. to 4:00 p.m. ET, about 6.5 hours) to open before completing the hedge. For ETF authorized participants (APs), however, arbitrage can be executed through the primary market's creation/redemption mechanism and hedged in the CME futures market. CME Bitcoin futures historically traded about 23 hours per day (Sunday 17:00 to Friday 16:00 CT, with an approximately 60-minute maintenance window each trading day), not the "about 18 hours" often cited previously; therefore, for APs, the true structural session gap against the 24-hour continuous trading of perpetual futures was about 1 hour, not about 6 hours. More important is the timeliness update: as of May 2026, CME has shifted to near-24/7 continuous trading (with only about 2 hours of maintenance on Saturday), the session mismatch between CME and perpetual futures has been essentially eliminated, and this channel no longer constitutes a substantive constraint for APs. On balance, the main body of this section's "session mismatch" argument should rest on **the mismatch between the ETF secondary market for ordinary investors and non-authorized participants (the NYSE's 6.5-hour cash session) and the 24-hour continuous trading of perpetual futures**—this "overnight void" outside the roughly 6.5-hour window is the true source of the current structural gap. It is precisely the existence of this risk exposure that makes arbitrage trades for secondary-market participants attractive only when the spread is large enough to compensate for the overnight risk.

The new fissure of the ETF era is therefore a structural anomaly still taking shape. It is not a temporary market failure but the inevitable product of institutional friction among different jurisdictions and market architectures as the global financial system takes in a wholly new, borderless digital asset. Unless a major institutional change occurs in the future—such as a 24-hour-trading ETF or the establishment of a globally unified clearing system for crypto assets—this onshore-offshore "dual-track pricing" pattern will persist over the long run. It will become a new arena of contest for macro traders and cross-market arbitrageurs and offers us a continually evolving, information-rich case for observing how two financial paradigms interact, compete, and merge.

18.7 CEX–DEX anomalies

Among the six major anomalies of the perpetual futures market, the efficiency gap between centralized exchanges (CEXs) and decentralized exchanges is perhaps the most structural. It reveals not only the vast microstructural difference between two different market architectures but also a fundamental question: does decentralization necessarily entail a loss of efficiency? This section analyzes the systematic deviations between CEXs and DEXs in price discovery, funding rates, and arbitrage opportunities. We argue that these anomalies stem partly from the limitations of the current DEX technology stack—such as oracle latency and high gas costs—which are "early-stage technical limitations" that can be gradually mitigated through technological iteration. Another part of the anomalies, however, is rooted in the intrinsic properties of decentralized architecture, such as the unavoidable physical latency of on-chain transactions. This

constitutes what we call the "structural arbitrage residual," an efficiency lower bound that seems impossible to eliminate entirely under the current technological paradigm. The distinctiveness of the CEX–DEX anomaly therefore lies in the fact that it displays, at once, the market's room for improvement and the fundamental efficiency cost that decentralization must pay.

18.7.1 Oracle latency and persistent spreads

The most intuitive anomaly between CEXs and DEXs is the price difference that persists for the same asset across the two markets. Academic research has quantified the significance of this gap. For example, Barbon and Ranaldo (2021) find that high and unstable on-chain gas fees impose significant costs on DEXs, especially for small trades, directly producing persistent arbitrage deviations [11]. This spread is not random noise but has a clear "information lag" feature: when the CEX price, the primary source of information, fluctuates violently, the CEX–DEX spread widens significantly, whereas in calm markets the spread converges. Empirical research on crypto markets shows that CEXs dominate in price discovery and that information flows in one direction from CEXs to DEXs [2], clearly indicating that the DEX price-discovery process is to some extent "anchored" to CEXs but with a pronounced lag.

The core source of this latency is the sixth-dimension constraint discussed in Chapter 17—smart-contract risk—which here takes the concrete form of the physical latency of oracles. Most DEX perpetual futures protocols, such as GMX, rely on oracles to fetch prices from CEXs or other off-chain data sources. This "fetch-broadcast-confirm" process inherently involves a time lag. As related research points out, even a 1-second delay can have large economic consequences in a violently fluctuating market [12]. For example, during an extreme market crash, the price of a major asset may move by more than 0.1% per second. For a $1 million position, a 1-second delay means more than $1,000 of potential risk exposure. This latency creates nearly risk-free profit opportunities for arbitrageurs, who monitor the spread between CEXs and DEXs and trade on stale price information on the DEX, systematically extracting returns from the protocol or ordinary users.

Figure 18-14 lays out several key dimensions of oracle latency: the top-left panel compares the typical oracle latency of different blockchains (ranging from Solana's sub-second level to the Ethereum mainnet's dozen-plus seconds); the top-right panel shows the magnitude of price deviation caused by oracle latency under different market environments, highlighting the severity of latency in high-volatility periods; the bottom-left panel characterizes the relationship between slippage and trade size, showing that slippage cost grows convexly with trade size (about a hundred-odd dollars for a $1 million trade, rising to about $2,000 at the $10 million level); and the bottom-right panel compares the vast difference in execution latency between DEXs of different architectures and CEXs, revealing the source of the efficiency gap.

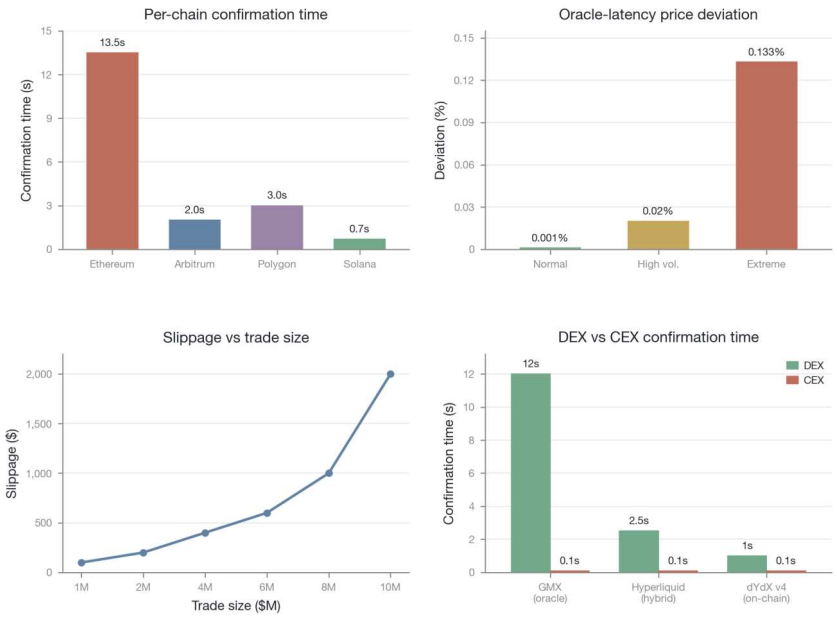


Figure 18-14. The multidimensional impact of oracle latency on DEX price discovery (Data source: constructed by the author; the per-chain latency and other values in the figure are representative orders of magnitude and, because of differing conventions—push/heartbeat versus pull/on-demand—are not measurements against a unified benchmark)

In analyzing oracle latency, it is necessary to distinguish two oracle-failure modes with starkly different economic consequences. The first is a passive stale price, in which the oracle has not yet updated to the latest price, creating a predictable, symmetric arbitrage window—the typical problem faced by protocols using Chainlink quotes, such as GMX. The second is active oracle manipulation, in which an attacker deliberately moves the index price by trading in a low-liquidity reference market and then extracts value on the DEX using the distorted oracle quote. The Mango Markets incident of October 2022 (a loss of about $116 million per the SEC's charging figure; different sources place it in the $110 million–$117 million range) is a typical case of the latter. Passive latency produces symmetric arbitrage opportunities, whereas active manipulation causes one-directional protocol losses, and the two contribute to the "structural arbitrage residual" through different mechanisms.

Oracle latency is not the only structural constraint DEXs face. Maximal extractable value (MEV) is another important factor that amplifies the CEX–DEX spread [13], and subsequent research has systematically quantified the scale of on-chain extractable value [14]. In the DEX environment, there is a "pending" window between a transaction's submission and its final confirmation, during which validators or searchers can extract value by reordering, inserting, or censoring transactions. Sandwich attacks (inserting buy and sell orders before and after a user's trade to capture the spread) and front-running (racing to execute a same-direction trade after observing a large order) are the most common forms of MEV. These behaviors have a dual effect: on one hand,

they directly increase the implicit trading cost for DEX users; on the other hand, they compound with oracle latency: when the oracle price lags the CEX price, MEV searchers exploit this information gap to conduct toxic arbitrage on the DEX, and this toxic arbitrage in turn indirectly weakens the DEX's market-making depth by depleting liquidity providers' funds. This vicious cycle makes MEV an important component of the structural arbitrage residual, and especially during on-chain congestion and high volatility, its impact may be comparable to or even greater than that of oracle latency.

The persistent spread between CEXs and DEXs can therefore be regarded as a direct, observable economic consequence of technical constraints such as oracle latency and MEV. It fully embodies the "constraint–anomaly mapping" framework: how a definite technical bottleneck (a constraint) translates directly into a persistent market anomaly. In terms of persistence, this is an anomaly with both cyclical and structural properties, and it is precisely a clear demonstration of the point stressed in Section 18.1.3 that "the spectrum is non-exclusive and the same broad category of anomaly can possess multiple properties." On one hand, as Layer 2 solutions and dedicated appchains rise, oracle latency is being compressed significantly, meaning that the "average level" of the spread will narrow with technological progress (the improvable component of the structural property); on the other hand, as long as DEXs rely on external information inputs or are subject to the physical time of on-chain confirmation, this latency cannot be eliminated entirely, forming an ineliminable structural efficiency lower bound.

18.7.2 The systematic deviation of DEX funding rates

Beyond price itself, CEXs and DEXs also show systematic deviation in the core mechanism of perpetual futures—the funding rate. As noted above, funding-rate arbitrage is supposed to "level" the rate across different markets. Between CEXs and DEXs, however, the efficiency of this arbitrage mechanism is markedly reduced. Empirical data show that the funding rate of DEX perpetual futures is not only far more volatile than that of CEXs but also often differs significantly from CEXs in its mean. Some studies even find that a simple funding-rate arbitrage strategy on a DEX (for example, shorting on a DEX with a persistently positive rate while going long on a CEX to hedge the risk) can earn an annualized return as high as 30%–40% (this figure comes mainly from the high-volatility period of DEX funding rates in 2022–2023 and had compressed significantly by 2024–2025). It is inaccurate, however, to treat this return as a signal of a "risk-free arbitrage opportunity." Applying the analytical framework established in Section 18.1.1, a considerable part of the CEX–DEX rate spread is in fact reasonable compensation for a basket of DEX-specific risks: the unhedgeable risk of smart contracts, the possibility of oracle manipulation, the security hazards of cross-chain bridges, and the regulatory-classification uncertainty that DEX protocols face. The 30%–40% return should therefore be understood, more accurately, as the composite risk premium for these risks, not as a signal of pure arbitrage failure. Only when the return significantly exceeds the reasonable compensation level for these risks does the excess portion constitute a genuine anomaly driven by arbitrage constraints.

The roots of this rate deviation are manifold. First, there are differences in the rate-calculation mechanism. Different DEX protocols adopt different funding-rate formulas

and parameters (such as the interest-rate composition and the way the premium index is computed) and different settlement cycles (for example, Hyperliquid's hourly settlement versus most CEXs' 8-hour settlement), which provides an "institutional basis" for rate differences. Second, there are differences in participant composition. CEXs have a vast retail base and mature institutional market makers, whereas the early users of DEXs are more often high-risk-seeking "DeFi yield-strategy participants" and professional arbitrageurs, and this difference in participant structure produces a systematic difference in the balance of long and short forces. Finally, and most critically, there is the substantial friction of cross-chain execution. Effective funding-rate arbitrage between a CEX and a DEX requires deploying funds, monitoring positions, and executing trades in two entirely different systems. This entails not only higher operational complexity, smart-contract risk, and potential cross-chain-bridge risk but also far lower margin utilization efficiency than arbitrage within a single CEX. Acting together, these constraints greatly limit the scale and efficiency of arbitrage capital, allowing the rate gap between CEXs and DEXs to persist over the long run.

Figure 18-15 shows the concrete manifestations of the CEX–DEX anomaly in spreads and funding rates: the top-left panel presents the typical level of the CEX–DEX spread under different market states (from 10 bps in a normal market to more than 100 bps in a crisis); the top-right panel simulates the paths of CEX and DEX funding rates over a period, between which a clear and persistent deviation exists; the bottom-left panel compares the "efficiency lower bound" of different DEX architectures—the minimum arbitrage residual determined by their architecture; and the bottom-right panel quantifies how arbitrage profit scales with trade size.

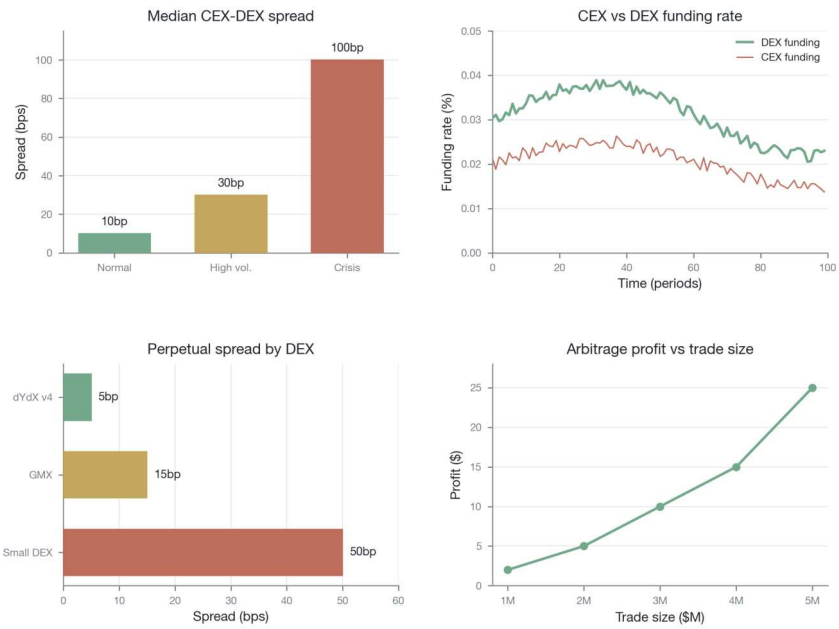


Figure 18-15. The systematic deviation of the CEX–DEX spread and funding rate (Data source: simulated data constructed by the author)

Compared with the rate splitting among CEXs (Section 18.3.4), the rate splitting between CEXs and DEXs is larger in magnitude and more fundamental in its constraint sources. The former stems mainly from institutional constraints such as margin and platform risk, whereas the latter is compounded by deeper technical constraints such as oracles, smart contracts, and cross-chain execution. The systematic deviation of DEX funding rates can therefore be regarded as a highly sticky structural anomaly produced by the combined action of multiple constraints.

18.7.3 The structural arbitrage residual

If oracle latency is a technical problem that "can be optimized," is there an unoptimizable, fundamental efficiency lower bound of decentralized architecture? The answer seems to be yes. This brings us to the core concept of this section, the "structural arbitrage residual." Its core idea is that even under the most ideal DEX design, the final confirmation of a transaction takes time. Whether relying on oracle updates or matching directly on an on-chain order book, there is a physical minimum time interval from the generation of a trading intention to its becoming irreversible on the blockchain. This interval, however brief, creates a "minimum arbitrageable window" for arbitrageurs.

Within this window, the price in the external world may already have changed while the on-chain state has not yet updated. Any arbitrage attempting to exploit this instantaneous spread must itself go through a "submit-package-confirm" on-chain process. This means that arbitrageurs are forever chasing a price change that has already occurred, and this chase itself involves a delay. There will therefore always be a spread small enough that the expected return of the arbitrage trade is insufficient to cover its execution costs (gas fees, time cost, failure risk, and so on). This spread, which cannot be eliminated entirely by arbitrage, is the "structural arbitrage residual."

Different DEX architectures have different "efficiency lower bounds" or "structural arbitrage residuals." As shown in Table 18-7, we can construct a clear hierarchy. The latency of an oracle-dependent DEX (such as GMX) is the sum of the oracle update frequency and the underlying chain's confirmation time; its latency is the longest, so its efficiency lower bound is the highest and its residual the largest. An on-chain order-book DEX (such as Hyperliquid), by placing the order book on a high-performance dedicated blockchain, frees itself from reliance on external oracles, and its latency comes mainly from the block time of its own network, giving it a middling efficiency lower bound. An appchain DEX (such as dYdX v4), by building a dedicated blockchain for a single application, can compress latency to the extreme, approaching the level of a centralized server; even so, however, its transaction confirmation must still be completed through distributed consensus, which means the latency remains nonzero. Its efficiency lower bound is the lowest of all DEXs, but it still exists.

DEX type	Representative project	Source of latency	Degree of decentralization	Efficiency lower bound
Oracle-dependent DEX	GMX	Oracle update frequency + underlying-chain confirmation time	High (relies on Ethereum/Arbitrum consensus)	Highest (largest residual)
High-performance order-book DEX	Hyperliquid	Block time of its own network	Medium-low (HyperBFT, small validator set)	Lower
Appchain DEX	dYdX v4	Distributed-consensus confirmation time	Medium (Cosmos SDK validator set, off-chain matching + on-chain settlement)	Low (but still nonzero)
Centralized exchange (reference)	Binance, OKX	Centralized-server matching	None	Approaching zero (sub-millisecond)

Table 18-7. Sources of latency and efficiency lower bounds for different DEX architecture types (Data source: constructed by the author from public market data and the literature)

Table 18-7 systematically compares the current mainstream DEX architecture types by source of latency and efficiency lower bound, using centralized exchanges as a baseline reference. From oracle-dependent DEXs (such as GMX) to on-chain order-book DEXs (such as Hyperliquid) to appchain DEXs (such as dYdX v4), the source of latency gradually narrows from the dual bottleneck of "oracle update + chain confirmation" to the single bottleneck of "distributed-consensus confirmation" alone, and the efficiency lower bound decreases accordingly. Even the appchain architecture with the lowest efficiency lower bound, however, still has latency significantly higher than the sub-millisecond matching speed of centralized exchanges, indicating that an ineliminable structural residual always exists. The "degree of decentralization" column newly added to Table 18-7 reveals an important finding: the DEX with the lowest efficiency lower bound at present (such as Hyperliquid) is precisely a DEX with a relatively low degree of decentralization, one that, through a small validator set under HyperBFT consensus (with a relatively concentrated number of validators and staking distribution), substantially approaches the speed of centralized matching. This observation in fact strengthens the argument for the efficiency-decentralization trade-off: under current technological conditions, approaching CEX execution efficiency on a DEX architecture seems to require a significant compromise in the degree of decentralization.

This "nonzero" residual has important theoretical implications. It clearly shows that decentralization has a cost. Here we must distinguish two related but different trade-offs. The blockchain trilemma proposed by Buterin (decentralization, security, scalability) describes a design constraint at the level of underlying infrastructure. This chapter's analysis reveals an independent trade-off along the dimension of financial-market pricing efficiency: the fundamental tension between efficiency and decentralization. The latter is not a mere restatement of the former but shifts the perspective from the design space of technical architecture to the efficiency boundary of market microstructure. In

this efficiency-decentralization trade-off, a CEX, by sacrificing decentralization and concentrating trust in a single entity, achieves sub-millisecond execution speed; a DEX, in order to achieve trustless, censorship-resistant trading, must accept the physical latency imposed by distributed consensus. The ineliminable "structural arbitrage residual" between CEXs and DEXs is precisely the quantifiable expression of this fundamental trade-off in market prices.

From the perspective of the persistence spectrum, therefore, this residual is a highly persistent structural anomaly under the current technological paradigm. It cannot be eliminated by simple technical upgrades unless the core value proposition of decentralization is abandoned. Understanding this is essential for viewing the development of DEXs objectively, assessing their competitive relationship with CEXs, and designing more effective market mechanisms. It tells us that pursuing complete DEX superiority over CEXs along every dimension may be an unrealistic goal; instead, genuine innovation lies in how to maximize the unique value that decentralization brings—such as self-custody of assets, transparent rules, and censorship resistance—while acknowledging and quantifying this efficiency cost. The qualifier "current technological paradigm" is essential, however. Emerging technical paths such as integrating trusted execution environments (TEEs) into validator nodes, encrypted mempools (such as Flashbots' SUAVE architecture), and commit-reveal schemes could in theory narrow this residual without abandoning decentralization, but these schemes themselves introduce new security assumptions and trust models whose effectiveness and safety have not yet been tested at scale in real-world conditions.

18.8 The crisis dynamics of anomalies

In the preceding sections, we have systematically examined the six categories of anomalies in the perpetual futures market. Under normal market conditions, these anomalies exist in relatively independent forms, their magnitude usually held within a controllable range by the force of arbitrageurs. The basis anomaly, the funding-rate anomaly, the cross-exchange anomaly, the cross-asset anomaly, the term-structure anomaly, and the CEX–DEX anomaly each reveal, along their own dimension, the boundary of market efficiency and the contours of arbitrage constraints. This relatively calm state, however, is not the whole truth of the market. When extreme stress arrives, the interaction among these seemingly independent anomalies intensifies, triggering a systematic pricing deviation that sweeps across the entire market.

The core argument of this section is that these anomalies do not amplify linearly in a crisis but display highly nonlinear, synchronously erupting dynamics. This is precisely the observable consequence, at the level of market price deviations, of the "risk resonance" model proposed in Chapter 17. When systemic risk is triggered and the seven-dimensional arbitrage constraints are activated simultaneously within a short time, arbitrageurs' capital and capacity are sharply weakened, and the market's arbitrage forces temporarily fail. As a result, all latent price deviations worsen sharply within the same time window and amplify one another, forming a synchronous eruption of anomalies that sweeps across all assets and platforms. This section examines this crisis dynamic, revealing—from the theoretical, empirical, and case levels—what

unstable characteristics the market takes on when all deviations erupt simultaneously.

Before proceeding, an important methodological statement is needed up front to delimit the nature and boundaries of this section's argument. The crisis-contagion path presented in this section (the six-stage table in Section 18.8.2) and the single crisis case (the LUNA/UST episode of May 2022 in Section 18.8.3) are a **theoretical-inference framework and a single-case illustration**, constructed from market-microstructure theory and observed crisis patterns, and **not a rigorous hypothesis test**: the step-by-step causal relationships among the stages of the contagion path have not been verified one by one by independent empirical research, and the LUNA/UST case only **illustrates** the observable manifestation of the risk-resonance mechanism in one real crisis rather than constituting a statistical proof of the mechanism. In reading the contagion table (Table 18-8) and the case analysis that follow, therefore, readers should understand them as "a representative narrative consistent with the risk-resonance model" rather than "an established general regularity." The falsifiability condition of the risk-resonance mechanism can be stated as follows: if, in a class of crises satisfying the model's premises (a systemic shock is triggered and multiple constraint dimensions tighten simultaneously), the anomalies of the various dimensions do not worsen in synchrony and cross-asset correlation does not break down, then the model does not hold for that class of cases. The goal of this section is to make this mechanism concrete as observable changes in indicators and to state clearly that it awaits testing through multi-case comparison.

18.8.1 The nonlinear amplification of anomalies

In theory, the nonlinear amplification of anomalies in a crisis is a direct consequence of the collapse of arbitrage capacity. In a normal market, arbitrage capital is ample, and the various arbitrage strategies (such as basis arbitrage, rate arbitrage, and cross-exchange arbitrage) serve as effective regulating mechanisms that keep price deviations within a narrow channel composed of transaction costs and risk premiums. When a systemic shock occurs (for example, a major macro event or the collapse of a core protocol), however, the risk-resonance mechanism described in Chapter 17 is activated. Here it is necessary to distinguish two types of shock that differ in their contagion dynamics. Endogenous crypto-native shocks (such as the LUNA collapse and the FTX bankruptcy) transmit primarily through the intra-chain contagion path described above, characterized by outward spread from a specific asset or platform. Exogenous macro shocks (such as a Fed rate surprise or the SVB bank run) strike the entire market simultaneously through the dollar-liquidity channel and often cause CEX and DEX pricing to deviate in synchrony before any intra-chain contagion even occurs. The six-stage path modeled in Section 18.8.2 is more applicable to the former; for the latter, the shock's entry point and propagation sequence may differ significantly. Constraints across multiple dimensions—liquidity, margin, platform risk, and smart-contract risk—tighten simultaneously, arbitrageurs' capital is depleted by losses, their risk exposure expands as hedges fail, and their capacity for market repair declines sharply (superlinearly) within a short time.

This nonlinear collapse of arbitrage capacity directly causes the nonlinear amplification

of anomalies. Once a price deviation exceeds a certain critical point, it is no longer effectively suppressed and may instead trigger a positive feedback loop. For example, a widening basis may trigger the liquidation of more leveraged positions, and the liquidations themselves widen the basis further. To capture this regime transition from calm to storm conceptually, one can envisage constructing a "composite anomaly index" as a diagnostic tool. The idea of the index is this: standardize the proxy indicators of the six categories of anomaly described above (for example, the absolute value of the basis, the absolute value of the funding rate, and the spread among leading exchanges) and then take a weighted average, yielding a single indicator that measures the market's overall "degree of deviation." This chapter does not actually construct and estimate this index, however: the precise definition of each proxy indicator, the standardization method, the source of the weights, the sample period, and the actual values in crisis periods all await determination by dedicated research, so it is here only a **constructible diagnostic-indicator concept**, not a completed empirical measurement. Under this concept, the risk-resonance model predicts that the index should fluctuate within a narrow band around a low mean under normal conditions but may experience a nonlinear "jump" during a crisis, with a magnitude far exceeding the standard deviation of normal fluctuation; if this jump were actually observed, it could serve as a candidate signal of the market's transition from a "normal state" to a "crisis state."

18.8.2 The contagion path among anomalies

The eruption of anomalies in a crisis does not occur independently but spreads through complex paths, forming a self-reinforcing network. A crisis usually begins with the sharp worsening of an anomaly in one dimension, but its impact transmits rapidly, within minutes, to all other dimensions through arbitrageurs' linked positions and the market's intrinsic connections. Understanding these contagion paths is the key to understanding crisis dynamics. The contagion path presented in the table below is a theoretical-inference framework constructed from market-microstructure theory and observed crisis patterns; the step-by-step causal relationships among its stages have not been verified one by one by rigorous empirical research, but the overall logic is highly consistent with the observations of historical cases such as LUNA/UST.

Contagion stage	Trigger mechanism	Direction of contagion	Concrete manifestation
Stage 1	External shock triggers mass selling	Shock → basis	Perpetual futures price plunges, producing a large negative basis
Stage 2	Negative basis causes losses on basis-arbitrage trades	Basis → closing of basis-arbitrage trades	Arbitrageurs are forced to sell spot and buy perpetual to stop losses
Stage 3	Large-scale closing depletes liquidity	Basis-arbitrage closing → liquidity	Bid-side depth plunges, slippage costs surge
Stage 4	Liquidity evaporation obstructs cross-exchange arbitrage	Liquidity → cross-exchange spread	Prices on different exchanges run independently, the law of one price breaks down
Stage 5	Extreme long-short imbalance	Liquidity → funding rate	The funding rate is pushed to an extreme negative value, intensifying long liquidations
Stage 6	Indiscriminate selling pressure	Deleveraging → cross-asset correlation	All assets are sold off, and correlation shifts from highly positive to disordered
Feedback loop	Hedge failure amplifies overall risk	Cross-asset → basis	Altcoin crashes force the liquidation of BTC hedge positions, intensifying basis deviation

Table 18-8. The contagion stages and paths of anomalies in a crisis (Data source: compiled by the author)

Table 18-8 presents how anomalies spread through a positive-feedback-loop network in a crisis: an external shock first tears open a gap in the basis dimension (the perpetual futures price plunges and the negative basis widens sharply), then transmits stage by stage within minutes along the chain "basis arbitrage forced to close → liquidity evaporation → cross-exchange spread breaks down → funding rate goes to extremes → cross-asset correlation collapses," with the worsening of each link in turn aggravating the previous one, finally closing into a self-reinforcing positive feedback loop (the trigger mechanism, direction of contagion, and concrete manifestation of each stage are detailed in Table 18-8). The key to this loop is its self-amplification: the collapse of correlation further increases the difficulty of portfolio hedging, amplifies overall market risk, and feeds back to the basis once more—for example, a fund may be forced to liquidate its BTC hedge position because its altcoin positions have crashed, thereby intensifying volatility and basis deviation in the BTC market.

The six-stage path above presents an idealized linear cascade; contagion in a real crisis often proceeds through multiple channels simultaneously. In the LUNA/UST episode, the stablecoin depegging simultaneously struck DEX liquidity pools (such as the violent rebalancing of the Curve 3pool) and the liquidation engine of the perpetual futures

market, and these two channels did not transmit in stage order. Moreover, the table above omits an independent contagion channel that was dominant in the FTX and Celsius/Voyager crises: an exchange freeze or bankruptcy that locks up user funds forces arbitrageurs to close cross-exchange hedge positions urgently at distorted prices on other platforms, and this "capital-freeze channel" can be triggered before or simultaneously with Stage 1, independently of organic market selling.

The contagion in this process is extremely fast, often completed within minutes. It reveals an important conclusion: market anomalies are not isolated events but an interconnected system. The collapse of one link can, through arbitrageurs' balance sheets and the market's microstructure, quickly escalate into a full-blown systemic crisis.

18.8.3 Case: the LUNA/UST crisis of May 2022

The exposition of theory and models needs to be illustrated with real-world cases. The LUNA/UST collapse of May 2022 [4][15] offers us a **single-case illustration** for observing the crisis dynamics of anomalies. The methodological statement at the head of this section bears repeating: the following analysis is a descriptive examination of one real crisis, used to illustrate the observable manifestation of the risk-resonance mechanism in that event, and is **not a hypothesis test of the mechanism**; a single case cannot rule out competing explanations, and establishing "the synchronous eruption of the six categories of anomalies in a crisis" as a general regularity still requires a systematic cross-sectional comparison of multiple crises (such as the FTX bankruptcy and the SVB shock). The following analysis is based on the public market data of major exchanges such as Binance, OKX, Bybit, and CME during May 5 to 13, 2022, and the indicators examined include the perpetual futures basis, the funding rate, the cross-exchange spread, the rolling correlation coefficient, and the CEX–DEX spread. Within just a few days, this algorithmic stablecoin ecosystem—once worth tens of billions of dollars—collapsed completely, triggering one of the most violent market upheavals in crypto history. The event demonstrated how the six categories of anomalies discussed above worsen in synchrony within the same time window.

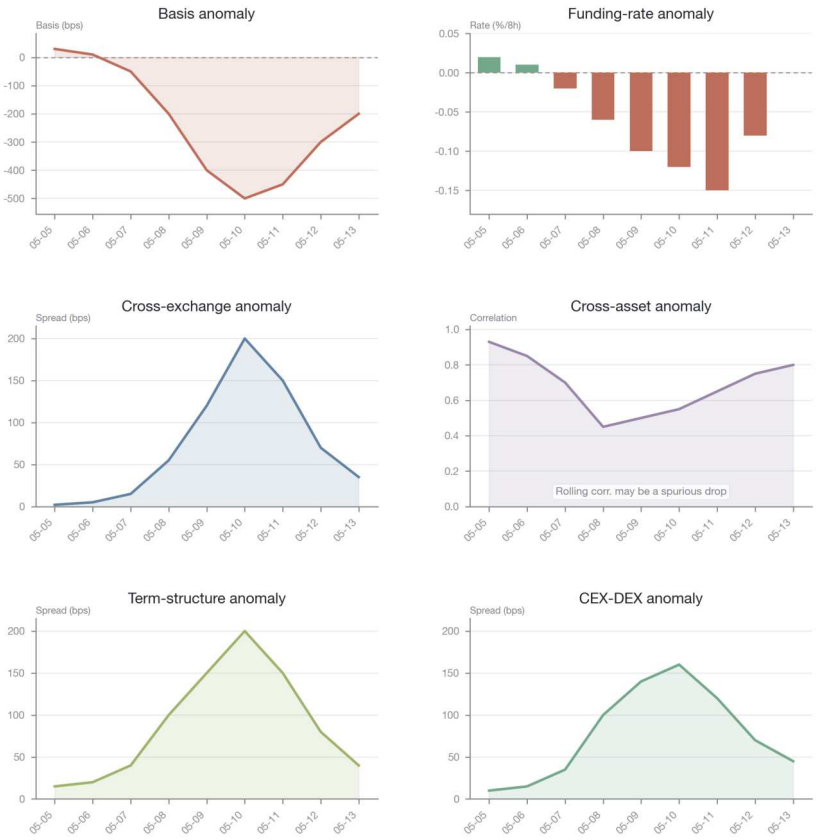


Figure 18-16. The synchronous eruption of the six categories of anomalies in May 2022 (illustrative/simulated: the data source is readings from historical Coinglass charts, and the specific basis-point values are reconstructed illustrative values rather than verifiable precise measurements; the cross-asset correlation panel may reflect a spurious decline under asynchronous price discovery and is illustrative only)

Figure 18-16 uses a multi-panel timeline to display the synchronous, drastic changes in the key indicators of the six anomalies during May 5 to 13, 2022. On the basis anomaly: before the crisis erupted, BTC perpetual futures maintained a positive basis of about 30 bps against spot; after the news of the UST depegging spread from May 8, it turned sharply negative, falling to a historic low of −500 bps on May 10 (data source: Coinglass). On the funding-rate anomaly: highly synchronized with the basis, it turned rapidly from about +0.02%/8h before the crisis to deeply negative, hitting an extreme level of −0.15%/8h on May 11, with the clustering effect of extreme negative rates emerging. On the cross-exchange anomaly: the BTC perpetual spread among leading exchanges widened from a normal 2–3 bps to more than 100 bps, peaking at 200 bps (tens of times the normal level), during the deepest phase of the crisis on May 10–11, and the market was split into relatively independent pricing zones. On the cross-asset anomaly: the correlation between BTC and ETH plunged from above 0.9 before the crisis to below 0.5 at the selling climax on May 10. (This correlation is estimated as the Pearson coefficient over a 24-hour rolling window, **without bias correction and for**

illustration only: a rolling correlation coefficient may produce a spurious downward signal in a high-volatility environment because of asynchronous price discovery, changes in the volatility level, and the like, so the "plunge in correlation" above should be interpreted with caution; a more rigorous contagion test should use the bias-correction method proposed by Forbes and Rigobon (2002) [16] or the DCC-GARCH dynamic conditional correlation model of Engle (2002) [10]—this is a suggested direction for robust testing rather than a method already adopted in this section.) On the term-structure anomaly: the two curves that should converge—the CME dated-futures basis and the offshore perpetual futures implied rate (funding rate)—diverged enormously, with the spread surging from a normal dozen or so basis points to nearly 200 bps. On the CEX–DEX anomaly: the spread widened sharply from a normal 10–15 bps to more than 150 bps, exposing the cost of the DEX's oracle latency, on-chain congestion, and lower liquidity depth, which leave it unable to keep up with CEX pricing efficiency in extreme conditions.

In this single case, the six panels together sketch a consistent picture: in the second week of May 2022, multiple efficiency dimensions of the market came under severe stress almost simultaneously. All deviations were no longer small, controllable noise but a manifestation of systemic risk that worsened in synchrony and reinforced one another. The collapse of LUNA/UST, as a major shock event, saw its local impact ultimately escalate into a market-wide crisis. The observations above are consistent with the predictions of the risk-resonance model; but as a single-case illustration they can only **support**, not **prove**, its generality; whether their consistency with the model recurs robustly in other crises is an open question awaiting further empirical testing.

18.8.4 The asymmetry of eruption and repair

The final, and highly practical, feature of crisis dynamics is the pronounced asymmetry between the eruption and the repair process. As we saw in the May 2022 case, the full eruption of anomalies can be completed within a few hours, whereas the market's repair process takes weeks or even months. This asymmetry—"the eruption completes within hours, while the repair often takes weeks"—reveals how slowly market confidence recovers and how gradually capital is rebuilt.

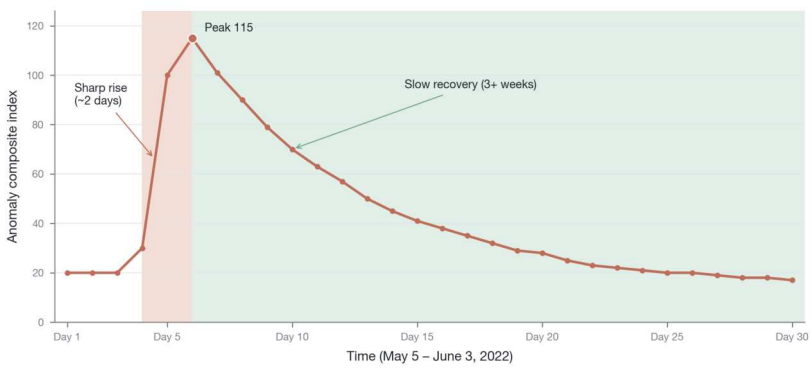


Figure 18-17. The eruption-repair asymmetry of the composite anomaly index (Data source: simulated data constructed by the author)

Through a simulated "composite anomaly index" time series, Figure 18-17 captures this asymmetry: the eruption phase of the crisis is steep and abrupt, with the index surging from a low level to its peak in a very short time. Behind this lies the positive feedback loop of risk resonance—liquidations trigger price declines, price declines trigger more liquidations, margin is depleted, and positions are forcibly liquidated, all completed in a very short time.

By contrast, the curve of the repair phase is gradual and drawn out. The market falls back from its peak far more slowly than it rose. This slow repair stems from its "rebuilding" nature, involving a gradual process at multiple levels. First is the rebuilding of capital: in a crisis, a large amount of arbitrage capital and market-maker capital is destroyed, and the remaining capital becomes extremely risk-averse because of severe losses, so new capital entering the market needs time to assess risk, raise funds, and build positions gradually. Second is the rebuilding of confidence: financial markets are built largely on confidence, and a violent crash severely damages investors' confidence in market stability, platform reliability, and even asset value itself; the recovery of confidence requires the market to display sustained stability over a period before participants restore their risk appetite. In addition, the repair speed of different anomalies diverges: anomalies that usually depend on high-frequency trading capability (such as the cross-exchange spread) repair first, because high-frequency trading firms can quickly redeploy their algorithms once the market stabilizes even slightly, whereas anomalies that depend on large-scale capital and lower risk tolerance (such as the basis and term-structure anomalies) take longer, and the capital-rebuilding and risk-assessment process of basis arbitrageurs and structural arbitrageurs is far more protracted after they suffer heavy losses.

This recovery ordering is not a universal rule, however; it depends on the nature of the crisis. In an organic market decline (such as deleveraging triggered by a macro shock), the cross-exchange spread does indeed repair first, thanks to the rapid redeployment of high-frequency arbitrageurs. In an exchange-bankruptcy crisis (such as the FTX episode), however, a large amount of arbitrage and market-making capital is locked on the bankrupt platform and cannot be withdrawn, and the recovery ordering may reverse entirely: the cross-exchange spread persists for weeks because arbitrage capital is frozen, while the basis anomaly instead recovers relatively quickly once liquidity normalizes on surviving exchanges such as CME and Binance. The more precise causal mechanism is the funding-liquidity constraint: high-frequency strategies need only intraday margin, so their funding constraint is loose, whereas basis arbitrage must lock up two-leg margin, so its funding constraint is tight; but when the capital of a specific platform is physically frozen, the distribution of the funding-liquidity constraint is fundamentally altered.

This asymmetry between eruption and repair has an important statistical implication: it systematically raises the long-run average level of market anomalies. Even if the market is in an efficient "normal state" for the overwhelming majority of the time, the few "crisis-state" periods and the long repair tails that follow significantly affect our overall assessment of market efficiency (if quantification is desired, the "crisis state" can be op-

erationally defined as the share of days on which the aforementioned composite anomaly index exceeds a certain threshold, but this chapter does not measure that share, and the "overwhelming majority / few" here is an illustrative statement rather than a precise statistic). It reminds us that the market's efficiency loss is lasting and that the impact of one crisis takes a long time to be eliminated. Moreover, the post-crisis regulatory response is an important factor prolonging the repair period. Take the LUNA/UST collapse: the regulatory uncertainty produced by the subsequent enforcement actions (including the arrest warrant for Do Kwon and the SEC's securities-classification ruling on LUNA/UST) suppressed arbitrage capital's willingness to re-enter, constituting an external recovery resistance beyond market microstructure. This slow repair process itself constitutes an important source of the frictions and deviations that persist in the market over the long run.

18.9 The evolution of anomalies and implications for market design

The analysis running through this chapter has revealed a series of persistent anomalies in the perpetual futures market that record the history of arbitrage forces being obstructed along different dimensions. These anomalies are not static, unchanging features, however. On the contrary, they form a panorama of dynamic evolution: some old deviations gradually fade as market infrastructure matures, some stubborn deviations persist over the long run because they arise from the market's deep structural constraints, and, at the same time, new market innovations continually give rise to new anomalies of varied forms. The purpose of this section is precisely to survey this dynamic picture systematically and to distill from it normative implications for future market design: what exactly can be repaired through institutional improvement, and what is a fundamental cost that must be accepted under the current architecture.

Understanding the evolutionary trajectory of anomalies is a core yardstick for assessing a market's maturity. A market's progress in efficiency is most intuitively reflected in the speed and breadth with which it eliminates price deviations. This section therefore unfolds along three temporal dimensions: reviewing the anomalies that are fading, examining those that persist, and looking ahead to those that are emerging (Table 18-9). Finally, we converge all of this on a practical question: how does one design a better market?

Evolution category	Representative anomaly	Current status	Driving factor
Fading	Cross-exchange spread among leading exchanges	Narrowed from about 3.5% (2017–2018 sample, Makarov-Schoar [17]) to 1–3 bps (2025)	Entry of high-frequency trading, API standardization
Fading	Calendar effects (weekend effect)	Essentially eliminated	24/7 quantitative funds and automated strategies
Fading	Excess returns of basis arbitrage	Annualized returns fell significantly from 30%–80%	Coincided with the expansion of protocolized arbitrage tools such as Ethena (causality pending proof)
Persistent	CEX–DEX structural residual	Shrinking but not to zero	Physical confirmation-time limit of the blockchain
Persistent	Long-term positive bias in funding rates	Changing slowly	Retail-heavy participant structure
Persistent	Regulation-segmented pricing zones	Requires regulatory change	Capital controls and a non-unified regulatory framework
Emerging	CME/ETF dual-track pricing	Formed after 2024	Approval and listing of Bitcoin spot ETFs
Emerging	Concentration risk of protocolized arbitrage	Developing	TVL concentration of protocols such as Ethena
Emerging	Microstructure effects of AI trading agents	Early stage	AI strategy convergence and algorithmic gaming

Table 18-9. A panorama of the evolution of perpetual futures market anomalies (Data source: constructed by the author from public market data and the literature)

Table 18-9 places the principal anomalies of the perpetual futures market into three categories according to their evolutionary trajectory: fading, persistent, and emerging. The anomalies in the fading category (such as the cross-exchange spread among leading exchanges narrowing from 3.5% to 1–3 bps) confirm, through the marked convergence of their magnitude, the gradual strengthening of arbitrage forces. The anomalies in the persistent category (such as the CEX–DEX structural residual and regulation-segmented pricing zones) mark the boundaries that arbitrage forces cannot cross under the current market architecture. The anomalies in the emerging category (such as CME/ETF dual-track pricing and the concentration risk of protocolized arbitrage) reveal how structural change in the market creates new efficiency gaps even as it eliminates old deviations. Together, the three categories form a panorama of the dynamic evolution of market efficiency.

18.9.1 Fading anomalies

The market is not standing still; the fruits of arbitrageurs' work are clearly visible over time. Some anomalies that were significant in the market's early stages have narrowed greatly in magnitude, or even nearly disappeared, as the arbitrage infrastructure has advanced and capital has deepened. These fading anomalies are the strongest evidence for testing the AIH framework of Chapter 16: arbitrage returns, as compensation for the infrastructure services arbitrageurs provide, tend to decline as the cost of those services falls.

The most typical example is the cross-exchange spread among leading exchanges. In the early days of the crypto market, the Bitcoin spread between different exchanges often reached several percent; Makarov and Schoar's study based on 2017–2018 data recorded cross-market spreads as high as several percentage points (about 3.5%) [17] (2020 is the year of that paper's publication), which provided rich profits for early cross-exchange arbitrageurs. As high-frequency trading firms entered, API interfaces were standardized, and features such as cross-exchange unified accounts became widespread, the latency and cost of executing cross-exchange arbitrage were compressed sharply. By 2025, the Bitcoin perpetual futures spread among first-tier exchanges such as Binance, OKX, and Bybit had, under normal market conditions, stabilized within an extremely narrow band of 1–3 bps (0.01%–0.03%), approaching the theoretical cost lower bound of cross-exchange arbitrage. The convergence of the spread from the percentage level to the basis-point level amply demonstrates the remarkable success of arbitrage capital and technology in "leveling" the surface of market prices.

Another anomaly that has been almost entirely "arbitraged away" is calendar effects, such as the "weekend effect" or holiday effect widespread in the early market. In the market's early days, because the participant structure was dominated by retail investors and institutional traders took weekends off, the market's liquidity and pricing efficiency declined systematically on weekends, causing volatility and spread patterns to show predictable intra-week variation. As 24/7 quantitative funds and automated trading strategies became the market's dominant force, however, any predictable pattern tied to a specific time was quickly identified and exploited until it was no longer profitable. Today, except during extreme events, the pricing behavior of major assets has essentially eliminated the significant difference between weekdays and weekends.

The excess returns of basis arbitrage likewise show a downward trend. In the bull market of 2020 to 2021, perpetual futures maintained a high positive basis against spot over the long run, providing arbitrageurs who shorted perpetual and bought spot with a significant annualized return as high as 30%–80%—this high return is itself a manifestation of insufficient arbitrage forces and the market's lack of enough capital to meet vigorous long demand. As protocolized arbitrage tools such as Ethena rose (see Section 18.2.4 for the mechanism), however, the market's short supply was greatly strengthened. Preliminary time-series evidence shows that after Ethena's TVL grew rapidly in 2024, the fluctuation center of both the basis and the funding rate of major assets' perpetual futures declined observably, consistent with the direction of the hypothesis that "protocolized arbitrage systematically compresses the bull-market positive basis." This co-movement, however, is not yet sufficient to establish causation: the same 2024

period was also overlaid with confounding factors such as the macro interest-rate environment and an overall market-regime shift, and attributing the decline in the basis center cleanly to Ethena still requires controlling for these confounders and supplementing the analysis with more rigorous econometric testing. On the premise of treating the direction of causation cautiously, one can say that a high-barrier arbitrage opportunity that once required professional hedge funds to access is being democratized by technological and protocol innovation, and one possible outcome is the fading of the related anomaly.

Behind every fading anomaly is a story of "strengthening arbitrage forces." Whether through the entry of more professional participants, the deployment of more abundant capital, or the application of more advanced technical tools, all are jointly filling the weak links in the market's efficiency network. These fading price deviations are the most powerful empirical support for the market's move toward maturity, verifying that arbitrage, as a corrective force, though delayed, ultimately takes effect.

18.9.2 Persistent anomalies

In sharp contrast to the anomalies that gradually vanish are certain structural anomalies that display pronounced stickiness, their form not having changed fundamentally over several years. The persistence of these anomalies does not mean that arbitrageurs are insufficient or that the market has failed; on the contrary, they delimit the boundaries that arbitrage forces cannot cross—the structural constraints determined by the market's core architecture, the laws of physics, or fundamental institutional barriers.

The CEX–DEX spread is the best example illustrating this class of anomaly. As described in Section 18.7.3, the "structural arbitrage residual" rooted in the blockchain's physical confirmation time cannot be reduced to zero—it is the quantifiable efficiency cost paid for "decentralization" and "trustlessness," and even as technological advances such as L2 scaling, high-performance appchains, and low-latency oracles continually compress latency, it can only be narrowed and not eliminated, and therefore stays sticky over the long run (its full mechanism and the efficiency-decentralization tradeoff are detailed in Section 18.7.3).

The long-term positive bias in funding rates is another stubborn structural anomaly. As analyzed in Section 18.3.1, its root is the "structural premium" determined by the retail-heavy participant structure, which compensates shorts for the counterparty risk and contrarian risk they bear; as long as the "retail-heavy, institution-neutral" participant structure does not undergo a fundamental reversal, this anomaly will persist (its causes are detailed in Section 18.3.1).

The segmented pricing zones created by regulatory barriers are the structural anomaly hardest to shake at the institutional level. As described in Section 18.4.3, the Korean "kimchi premium" persists because of strict capital controls and foreign-exchange restrictions. South Korea's Virtual Asset User Protection Act of 2023 further regulated the operations of domestic exchanges, but the reporting and approval obligations for large cross-border transfers under the Foreign Exchange Transactions Act were not relaxed as a result; moreover, this control does not treat all sizes alike—small personal transfers face lower friction, whereas the large, frequent cross-border capital deploy-

ment needed to support institutional-scale arbitrage remains costly in compliance terms. This "scale-dependent barrier" precisely explains why the kimchi premium, although it does not reach a fully segmented level thanks to the presence of small arbitrageurs, still significantly exceeds what transaction costs can explain. The elimination of this class of anomaly (together with the systematic difference between CME/Coinbase and Binance/OKX caused by U.S. investors being restricted from accessing offshore exchanges) depends not on technology or capital but entirely on the coordination and unification of regulatory frameworks among the world's major economies, and a fundamental breakthrough is unlikely in the foreseeable future.

In addition, as described in Section 18.5.1, the "inefficiency zone" of altcoin perpetual futures is a structural anomaly stemming from the "long-tail distribution" of liquidity—the order books of the vast number of small- and mid-cap tokens are extremely shallow, so even when arbitrageurs spot a significant deviation, they cannot execute effectively because of the enormous trading-impact cost; this efficiency stratification determined by liquidity stratification is a structure that forms spontaneously in the market and is not easily changed by external intervention (see Section 18.5.1).

These persistent anomalies remind us that we cannot simply regard every price deviation as "market failure." They should instead be seen as "equilibrium outcomes" under specific constraints. They are the monetized expression of the "friction cost" the market must pay to keep functioning under its existing technology, institutions, and participant structure. Eliminating them requires not "moral suasion" of arbitrageurs but systematic change to the constraints themselves.

18.9.3 Emerging anomalies

The market's evolution not only eliminates old anomalies but also creates new ones. Every major change in market structure—such as the introduction of a new product, the application of a new technology, or the entry of a new participant—may open new "fissures" even as it reshapes the efficiency map. We are currently in an era of change driven jointly by ETFs, protocolized arbitrage, and AI trading, and a series of unprecedented new anomalies is emerging.

The approval and listing of Bitcoin spot ETFs in the United States in early 2024 was an important structural event for the crypto market in recent years. While opening a compliant gateway to the mainstream financial world, it also created a wholly new, still-forming structural anomaly: the "dual-track pricing" of the CME/ETF system and the offshore perpetual futures system. As analyzed in Section 18.6.2, these two systems differ fundamentally in regulatory framework, trading hours, settlement currency, and settlement cycle: the segmentation between SEC regulation and offshore regulation, the mismatch between the New York trading session and round-the-clock trading, the difference between dollar settlement and USDT/USDC settlement, and the lag between T+1 settlement and instant settlement. Together, these differences form an "institutional obstacle" that is hard to surmount, obstructing effective cross-system arbitrage. As a result, the two markets frequently give inconsistent pricing signals for the same underlying asset (Bitcoin). Especially while the U.S. market is closed, the offshore perpetual futures market becomes the world's only continuous pricing center,

and lacking the "price anchor" of the ETF and CME futures, its price is more prone to deviation from regional news or liquidity shocks. This new fissure caused by institutional segmentation and time mismatch did not exist before 2024, and its evolution will become a core topic of market-structure research in the coming years.

The rise of protocolized arbitrage likewise cuts both ways. Protocols represented by Ethena, through their enormous capital volume, effectively play the role of an "anomaly-suppression mechanism" under normal market conditions, systematically reducing the fluctuation of the basis and the funding rate. This model of "protocolizing" and "centralizing" arbitrage activity, however, also carries the risk of creating a new type of crisis dynamics: once the market shifts into a sustained, deep negative funding-rate environment and triggers a large-scale wave of redemptions, the protocol is forced to close out its enormous short perpetual futures positions en masse within a short time, which could push the basis to an unprecedented, extreme one-directional level, forming a new type of transient anomaly triggered by the protocol's own mechanism (the full exposition of this "redemption → forced unwinding → extreme basis" mechanism is in Section 18.2.4). This risk of the "protocol being forced into large-scale unwinding" has not yet occurred in reality but is entirely possible in theory, and it reveals that market innovation, while raising efficiency, may redistribute and concentrate tail risk in more complex forms.

The gradual spread of artificial intelligence (AI) trading agents portends another important microstructural change. On one hand, AI can identify and eliminate existing arbitrage opportunities—especially the tiny ones hidden in vast data—through greater speed and more complex models, further accelerating the fading of traditional anomalies. On the other hand, when the market's main participants shift from human traders to AI agents of every kind, the gaming among them may create wholly new microstructural patterns that humans struggle to understand. For example, strategy convergence and resonance may arise among AIs. This risk is not entirely new; the literature has documented the instability that algorithmic-strategy crowding caused in traditional markets. Kirilenko et al.'s analysis of the 2010 "Flash Crash" shows that the same-direction behavior of algorithmic traders was a key amplifier of the price collapse [18] ("2010" in the text refers to the year of the event; the authoritative analysis is the version published in The Journal of Finance in 2017). But AI agents introduce a new dimension: when major models are trained on overlapping training datasets, they may share a structural bias regarding market-regime shifts. This shared bias is invisible under normal conditions, but when a real macro-regime shift occurs, all models may emit the same trading signal simultaneously, producing a "slow-motion" liquidity evaporation—unlike a millisecond-scale flash crash, this is a correlated flow shock lasting several hours, and it may be more destructive because it operates below traditional anomaly-detection thresholds; alternatively, the complex camouflage or deception strategies that AIs adopt to game one another may leave non-random "fingerprints" in the price series, constituting a wholly new anomaly based on algorithmic gaming. The dual-sided effect of AI—both terminator of anomalies and potential creator of new ones—marks the frontier of the evolution of market efficiency.

The core insight is that anomalies form a dynamically balanced ecosystem. Every advance of the market—whether product innovation, protocol innovation, or technological

innovation—simultaneously eliminates some old arbitrage space and, by changing market structure, creates new efficiency gaps. The assessment of market efficiency therefore cannot rest on the analysis of a static snapshot in time but must turn to the continual tracking of the dynamic evolution of anomalies.

18.9.4 From anomalies to institutional improvement

The ultimate purpose of analyzing anomalies is not only to understand the world but to change it. Translating the empirical analysis of the causes and evolution of anomalies above into normative recommendations for market design is the final destination of this chapter. For market designers and regulators, the core question is this: faced with the many and varied market anomalies, which are defects that can be "repaired" through institutional and technical improvement, and which are "costs" we must accept that stem from fundamental trade-offs in the market's architecture?

On repairable anomalies: the root of some anomalies lies in the imperfection of current institutional design or technical facilities, and they can be mitigated or eliminated through a clear path. We can rank these institutional-improvement measures along the two dimensions of "scope of impact" and "feasibility," forming a pragmatic reform roadmap (Table 18-10).

Anomaly type	Core cause	Institutional/technical improvement required	Scope of impact	Feasibility/timeline
8-hour settlement pulse	Discrete funding-rate settlement cycle	Shorten the settlement cycle to 1 hour or continuous settlement	Medium	High / already under way
DEX oracle latency	Oracle update frequency and on-chain confirmation time	Low-latency oracles (such as Pyth) + high-performance appchains	High	High / short-to-medium term (1–3 years)
Cross-exchange spread and rate splitting	Capital fragmentation and low margin efficiency	Cross-exchange margin mutual recognition or unified clearing (such as ClearToken)	High	Medium / medium term (3–5 years)
Slow negative-basis repair	High friction in the spot coin-borrowing market	Develop more efficient, more liquid centralized or decentralized lending markets	Medium	Medium / medium term (3–5 years)
Core-periphery pricing gradient	Counterparty risk of small and mid-sized exchanges	Introduce an independent exchange credit-rating system and proof-of-reserves standards	Medium	Medium / medium term (3–5 years)
Regulation-segmented pricing zones	Capital controls and a non-unified regulatory framework	Coordination and mutual recognition of regulatory frameworks among the world's major economies	Very high	Low / long term (more than 5–10 years)

Table 18-10. Repairable anomalies and the corresponding institutional-improvement roadmap (Data source: constructed by the author from public market data and the literature)

A priority sequence is clearly visible from this roadmap. Upgrading oracle technology and shortening the settlement cycle are the improvement measures with the highest feasibility at present, and their impact is significant. Emerging DEXs such as Hyperliquid have, through hourly or even more frequent settlement, significantly weakened the "settlement-timing effect" brought by traditional 8-hour settlement. Likewise, the new generation of oracles and appchain architectures is compressing the CEX–DEX latency gap from the second level to the sub-second level. These are technology-driven improvements already under way.

In the medium term, cross-exchange margin mutual recognition is the core measure

for solving capital fragmentation and the amplification of anomalies during crises. At present, arbitrageurs must post large amounts of margin at each exchange, greatly reducing capital efficiency. If net settlement of cross-exchange margin could be achieved through a third-party clearinghouse (analogous to LCH or CME Clearing in traditional finance), it would greatly free up arbitrage capital and strengthen the market's resilience under stress. Projects such as ClearToken are exploring in this direction. Achieving cross-exchange margin mutual recognition, however, requires not only overcoming the coordination game and trust problems among exchanges but also having the participating institutions obtain central counterparty (CCP) clearing qualification in major jurisdictions (such as CCP recognition under the EU's EMIR framework or DCO registration in the United States), and these regulatory-approval processes are themselves highly uncertain. In an optimistic scenario, preliminary bilateral netting arrangements may appear within 3–5 years, but the timeline for establishing a comprehensive multilateral clearing system is hard to predict.

In the long term, the coordination of global regulatory frameworks is undoubtedly the most far-reaching reform, capable of fundamentally eliminating the segmented pricing zones caused by regulatory arbitrage and market segmentation. This touches the core of each nation's financial sovereignty, however, and is the most difficult, beyond what market forces themselves can determine. Table 18-11 summarizes the expected progress of the improvement measures above along the temporal dimension.

Temporal dimension	Eliminable anomaly	Improvement required	Current progress
Short term (1–2 years)	8-hour settlement pulse	Shorten the settlement cycle to 1 hour or continuous settlement	Hyperliquid and others have implemented more frequent settlement
Short term (1–2 years)	DEX oracle latency	Low-latency oracles + high-performance appchains	New-generation oracles such as Pyth and Stork have been deployed
Medium term (3–5 years)	Cross-exchange spread and rate splitting	Cross-exchange margin mutual recognition or unified clearing	Projects such as ClearToken are exploring
Medium term (3–5 years)	Slow negative-basis repair	More efficient lending markets	Decentralized lending protocols continue to develop
Medium term (3–5 years)	Core-periphery pricing gradient	Independent credit-rating system and proof of reserves	Third-party ratings such as Kaiko are gradually improving
Long term (more than 5–10 years)	Regulation-segmented pricing zones	Coordination and mutual recognition of global regulatory frameworks	Regulatory progress varies by country; coordination is extremely difficult

Table 18-11. The temporal dimension of anomaly elimination and current progress (Data source: constructed by the author from public market data and the literature)

Table 18-11 reorganizes the institutional-improvement measures of Table 18-10 along the temporal dimension, dividing them into three tiers—short term (1–2 years), medium term (3–5 years), and long term (more than 5–10 years)—and noting the current imple-

mentation progress of each. Tables 18-10 and 18-11 cover the same set of six improvement measures, the former emphasizing "scope of impact / feasibility" and the latter "temporal dimension / current progress"; the two are complementary rather than independent conclusions. The specific time spans given in the tables (such as 1–2 years and 3–5 years) are the author's subjective judgment and ordinal ranking (near/medium/far term) based on the current state of the industry, not the output of any forecasting model, and readers should understand them as relative priorities rather than precise time predictions. A priority sequence can be read from them: technology-driven improvements (such as oracle upgrades and shortened settlement cycles) are already in the implementation stage and are expected to produce observable efficiency gains in the relatively near term; institution-coordination improvements (such as cross-exchange margin mutual recognition) must overcome the coordination game within the industry and are in a medium-term exploratory stage; and reforms involving the unification of global regulatory frameworks are, because of their inherent political complexity, unlikely to achieve a fundamental breakthrough in the foreseeable future.

On the other hand, some anomalies are irreparable. Not all anomalies can be repaired. Some are fundamental costs the market must pay in pursuing particular objectives (such as decentralization and censorship resistance). They stem from an irreconcilable "trade-off triangle," not a temporary technical defect.

A typical example is the CEX–DEX structural arbitrage residual. As described in Sections 18.7.3 and 18.9.2, as long as the blockchain's physical confirmation time is nonzero, a decentralized system can never surpass a centralized one in speed, and this ineliminable tiny latency constitutes the efficiency lower bound of decentralized finance. Its normative implication is that pursuing extreme decentralization and trustlessness requires accepting this efficiency "discount," and trying to eliminate it entirely is tantamount to abandoning decentralization itself. This is a philosophical choice about "what one wants" and "what one is willing to give up," not an engineering problem.

Second, the theoretical lower bound of information asymmetry dictates that arbitrage opportunities never fully disappear in theory. The paradox of informational efficiency proposed by Grossman and Stiglitz in 1980 [19] rests on a key premise: acquiring and processing information is costly. Under this premise, if market prices already fully reflect all information, then no trader can earn excess returns by collecting and analyzing information, and so no one has an incentive to engage in this costly activity. Yet once no one collects information, prices no longer reflect information and the market becomes inefficient again. The equilibrium outcome of this paradox is that the market must retain an "inefficiency" gap large enough that information collectors' expected returns just cover their information costs. Mapping this logic onto the crypto perpetual futures market, we note an important difference: the information-cost structure of the crypto market differs fundamentally from that of the traditional equity market. The openness and transparency of on-chain data lower the acquisition cost of some information, but the market's fragmentation (dozens of exchanges, multiple blockchains) and 24/7 uninterrupted trading greatly raise the cost of monitoring and processing information. The net effect is that the equilibrium point of informational efficiency in the crypto market may be higher than in traditional markets—that is, the "minimum arbitrage return" required to maintain market efficiency is higher. The Grossman-Stiglitz

paradox therefore theoretically constrains the possibility of fully eliminating anomalies: complete efficiency is a logically unattainable ideal, and the crypto market's distinctive information-cost structure further raises this ineliminable efficiency lower bound.

Finally, participant heterogeneity is an endogenous feature of the market ecosystem, hard and unnecessary to eliminate. The participant composition of different markets (retail investors, institutions, arbitrageurs, industrial capital) differs naturally, and their risk appetites, information channels, trading motives, and behavioral patterns vary widely. This heterogeneity is the source of the market's depth and breadth, but it also inevitably causes systematic differences in pricing logic across different markets (such as CEX vs. DEX, CME vs. offshore perpetual). For example, the long-term positive bias in funding rates is rooted in the persistent preference difference between retail investors and institutions. Attempting to "unify" the behavior of all participants in some way is not only unrealistic but would also harm the market's diversity and vitality.

Reflecting from anomalies to institutional improvement ultimately leads us to recognize the boundaries of market design. It requires us to distinguish "the doable" from "the undoable," to focus our energy on the areas where technical and institutional innovation can genuinely improve market resilience and efficiency, while also understanding and respecting the ineliminable efficiency losses determined by the market's fundamental trade-offs.

18.10 Chapter summary

This chapter has systematically analyzed the six categories of persistent market anomalies that arise in the perpetual futures market from constrained arbitrage activity: the basis anomaly, the funding-rate anomaly, the cross-exchange anomaly, the cross-asset anomaly, the term-structure anomaly, and the CEX–DEX anomaly. The analysis is not intended to declare the market "inefficient" but to reveal a more accurate understanding: market efficiency is always in a contest with various inherent, dynamic constraints, and anomalies are the most direct reflection of that contest.

The core conclusions of this chapter can be distilled into five interconnected layers. First, at least six categories of persistent anomalies exist in the perpetual futures market that can be systematically quantified and attributed; they are not isolated, random market noise but a stable reflection of market microstructure and macro dynamics. Second, the "constraint–anomaly mapping" method proposed in this chapter is the cornerstone of the entire analysis: **most** market anomalies can, to a large extent, be **traced to** the seven-dimensional constraint framework laid out in Chapter 17, and understanding the roots of an anomaly often requires returning to an analysis of the constraint itself; at the same time, it must be acknowledged that for **some** anomalies (such as the long-term positive bias in funding rates), the constraint merely **amplifies** an equilibrium deviation that already exists rather than creating the anomaly out of nothing—the method is therefore a mechanistic attribution, not a rigorous causal identification for each case. Third, the "anomaly persistence spectrum" forms a **non-exclusive continuum** along the three characteristic forms of transient, cyclical, and structural (rather than three mutually exclusive bins), revealing that different anoma-

lies have markedly different temporal characteristics and repair mechanisms, and that the same broad category of anomaly may fall at different positions on the spectrum depending on state or direction. The transient form reflects the market's instability under extreme stress, the cyclical form reveals the predictable friction endogenous to institutional design, and the structural form reflects the fundamental trade-offs and costs in the market's architecture. Fourth, the crisis-dynamics analysis uses the single case of LUNA/UST in May 2022 to illustrate the core prediction of Chapter 17's risk-resonance model: during extreme market turmoil, multiple categories of anomalies worsen almost in synchrony and infect one another, forming a positive feedback loop that causes the nonlinear amplification of price deviations—this mechanism is consistent with the single-case observation, but its generality awaits testing through multi-case comparison. Fifth, the landscape of anomalies is itself in a process of continual dynamic evolution: some anomalies are gradually fading with technological progress and capital deepening, others persist over the long run as durable features of structural constraints, and new anomalies keep emerging as market structure evolves.

Specifically, different anomalies correspond to different constraint forms and temporal characteristics. The basis anomaly presents a cyclical positive bias in bull markets and manifests as transient negative-basis stickiness in crises, the two pointing respectively to different combinations of margin and liquidity constraints. The long-term positive bias in funding rates stems from a structural asymmetry in participant structure, while its 8-hour settlement effect is rooted in a specific institutional-design parameter. From the residual spread among leading exchanges that approaches the theoretical lower bound, to the segmented pricing zones formed by regulatory barriers, to the permanent efficiency gap between CEXs and DEXs arising from the physical confirmation-time limit, each category of anomaly corresponds to a specific constraint form and its unique manifestation along the time dimension.

As the final chapter of Part 6, "Arbitrage: Boundaries and Limits," this chapter, together with the two preceding chapters, completes the tracing of the arbitrage mechanism from theory to reality across its entire course. Chapter 16 displayed the precision of the mechanism, including six categories of strategy, a layered ecosystem, and three-layer efficiency transmission; Chapter 17 displayed the instability of the mechanism, covering the seven-dimensional constraints, risk resonance, and nonlinear collapse; this chapter has revealed the domains the arbitrage mechanism cannot cover: persistent price deviations, pricing deviations sharply amplified in crises, and the structural efficiency loss that is a permanent cost of the market's architecture. The evolution of the anomaly landscape shows that the market as a whole is developing toward higher efficiency: from the roughly 3.5% cross-exchange spread recorded by Makarov-Schoar on 2017–2018 data to the single-digit-basis-point residual of 2025, from manual cross-exchange arbitrage to protocolized arbitrage, and from isolated CEXs to a CEX–DEX efficiency-transmission channel, arbitrageurs continually mend the weak links in the efficiency-transmission network. But some links, because of the fundamental limits of the laws of physics or the institutional framework, may never be eliminated entirely. Identifying and distinguishing these different layers of constraint and anomaly is precisely the core analytical framework that the three chapters of this part have sought to provide.

The analysis of this chapter will be extended in later parts. The phenomenon of a sharp decline in liquidity caused by arbitrageurs' exit will receive a more systematic theoretical analysis in Part 7's treatment of "liquidity"; the volatility clustering and cross-market contagion during the eruption of anomalies will be placed in a broader theoretical context in Part 8's analysis of "volatility"; and this chapter's tracing of the evolution of anomalies also provides the necessary empirical material for Part 9's analysis of market quality.

References

[1] Seo, M. H., Koo, B., & Yang, Y. F. (2024). Nonlinear dynamics of Kimchi premium. *Economic Modelling*, *135*, 106726. https://doi.org/10.1016/j.econmod.2024.106726

[2] Zhivkov, P. (2026). The two-tiered structure of cryptocurrency funding rate markets. *Mathematics*, *14*(2), 346. https://doi.org/10.3390/math14020346

[3] Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35–55. https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

[4] Liu, J., Makarov, I., & Schoar, A. (2023). Anatomy of a run: The Terra Luna crash (NBER Working Paper No. 31160). National Bureau of Economic Research. https://doi.org/10.3386/w31160

[5] Christin, N., Routledge, B. R., Soska, K., & Zetlin-Jones, A. (2022). *The crypto carry trade* (Working Paper). Carnegie Mellon University. https://www.andrew.cmu.edu/user/azj/files/CarryTrade.v1.0.pdf

[6] He, S., Manela, A., Ross, O., & von Wachter, V. (2024). Fundamentals of perpetual futures (Working Paper, Revise & Resubmit at *Review of Financial Studies*). Washington University in St. Louis. https://doi.org/10.48550/arXiv.2212.06888

[7] Kaiko. (2026). *Kaiko exchange ranking*. Retrieved March 6, 2026, from https://www.kaiko.com/indices/exchange-ranking

[8] Agio Ratings. (2026, February 18). *Crypto credit ratings: How institutions measure counterparty and asset risk*. Agio Ratings. https://www.agioratings.io/insights/crypto-credit-ratings-how-institutions-measure-counterparty-and-asset-risk

[9] McGinley, I. (2023, September 11). *Enforcement by enforcement: The CFTC's actions in the derivatives markets for digital assets* [Keynote speech]. U.S. Commodity Futures Trading Commission. https://www.cftc.gov/PressRoom/SpeechesTestimony/opamcginley1

[10] Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, *20*(3), 339–350. https://doi.org/10.1198/073500102288618487

[11] Barbon, A., & Ranaldo, A. (2021). On the quality of cryptocurrency markets: Centralized versus decentralized exchanges. *arXiv*. https://doi.org/10.48550/arXiv.2112.07386

[12] Nadler, M., Schuler, K., & Schär, F. (2025). Blockchain price oracles: Accuracy and violation recovery. *Journal of Corporate Finance*, *95*, 102908. https://doi.org/10

.1016/j.jcorpfin.2025.102908

[13] Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., & Juels, A. (2020). Flash Boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. In *2020 IEEE Symposium on Security and Privacy (SP)* (pp. 910–927). IEEE. https://doi.org/10.1109/SP40000.2020.00040

[14] Qin, K., Zhou, L., & Gervais, A. (2022). Quantifying blockchain extractable value: How dark is the forest? In *2022 IEEE Symposium on Security and Privacy (SP)* (pp. 198–214). IEEE. https://doi.org/10.1109/SP46214.2022.9833734

[15] Briola, A., Vidal-Tomás, D., Wang, Y., & Aste, T. (2023). Anatomy of a stablecoin's failure: The Terra-Luna case. *Finance Research Letters*, *51*, 103358. https://doi.org/10.1016/j.frl.2022.103358

[16] Forbes, K. J., & Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance*, *57*(5), 2223–2261. https://doi.org/10.1111/0022-1082.00494

[17] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, *135*(2), 293–319. https://doi.org/10.1016/j.jfineco.2019.07.001

[18] Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, *72*(3), 967–998. https://doi.org/10.1111/jofi.12498

[19] Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, *70*(3), 393–408. https://www.jstor.org/stable/1805228

Part VII
Liquidity

Core question: Where does liquidity come from, why does it vanish, and how should it be governed?

Liquidity Provision and Strategic Games

Consider the operational practice of a global cryptocurrency market maker. (The opening scenario of this chapter is constructed from public market data and industry interviews; it is intended to illustrate the multidimensional pressure structure that perpetual-futures market makers face, and it does not refer to any specific institution. The concrete figures in the text—$50 million of net long inventory and a 150% annualized funding rate—are illustrative settings rather than measured statistics.) Within a single trading window, its trading system faces three types of structural pressure at once. On the centralized exchange (CEX) perpetual-futures order book, a sudden drop in ask-side depth triggers a toxic-order-flow alert, and distinguishing liquidation flow from informed flow becomes the key to staying profitable. At the funding-rate management level, the annualized BTC funding rate spikes to 150% (referencing the extreme rates during BTC's break to an all-time high in March 2024; here the annualization convention follows "annualized = per-settlement rate × settlements per day × 365," so 150% annualized corresponds to about 0.137% per settlement); combined with roughly $50 million of net long inventory, this generates a substantial holding cost in every settlement period, and unlike the overnight interest of traditional markets, this cost variable is highly random and directionally uncertain. In on-chain markets, the net asset value of the liquidity pool declines steadily in the absence of any obvious triggering event, revealing an additional risk dimension that market makers on decentralized exchanges (DEXs) face under maximal extractable value (MEV) attacks. This scenario reveals the core complexity of the perpetual-futures market-making business: its challenge arises not from a single dimension but from the simultaneous superposition of adverse-selection risk, inventory-cost risk, and institutional risk, and it takes fundamentally different forms across the two market architectures of CEX and DEX.

This is the typical operating environment confronting perpetual-futures market makers in 2026. Cross-platform arbitrage and the fragmented trading structure of crypto markets [1], together with the pricing mechanism of perpetual futures [2], jointly shape this environment: market makers must operate continuously for 72 hours across two core market architectures—the CEX central limit order book (CLOB) and the on-chain order book of DEXs. They must contend with "toxic" order flow from four distinct sources—informed trading, arbitrage extraction, liquidation cascades, and MEV attacks—while managing a risk inventory subject to periodic funding-rate shocks, all within a market that runs 24/7. Compared with market making in traditional financial

markets, this difference in difficulty is not one of degree but one of dimension.

In Part Six of this book, we have repeatedly treated "liquidity" as a key background variable, using it to explain the boundaries of arbitrage opportunities and the causes of market anomalies—whether the "liquidity black hole" examined in Chapter 17 or the altcoin "efficiency wasteland" analyzed in Chapter 18. In those chapters, however, liquidity was a macro-level, given presence. This chapter shifts the analysis from the demand side of liquidity to the supply side and investigates a core question: where does liquidity come from? By whom, and at what cost and risk, is it "produced"? Why are market makers willing to bear the risks of adverse selection and inventory fluctuations on an ongoing basis? And what strategic framework have they developed within this game?

These questions are foundational. Understanding how liquidity is produced is the prerequisite for understanding why it fails, how it might be repaired, and why it vanishes precisely when it is most needed—and thus for designing more resilient liquidity-incentive mechanisms. Market making in perpetual futures rests on a multidimensional interplay of risk pricing, information games, and institutional constraints, and its cost structure differs fundamentally from that of traditional markets.

19.1 The economics of liquidity provision

In any active financial market, traders expect to be able to buy or sell an asset at a "reasonable" price at any time. Behind this expectation lies a market quality called "liquidity," which directly determines whether traders can complete their transactions at a reasonable cost. When liquidity is abundant, the market runs smoothly, and traders can easily execute their intentions without exerting excessive impact on prices; when liquidity dries up, the market becomes fragile and sluggish, and even small trades can trigger sharp price swings. Yet liquidity does not arise from nothing. It is not a natural state but a product that is carefully "manufactured." The professional participants who manufacture this product are market makers.

The market maker's role is often misunderstood. In the narratives of some market participants, they are mysterious "manipulators" with an information advantage who profit from the losses of ordinary traders. In other, overly simplified academic models, they are portrayed as "automata" that passively respond to order flow and mechanically adjust quotes. Both portrayals depart from reality. The essence of market making is neither a predation based on information asymmetry nor a passive public service. It is a rigorous, high-risk business whose core lies in the trade-off between risk and return. Market makers are willing to act as counterparty on an ongoing basis because they expect to earn a statistically positive spread return by bearing specific, manageable risks. Their presence allows those liquidity demanders who are unwilling to wait and want to trade immediately to externalize their "cost of waiting" and their "cost of finding a counterparty."

This section analyzes the economics of the market-making business and answers a central question: why is anyone willing to bear risk in order to supply liquidity to the market continuously? The analysis begins with the revenue and cost structure of the

business and then examines two classic risks: the adverse-selection risk that stems from information asymmetry and the inventory risk that stems from price fluctuations. These two risks form the general theoretical pillars for understanding all market making.

Yet the perpetual future, as a distinctive financial instrument—with its lack of an expiration date, high leverage, funding rate, and forced-liquidation mechanisms—adds a new dimension to the market maker's traditional risk profile. One of this chapter's integrative frameworks addresses exactly this: it consolidates and names the "third pillar" of perpetual-futures market making, institutional cost. We analyze in detail how the randomness of the funding rate, the punitive nature of the liquidation mechanism, the lag of the mark price, and the operational pressure of 24/7 continuous trading together constitute a set of structural costs that are endogenous to the product's design and difficult to hedge away completely. Finally, we integrate these three pillars into a unified profit equation, thereby revealing the market maker's "survival condition." This equation not only explains how market makers profit but, more importantly, foreshadows the conditions under which market makers will choose to contract or even abandon market making, giving rise to the "liquidity black hole" phenomenon we will examine in later chapters. Understanding this profit equation is the key to understanding liquidity-provision behavior and, indeed, the stability of the entire market microstructure.

19.1.1 The nature of market making

The market maker's core business model can be distilled as follows: by continuously posting competitive bid and ask prices to the market, it profits from the spread between the two. Its function resembles that of an intermediary in financial markets—sourcing assets from deeper liquidity sources and offering them, at a price that includes a spread, to traders who need immediate execution. This seemingly simple model in fact performs several essential market functions.

First, the market maker supplies immediacy. In a purely order-driven market without market makers, a trader wishing to buy immediately must wait for a counterparty willing to sell at an acceptable price to appear. This wait can be long, and its outcome is uncertain. By committing to be ready to trade at any time, the market maker greatly shortens this matching time, connecting an asynchronous market into a seemingly synchronous whole. It links buyers and sellers who arrive at different moments along the dimension of time.

Second, the market maker reduces transaction costs. Competition among market makers is expressed primarily in who can offer a narrower bid-ask spread and greater quoting depth. A market with a narrower bid-ask spread means lower costs for liquidity demanders seeking immediate execution. A market with greater depth means that large trades can be executed with smaller price impact. The existence of and competition among market makers therefore directly determine the trading efficiency and capacity of the market.

Third, the market maker participates in and contributes to price discovery. The prices it posts are not set arbitrarily but reflect an integrated judgment of current market information, order-flow pressure, its own inventory level, and expectations of future volatility. The dense queue of quotes on an order book itself embodies the market's

collective assessment of an asset's fair value. When new information enters the market, market makers quickly adjust their quotes, thereby guiding the price toward a new equilibrium. Their quoting behavior becomes an important link in the price-discovery process.

Understanding the market-making business hinges on distinguishing the source of its profit. Unlike position traders who depend on predicting an asset's long-term price direction, a pure market maker aims to profit from the turnover of two-way trading rather than from the price movement of a one-sided position. Its basic profit formula can be expressed as:

$$\text{Profit} \approx (\text{average bid-ask spread} \times \text{total volume}) - \text{transaction costs}$$

Here, "transaction costs" are not simply fees but a complex combination that mainly comprises two core risks: the loss from adverse selection by better-informed traders, and the risk of exposure to price fluctuations from holding inventory. A market maker's strategy, algorithms, and technology investments all revolve around how to price these two risks precisely and internalize their cost into the bid-ask spread. The core competitive advantage of a successful market maker lies not in predicting market moves more accurately but in measuring and managing risk more precisely than its competitors, so that it can offer better quotes at a given level of risk and achieve a statistically positive expected return across a large number of trades. From an economic standpoint, a market maker's profit is essentially the risk premium the market pays for two scarce services: immediacy and price certainty.

19.1.2 Adverse selection and inventory management

The market maker's profit model is, fundamentally, a search for certainty in a game against uncertainty. After decades of research, the academic literature has traced the core uncertainty faced by market makers to two sources, which constitute the twin pillars of modern market-making theory: adverse selection and inventory management. Understanding these two pillars is the basis for understanding how market makers price, how they manage risk, and why they sometimes choose to exit the market.

The first pillar, adverse selection, arises from the structural information asymmetry in the market [3]: the order flow a market maker faces is a mixture of two types of traders—"noise traders," whose behavior is random and driven by allocation or hedging needs, and "informed traders," who possess undisclosed information about an asset's future price. In nearly every trade with an informed trader, the market maker stands on the wrong side and bears a nearly certain loss—the adverse-selection cost. Its response is to spread this expected loss across every trade through a wider bid-ask spread, so that the width of the spread reflects the market maker's judgment about how active informed trading is in the market. A full characterization of this risk—the classic models of Glosten and Milgrom (1985) [4] and Kyle (1985) [5]—is developed in Sections 19.2 and 19.2.1.

The second pillar, inventory management, arises from the fact that a market maker inevitably accumulates net inventory through passive fills: buy and sell orders are hard to match perfectly over the short term, so inventory departs from neutral and forms

a net long or net short position, turning the market maker partly from a pure spread trader into a position holder exposed to price fluctuations. The core response tool is "quote skewing"—dynamically adjusting quotes on the bid and ask sides according to inventory direction so as to steer order flow and accelerate the return of inventory to neutral. In essence, this is a trade-off between "closing the position quickly to reduce risk" and "waiting for a better price to profit"; the longer the holding period, the greater the risk, so market makers generally favor high turnover. The classic modeling of this mechanism—the inventory models of Ho and Stoll (1981) [6] and Amihud and Mendelson (1980) [7]—is developed in Sections 19.3 and 19.3.1.

Together, the two risks of adverse selection and inventory management constitute the core of the market maker's cost structure. The market maker's quoting decision is, in essence, the solution to an optimal spread and depth subject to the constraints of these two risks. The spread must be wide enough to compensate for the expected adverse-selection loss and the risk premium of holding inventory, yet narrow enough to attract sufficient order flow in competition. This is a dynamic optimization process that continually seeks a moving balance between risk and return, safety and growth.

19.1.3 Institutional cost

The classic twin-pillar theory provides a solid foundation for understanding market-making behavior in all markets. Yet when we turn to the perpetual future—a financial instrument native to the crypto world—we find that adverse selection and inventory risk alone are no longer sufficient to explain market-maker behavior. The distinctive institutional design of perpetual futures—including the funding rate, leverage and liquidation, the mark-price mechanism, the absence of an expiration date, and 24/7 uninterrupted trading—together gives rise to a third systematic source of cost, which we term "institutional cost." This consolidated naming reveals why market making in perpetual futures differs in kind from market making in traditional equities or futures.

Institutional cost arises not from information asymmetry or price fluctuations but is embedded in the "rules of the game" of the perpetual future as a product. As long as one participates in this market, one must bear these costs. They comprise the following main components.

The first is funding-rate exposure. The funding rate is the core mechanism by which perpetual futures anchor to the spot price [8]; it can be positive or negative and fluctuates sharply. Two related but distinct concepts must be distinguished: the spot-futures basis between the perpetual future and the spot (referred to in the academic literature as "crypto carry"), and the funding rate itself. Schmeling et al. (2023) show that crypto carry sometimes exceeds 40% annualized, far above the level of traditional markets, largely owing to the structural friction between leveraged retail demand and scarce arbitrage capital [9]. The basis can be captured by arbitrageurs through a cash-and-carry strategy, whereas the funding-rate exposure a market maker faces is of a different nature: it is the funding payment or receipt in which the market maker is forced to participate every settlement period after accumulating inventory through passive fills. The funding rate fluctuates far more than traditional funding costs: during BTC's break to an all-time high in March 2024, a single 8-hour settlement rate once exceeded

0.1% (using the convention "annualized = per-settlement rate × settlements per day × 365," a single rate of 0.1% per 8 hours corresponds to about 109.5% annualized, while an annualized figure of over 400% would correspond to about 0.365% per settlement), which means that a market maker holding $10 million of net long inventory would pay more than $10,000 in a single settlement period. More critically, the direction and magnitude of the funding rate are highly random and can flip from an extreme positive value to a negative one within hours. This stands in sharp contrast to the relatively stable, predictable funding costs (such as the overnight rate) in traditional futures market making. A market maker must manage not only the risk of price fluctuations but also the risk of funding-rate fluctuations—a risk dimension unique to perpetual futures.

The second is the liquidation risk premium. Because of the high leverage of perpetual futures, the forced-liquidation mechanism is triggered frequently during sharp market swings. A market maker's own inventory also faces the risk of forced liquidation. Forced liquidation differs fundamentally from a market maker's active inventory management: it occurs at the moment when market liquidity is worst and prices are most unfavorable, a punitive, passive purging of risk. To avoid this, a market maker must maintain a higher margin level than theory would require, or be forced to close positions actively at very poor prices as it approaches the liquidation threshold. The additional cost incurred, or potential return forgone, in order to avoid forced liquidation constitutes the liquidation risk premium.

The third is mark-price deviation risk. Margin and profit-and-loss (P&L) calculations for perpetual futures are based on a smoothed mark price (typically a moving average of a multi-exchange index price plus the basis). The primary design purpose of this smoothing mechanism is to prevent erroneous liquidations caused by short-term price manipulation; in this sense it is protective. However, in a sustained one-directional price move, the mark price systematically lags the true, latest market price, and this lag produces two risks of opposite direction but equal harm. First, liquidation-delay risk: when the price moves continuously in one direction, the lag in the mark price means that positions that should have triggered liquidation earlier are "tolerated" for longer, so that the eventual liquidation loss is far greater than it would have been under timely liquidation; this additional loss must be borne by the insurance fund or through a socialized-loss mechanism, indirectly affecting all market participants, including market makers. Second, hedge-basis risk: when a market maker hedges its perpetual-futures position, the price of the hedging instrument (such as spot) moves in real time with the latest traded price, whereas the P&L calculation of the perpetual future is anchored to the lagging mark price. In fast markets, the deviation between the two can widen significantly, so that a market maker's hedge appears balanced in nominal terms while the actual P&L shows unexpected deviations caused by the mark-price lag.

In addition, the indefinite exposure created by the lack of an expiration date increases the difficulty of inventory management: perpetual futures lack the time anchor of traditional futures' "natural settlement at expiration," so an imbalanced inventory can in theory be held indefinitely, remaining continuously exposed to price and funding-rate risk, forcing market makers to rely more heavily on costlier active hedging (the full mechanism of its missing "center of gravity" is discussed in Section 19.3.2).

Finally, 24/7 uninterrupted trading brings enormous operational costs and risk-monitoring pressure. Traditional markets have closing and holiday hours, which give market makers necessary downtime for system maintenance, model calibration, risk review, and strategy adjustment. In the perpetual-futures market, risk is ever-present. Market makers must invest heavily in building trading systems, risk-control systems, and monitoring teams capable of running stably around the clock, so as to respond to any market disruption or technical failure that may occur at any moment. This continuous operational pressure and labor cost constitute another important component of institutional cost.

Integrating the five subcomponents above, we arrive at a more complete cost structure for a perpetual-futures market maker. Institutional cost is a composite category, encompassing all costs induced by the institutional design of the perpetual future itself—from funding-rate exposure, the liquidation risk premium, and mark-price deviation risk to the indefinite exposure of no expiration and the pressure of 24/7 uninterrupted operation. Among these, the system-building, staffing, and continuous-monitoring expenditures brought by 24/7 operation, though "operational" in form, are rooted in the institutional design of perpetual futures' uninterrupted trading rather than in the expansion of a market maker's own business scale; they are therefore classified under institutional cost rather than as a separate operational-cost category in this chapter's analytical framework. Compared with spot or delivery-futures market making, a perpetual-futures market maker must not only face more severe adverse-selection and inventory risks (amplified by leverage and high volatility) but also bear this set of institutional costs, brought by the product's institutional design itself, that cannot be fully eliminated through traditional risk-management tools. This explains why liquidity-provision behavior in perpetual-futures markets exhibits characteristics markedly different from those of traditional markets.

To give the concept of institutional cost a structured analytical expression, we can decompose it into the following four subcomponents:

$$C_{\text{inst}} = \mathbb{E}[|FR_t| \cdot |Q_t|] + P(\text{liq}) \cdot L_{\text{penalty}} + f(\Delta_{\text{mark}}) + C_{\text{ops}}$$

Here, the first term is the expected cost of funding-rate exposure, where FR_t is the funding rate at time t and Q_t is the size of the market maker's net inventory; the interaction of the two determines the funding payment or receipt in each settlement period, and taking the expectation of the absolute values reflects the risk exposure under directional uncertainty of the rate. The second term is the liquidation risk premium, where $P(\text{liq})$ is the probability of triggering forced liquidation at a given inventory and leverage level, and L_{penalty} is the additional loss from forced liquidation relative to active position closing. The third term is a function of mark-price deviation risk, where Δ_{mark} is the magnitude of the deviation between the mark price and the latest traded price; this function grows nonlinearly when prices move rapidly in one direction. The fourth term, C_{ops}, is the fixed and semi-fixed cost of 24/7 uninterrupted operation. This expression is a conceptual classification framework, not a directly calibratable quantitative model: the parameters of each subterm (such as the computation of $P(\text{liq})$) depend heavily on a specific exchange's liquidation-engine rules, margin system, and microstructural state, and in practice are usually estimated by Monte Carlo simulation

rather than by an analytical expression. Its value lies not in providing precise numbers but in revealing the multiple sources of institutional cost and the structural relationships among its subcomponents; in particular, the randomness and state-dependence of the first three terms mean that institutional cost can expand sharply under extreme market conditions.

19.1.4 The profit equation and survival conditions

Combining the above analysis, we can construct a more comprehensive profit equation for the market maker that systematically characterizes its sources of revenue and its cost components. This equation is not merely an accounting identity but an analytical framework that guides a market maker's day-to-day decisions and strategic choices.

$$\text{Total profit} = \text{spread revenue} - (\text{adverse-selection cost} + \text{inventory-risk cost} + \text{institutional cost})$$

Here, "spread revenue" is a broad concept in this chapter's framework: it includes not only the profit from the market-making spread itself but also the fee-rebate revenue that exchanges pay to market makers (detailed in Section 19.5.4) and the funding-rate revenue that market makers *collect* by strategically managing their inventory direction (Section 19.3.3 will analyze the dual nature of the funding rate as both cost and revenue). The accounting attribution of the funding rate must be clearly distinguished: the funding revenue a market maker strategically *collects* at the margin through quote skewing is booked under the broad "spread revenue," whereas the funding payment it is forced to *bear* after accumulating inventory through passive fills is booked under "institutional cost." The two are in fact manifestations of the same variable in opposite directions (single-period funding P&L $= -FR_t \cdot Q_t$, with a positive value indicating net revenue to the market maker): when the inventory direction is opposite to the funding direction, it is positive revenue; when it is in the same direction, it is a cost. On mainstream pairs, rebate revenue can account for a significant proportion of a market maker's total revenue. The three costs in parentheses—corresponding to information risk, position risk, and product-institutional risk, respectively—constitute the cost threshold that a market maker must clear. This equation reveals the market maker's "survival condition": the revenue it captures through spreads must, in a statistical sense, be sufficient to cover all the risk costs it faces. When this inequality cannot be satisfied, a rational market maker has only three options: widen the bid-ask spread to increase revenue per trade, reduce quoting depth to lower risk exposure, or, in the most extreme case, withdraw from the market entirely and stop quoting.

The shifting balance of this profit equation explains the dynamic changes in market liquidity. In an environment that is calm, information-transparent, and highly competitive, a market maker's various costs are low, and it has both the ability and the willingness to offer narrow spreads and deep liquidity, so that the market appears efficient and stable. However, when the market enters a period of sharp volatility, when major news breaks frequently, or when the institutional mechanisms of perpetual futures (such as the funding rate or liquidation pressure) become extreme, a market maker's costs surge. Adverse-selection risk rises as information asymmetry intensifies, inventory risk increases as price volatility is amplified, and institutional cost becomes salient

as funding rates turn extreme and liquidation risk looms. At this point, to maintain its survival condition, market makers are forced to collectively and defensively widen spreads and withdraw depth. This is precisely the microeconomic root of the "sharp liquidity contraction" phenomenon we will examine in depth in later chapters.

Recent empirical research further corroborates the theoretical analysis above. Using the natural experiment of Binance's July 2022 removal of maker-taker fees on BTC pairs, Galati (2024) finds that zero fees actually led market makers to widen bid-ask spreads and reduce quoting depth, raising total market transaction costs [10]. The mechanism is that, once the maker rebate was removed, market makers lost the explicit incentive to maintain narrow spreads and instead compensated with wider spreads, so that the implicit spread cost exceeded the explicit fees that had been eliminated. Easley et al. (2024), using machine-learning methods to analyze microstructure data for several mainstream crypto assets on a major crypto exchange, demonstrate that microstructure indicators such as liquidity and price discovery have significant predictive power for crypto-asset price dynamics [11]. These studies show that the impact of market-maker behavior on market quality is far more complex than a simple fee incentive would suggest.

A market maker's survival condition varies markedly across asset types. For highly liquid, high-market-cap assets such as Bitcoin or Ethereum, market depth is sufficient to absorb large volumes of trading, and information is relatively dispersed and transparent. Although informed trading still exists, its share of total order flow is low. A market maker can therefore be profitable at a relatively narrow spread through a model that combines lower profit margins with high volume. By contrast, for the large universe of "altcoin" perpetual futures, markets are typically thin in liquidity, highly concentrated in information (with project teams or whales holding an absolute information advantage), and more volatile in price. This means market makers face far higher adverse-selection and inventory-risk costs in these markets. To survive, they must set very wide bid-ask spreads. This in turn suppresses trading activity, forming a negative feedback loop of insufficient liquidity. This explains the microeconomic mechanism behind the "altcoin efficiency wasteland" phenomenon observed in Chapter 18.

Using a real-time order-book snapshot of ten major perpetual-futures pairs on Binance, the following figure illustrates the negative correlation between market depth and the bid-ask spread: BTC has the greatest depth and the lowest spread, ETH follows, and for small- and mid-cap assets such as ADA and DOT, depth drops sharply while spreads surge (see the figure for specific values). Over this sample period, liquidity is highly uneven across assets, consistent with the direction of the theoretical analysis above.

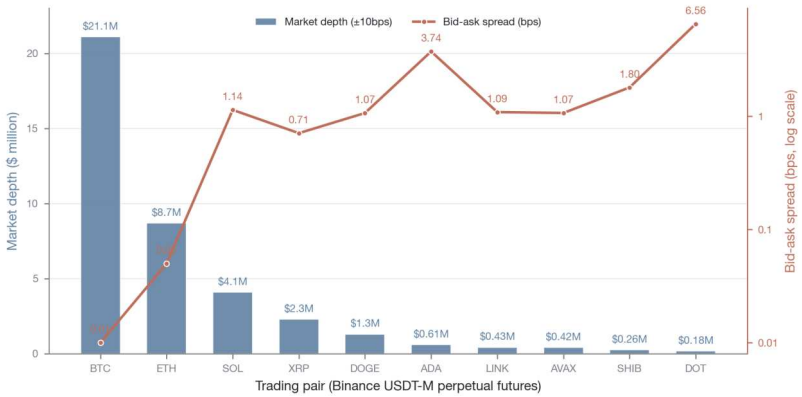


Figure 19-1. Comparison of market depth and bid-ask spread across perpetual-futures pairs (Data source: real-time order-book data for Binance USDT-M perpetual futures, March 2025)

Ultimately, we can understand a market maker's profit as the fee the market pays to obtain liquidity as a "public good." Liquidity is not free precisely because it is "produced" by professional risk-bearers who commit capital, technology, and labor and who bear the multiple risks of adverse selection, inventory, and institutional (including operational) cost. A market maker's profit is the reasonable compensation for this production process. Only by understanding this can we appreciate why liquidity is valuable and why, under specific conditions, it becomes fragile and scarce.

19.2 Adverse selection

In the market maker's profit equation, inventory risk is a persistent risk exposure, whereas adverse selection is an episodic risk factor with pronounced impact. It occupies a central place in the market maker's cost structure. The essence of adverse selection stems from the ever-present information asymmetry in markets. In an ideal world where all participants held exactly the same information, a market maker's risk would be confined solely to the price fluctuations of its inventory. In the real world, however, there is always a subset of traders (whom we call "informed traders") who, sooner or later and to a greater or lesser degree, possess nonpublic information about the future direction of an asset's price. When these informed traders enter the market, they do so not to satisfy a liquidity need but to profit from their information advantage. And the market maker, as the liquidity supplier that always stands ready to quote, inevitably becomes the primary counterparty through which informed traders monetize the value of their information.

This game is asymmetric from the outset: the market maker cannot tell before a trade whether its counterparty is an informed trader or a noise trader, yet it always ends up on the wrong side in trades with the former—the price it pays for information asymmetry is the adverse-selection cost. How great the adverse-selection risk of a market is therefore directly determines the "floor" of its liquidity cost; the bid-ask spread is, in a sense, the insurance premium the market pays for the expensive commodity of "information."

In the distinctive ecosystem of perpetual futures, this classic problem takes on new

dimensions and greater intensity. The classic models used to analyze adverse selection in traditional markets, such as the Glosten-Milgrom model, retain their core insight, but the model's parameters and dynamics are greatly amplified by the institutional features of perpetual futures (especially leverage, 24/7 uninterrupted trading, and the distinctive on-chain information environment). Market makers must contend not only with informed traders from traditional channels but also with new sources of information advantage arising from the on-chain world. Their risk-monitoring systems must be more sensitive—able not only to detect the occurrence of a risk event but also to distinguish its type. This section analyzes in depth the particularities of adverse selection in perpetual-futures markets: it begins by reinterpreting the classical theory, identifies the sources of informed trading, analyzes its amplifiers, and finally examines the quantitative tools market makers use to protect themselves, revealing the full picture of this ongoing game.

19.2.1 The applicability of the Glosten-Milgrom model

One of the cornerstones of market-microstructure theory is the information-asymmetry model proposed by Glosten and Milgrom (1985) [4]. Although this model was born in the pre-digital-currency era, its profound insight remains the starting point for understanding market-maker behavior today. The core intuition of the Glosten-Milgrom model can be depicted as a simple probability game. When setting bid and ask quotes, a market maker faces two types of counterparties: noise traders, who trade for various reasons unrelated to the asset's fundamentals, such as liquidity needs, portfolio rebalancing, or irrational market sentiment; and informed traders, who possess private information about the asset's future value.

The market maker's dilemma is that it cannot tell a counterparty's true identity before a trade occurs. It can only estimate, based on historical experience, the probability that the next order comes from an informed trader (usually denoted by the Greek letter μ). When a buy order arrives, the market maker knows there is a probability μ that it comes from an informed trader who knows the asset is about to appreciate, and a probability $1 - \mu$ that it comes from a random noise trader. If the trade goes through and the counterparty happens to be informed, then the price of the asset the market maker sold will rise in the future, and it will suffer a loss. If, on the other hand, the counterparty is a noise trader, the future direction of the price is random, and the market maker earns the bid-ask spread securely. To avoid going bankrupt in a repeated game against informed traders, a rational market maker must set a bid-ask spread wide enough that the profit earned from noise traders can offset the expected loss suffered on informed traders. The size of the spread therefore directly reflects the market maker's judgment about the degree of information asymmetry in the market: the higher μ, the wider the spread.

Applying this classic framework to perpetual-futures markets, we find that its basic logic still holds, but the model's key variables are given new, more violent dynamics by the institutional design of perpetual futures. First, the leverage mechanism becomes a powerful amplifier of adverse selection. In the original Glosten-Milgrom model, the impact an informed trader can inflict using an information advantage is limited by the

size of its own capital. But in the perpetual-futures market, an informed trader with $1 million of capital can easily open a position at 10x or even 100x leverage, instantly scaling its notional principal to $10 million or even $100 million. This means the same piece of information can exert an impact on the order book several times greater in the perpetual-futures market than in the spot market. The market maker no longer faces a counterparty of limited capital but a trader whose impact, magnified by leverage, is several times its original capital. This forces the market maker to treat the leverage level as a core variable in its assessment of adverse-selection cost, systematically setting wider spreads in high-leverage markets or on high-leverage assets.

Second, 24/7 uninterrupted trading fundamentally changes the pattern of information arrival and release. In traditional equity markets, information typically accumulates after the close and is released collectively at the next day's open. Market makers (or specialists) can make a preliminary judgment about the degree of information asymmetry during the opening call auction by observing order flow, and adjust the opening spread accordingly. In the perpetual-futures market, however, trading runs without interruption, and information can pour in at any moment from anywhere in the world. Whether U.S. regulatory news, project developments in Asia, or macroeconomic data from Europe, any of these can be instantly translated into trading instructions that hit the order book. Market makers lose the important assessment-and-adjustment window of traditional markets; they must remain vigilant at all times, and their risk-management systems must be automated and run around the clock. This continuous state of alertness is itself a high operational cost and risk premium.

Finally, some mechanisms unique to perpetual futures, such as funding-rate settlement, also create "micro-hotspots" of information trading. The funding rate settles every 8 hours (or more frequently), and around the settlement point, market participants trade based on their expectations of the future rate, which itself can attract arbitrageurs and informed traders. For example, if the market broadly expects the funding rate to flip from positive to negative, traders holding long positions have an incentive to close before settlement, while informed traders who expect the price to come under pressure as a result may preemptively build short positions. These games around funding-rate settlement make the periods before and after the settlement point a hotbed of information trading, and the adverse-selection risk a market maker faces during these periods spikes periodically. This mechanism makes information-risk concentration periods—analogous to the "open" and "close" of traditional markets—recur every 8 hours.

The reinterpretation of the Glosten-Milgrom model in perpetual futures is therefore no longer a static probability calculation but a dynamic, multivariate risk-assessment process. A market maker's quoting decision must reflect, in real time and in an integrated way, factors such as the leverage level, global information flow, and the funding-rate cycle, which makes the asymmetric game in perpetual-futures markets significantly more complex than in traditional markets.

19.2.2 Sources of informed trading in perpetual futures

For a perpetual-futures market maker, the core problem is to identify the order-flow attributes of every fill. Identifying the intent behind order flow—distinguishing "toxic" informed trading from "harmless" noise trading—is the key to its survival. Compared with the relatively mature and well-regulated traditional financial markets, the spectrum of "informed traders" in perpetual-futures markets is broader and more complex, and their information sources are more diverse. Classifying these informed traders helps us map the threats a market maker faces.

The first type, and the most traditional, is the macro- or event-driven informed trader. These traders possess nonpublic information about to have a major impact on the market. Such information may come from the regulatory level, such as inside information that the U.S. Securities and Exchange Commission (SEC) is about to approve or reject a crypto exchange-traded fund (ETF); it may also come from project fundamentals, such as a core developer of a major public chain discovering a serious vulnerability, or a DeFi protocol about to announce a major partnership. In the brief window before this information becomes public, informed traders use the high leverage of perpetual futures to build large positions quickly. Their trading is decisive and firm, typically taking the form of continuous, large market orders aimed at completing the position build as fast as possible rather than quibbling over small slippage. For a market maker, this type of order flow is the most toxic, because it stems directly from a certain future change in price. Once it trades with such a trader, the market maker is almost destined to be on the wrong side.

The second type is the cross-market information-relay trader. The "information" advantage of these participants stems not from inside information but from speed. The crypto market is a global, fragmented market, and the same asset may trade simultaneously on hundreds of spot exchanges, futures exchanges, and ETF markets. Because of information-transmission delays and microstructural differences among markets, price changes do not occur synchronously across all markets. For example, when the BTC spot price on Coinbase suddenly jumps because of a large institutional buy order, a high-frequency trading firm can, using its low-latency servers and optimized trading algorithms, capture this signal within milliseconds and immediately build a long position in the BTC perpetual-futures market on Binance. For market makers on Binance, this order flow is likewise "informed," because it foreshadows a follow-on rise in price. These traders act as "information relays" across markets, earning the profit from differences in the speed of price discovery among markets, while market makers pay the adverse-selection cost for that speed difference.

The third type, unique to the crypto world, is the on-chain-information-parsing trader. The transparency of the blockchain means that a large amount of valuable information is publicly recorded, but interpreting this information requires expertise and powerful data-analysis capabilities. These informed traders predict future price movements by monitoring on-chain activity. For example, they can track large transfers by "whale" addresses; when a long-dormant whale address suddenly transfers a large quantity of tokens to an exchange, this may foreshadow imminent selling pressure. They can also analyze the liquidation queues of DeFi protocols; when the number of under-collateralized

positions surges, this signals that a wave of liquidations may be imminent, leading to a sharp price drop. In addition, analyzing changes in a protocol's total value locked (TVL), the voting trends of governance tokens, and even the activity patterns of MEV bots can all provide clues for price prediction. Once effectively parsed, these on-chain signals constitute a powerful information advantage, allowing such traders to conduct adverse selection against market makers with weaker on-chain-data-analysis capabilities in the CEX perpetual-futures market.

The fourth type is the whale or large holder with the intent to manipulate the market. Unlike the first three types, who passively exploit information, these participants actively create information. Using their ample capital, they artificially manufacture sharp price swings—through continuous large market orders during relatively thin trading periods (such as the early hours of a weekend) or on small-cap assets such as altcoins. The purpose may be to dump spot holdings on retail traders after pushing the price up, or to trigger other traders' stop-losses or liquidations and profit amid the chaos. Although this behavior constitutes market manipulation in many jurisdictions, it remains common in the still-underregulated crypto market. For a market maker, this type of order flow is extremely hard to distinguish from genuinely fundamentals-driven informed trading, yet its risk is equally large. A market maker's quoting depth is easily punched through under such large impacts, leading to a significant accumulation of inventory risk.

The four sources of informed trading above differ markedly in market share, type of information advantage, and dependence on speed. Macro-event-driven informed traders are the fewest in number but have the greatest single-shot impact; their information advantage stems from nonpublic fundamental information, their demand for speed is relatively low, and the information window may last for hours. Cross-market information-relay traders are more numerous and highly dependent on speed; their profit depends entirely on a millisecond-level latency advantage, but the information content of each trade is low. On-chain-information-parsing traders are growing rapidly in number as on-chain analytics tools become widespread; their information advantage stems from the specialized interpretation of public data and lies between the first two. Manipulative large holders are the fewest but the hardest to predict, because they not only exploit information but actively create it. Together, these four sources constitute the multi-source threat network a market maker faces.

Understanding these diverse sources of informed trading directly shapes the design of a market maker's defensive strategy. Adverse selection is not a single problem but a composite of risks of different natures. A market maker's countermeasures must therefore be multilayered and differentiated. Countering macro-event-driven traders may require more macro-level management of risk exposure; contending with cross-market information relays is an ongoing competition over technology and speed; and defending against on-chain-information parsers requires the market maker itself to possess powerful on-chain-data-analysis capabilities. In this ongoing game, a market maker's survival strategy lies in continuously optimizing its risk-monitoring system so as to identify the true state of the market faster and more accurately than its opponents.

19.2.3 Amplification mechanisms of adverse selection

If the diverse sources of informed trading lay a complex web of threats before a market maker, then several core institutional features of perpetual futures act as powerful amplifiers, significantly heightening the impact of each risk in that web. Leverage, 24/7 uninterrupted trading, and structural information asymmetry together push the adverse-selection problem of perpetual-futures markets to a sharper edge than in any traditional financial market.

Leverage is the first and most direct amplifier. It fundamentally alters the cost-benefit structure of informed trading. An informed trader with a $1 million information advantage, in an unleveraged spot market, has both potential profit and market impact capped at its principal. But in a perpetual-futures market offering 10x leverage, it can use the same $1 million in margin to control a position with $10 million in notional value. This means its ability to attack the market maker's quotes using its information advantage is amplified tenfold. The market maker's bid and ask quotes, originally set to withstand million-dollar-scale impacts, must now face ten-million-dollar-scale impacts. This disproportionate risk exposure forces market makers to treat the leverage level as a key parameter in their pricing models. As a result, all else equal, the bid-ask spread of perpetual futures (measured in basis points) is almost always systematically wider than that of the spot market for the same asset. This extra spread is precisely the risk premium the market pays for leverage's "double-edged" effect.

This amplification is not linear but exhibits superlinear (convex) growth. As the market's average effective leverage rises from low (1–5x) to medium (5–10x) and then to high (10–20x and above), the implicit cost markup a market maker requires to compensate for the amplified impact of informed trading rises faster than the linear increase in the leverage multiple itself. This convexity arises from a key dual mechanism: higher leverage not only amplifies the notional impact of a single informed trader (a 10x-leveraged informed trader's impact on the order book is 10 times its principal) but also lowers the capital threshold for informed traders to enter the market, thereby attracting more potential informed traders. The market maker faces not only a greater impact per informed order but also a rise in the probability of informed trading, μ, itself. The amplifying effect of leverage therefore has a "both-rise" character: each informed trader's impact is larger, and the total number of informed traders is greater. This structurally explains why high-leverage perpetual-futures markets have higher transaction costs.

Second, 24/7 uninterrupted trading, while convenient for users, also creates structural windows of fragility for market makers—the so-called "trough periods." The intensity of activity in global financial markets follows an intraday pattern dominated in turn by the Asian, European, and North American sessions. In the gaps where trading activity in different time zones hands off—for example, the deep night of the Asian time zone and the early morning of the North American time zone—overall market volume drops significantly, and the number of market makers participating in quoting also falls. This causes order-book depth and liquidity to decline sharply, forming "liquidity troughs." These trough periods become ideal operating windows for informed traders and market manipulators. The same large order that might have only limited impact during peak trading hours can, during a liquidity trough, exhaust one side of the order book's depth

and cause a sharp price impact. The adverse-selection risk a market maker faces during these periods is sharply amplified. In response, a rational market maker systematically widens spreads, reduces quoting depth, and even withdraws from the market entirely during trough periods to avoid risk. This behavioral pattern ultimately manifests as a clear time-zone-based intraday pattern in the bid-ask spread, which peaks during liquidity troughs.

The following figure simulates the typical variation of the BTC perpetual-futures bid-ask spread over a 24-hour cycle: the spread is widest in the early Asian morning—the period of thinnest global liquidity—and narrowest when the European and North American sessions overlap and quoting competition is most intense, so the entire curve exhibits a single-peak, single-trough shape. This corroborates the earlier mechanistic account that "liquidity falls during trough periods and spreads widen accordingly," and provides an intuitive illustration that "24/7 trading increases adverse-selection risk."

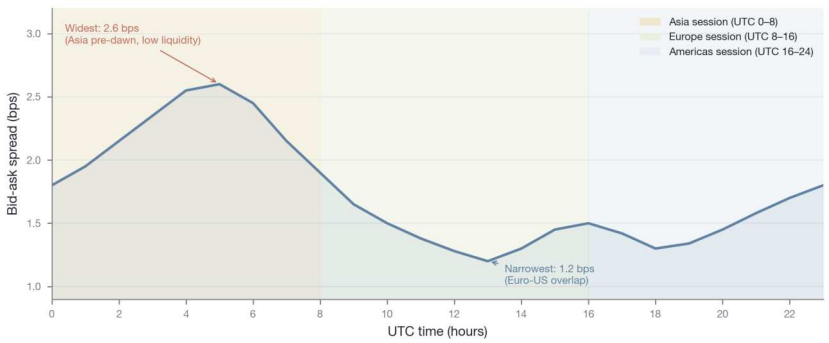


Figure 19-2. Intraday variation pattern of the BTC perpetual-futures bid-ask spread (Data source: Kaiko Research, Binance BTCUSDT perpetual-futures order-book statistics, December 2024–March 2025; this figure shows an illustrative intraday pattern, and the plotted curve is a representative shape rather than an hour-by-hour empirical sample)

Finally, the very position of the perpetual-futures market within the crypto ecosystem also creates a form of structural information asymmetry. As argued in Chapter 13 of this book (price discovery), because of their high liquidity, low transaction costs, and high leverage, the perpetual-futures markets of mainstream crypto assets are often the center of price discovery for the entire crypto world. This means that the latest information about these assets is usually reflected first in the perpetual-futures market. Those who trade first in the perpetual-futures market therefore hold a systematic information advantage relative to traders in other markets (including spot market makers). Perpetual-futures market makers, though situated at the heart of price discovery, are passive liquidity suppliers by role, which dictates that they are always the ones who learn information "second," while the first to know are the informed traders who initiate the trades. This information disadvantage, endogenous to the level of market microstructure, is a structural constraint on perpetual-futures market makers and constitutes a stable, persistent component of their adverse-selection cost.

Taking the three amplification mechanisms together, we can decompose a market maker's total adverse-selection cost into three layers: the baseline adverse-selection

cost (the information-asymmetry cost that exists even in an unleveraged market with fixed trading hours), the leverage-amplification layer (the additional cost induced by the high leverage of perpetual futures), and the structural-information-asymmetry layer (the cost jointly produced by 24/7 trading, the funding-settlement cycle, and the perpetual future's position as a center of price discovery). When the market operates calmly, the relative weights of these three layers are roughly balanced, and the total adverse-selection cost remains at a level a market maker can cover with normal spreads. However, when the market enters a high-volatility state—especially the extreme conditions of a liquidation cascade—the costs of the leverage-amplification layer and the structural-information-asymmetry layer expand sharply and nonlinearly. The leverage-amplification effect is further intensified by the forced liquidations that liquidation triggers (liquidation itself creates new one-directional order flow), and information asymmetry also widens as market participants' ability to interpret information diverges amid panic. The resonance of these three layers pushes the total adverse-selection cost to extreme values far above normal levels.

In summary, leverage, 24/7 trading, and structural information asymmetry together constitute the triple amplification mechanism of adverse-selection risk in perpetual-futures markets. They increase the absolute cost of adverse selection and make it vary in complex, dynamic ways, posing a systematic challenge to a market maker's risk-management and pricing capabilities beyond anything in traditional markets.

19.2.4 Toxicity quantification and risk monitoring

Faced with an adverse-selection threat that is amplified by multiple mechanisms and complex in origin, a market maker relying on intuition and manual adjustment alone cannot effectively cope with these high-frequency, multi-source risks. To survive, modern market makers—especially institutions engaged in high-frequency market making—must build a sophisticated quantitative risk-monitoring system to detect and assess the "toxicity" of order flow in real time and adjust their quoting strategy accordingly. Order-flow toxicity is, in essence, another way of describing adverse-selection risk; it measures the likelihood that trading with the current order flow will lead to a loss over some future period. This section introduces several core toxicity metrics, particularly VPIN, and examines how they are integrated into a market maker's decision framework.

Among the many toxicity metrics, the volume-synchronized probability of informed trading (VPIN), formally developed by Easley, López de Prado, and O'Hara (2012), is the most widely studied and applied [12]. The core idea of VPIN is to modify the traditional PIN (probability of informed trading) model to make it more suitable for the high-frequency world. The traditional PIN model infers the probability of informed trading by analyzing the arrival rates of buy and sell orders. But this "clock-time"-based sampling method runs into problems in the high-frequency world: trading activity is extremely uneven over time, with a large volume of trades possibly erupting within seconds and calm running for minutes. Analysis based on fixed time intervals can easily make wrong judgments during sparse-trading periods or be overwhelmed by information during dense-trading periods.

VPIN's core methodological innovation is that it replaces "clock time" with "volume

time." Rather than partitioning data by fixed time intervals (such as every minute), it partitions data by fixed volume, forming so-called "volume buckets." Within each volume bucket, VPIN computes the absolute value of the difference between buy-side volume and sell-side volume—the "order-flow imbalance." The logic behind this is that informed traders, in order to build positions quickly, typically consume liquidity on the order book continuously and in one direction, thereby causing a significant order-flow imbalance. Conversely, the buying and selling of noise traders is more likely to offset over the short term, resulting in a smaller order-flow imbalance. By continuously computing the average order-flow imbalance across a series of volume buckets, VPIN ultimately produces a value between 0 and 1. A high VPIN value means a high degree of order-flow imbalance in the recent market, a high probability of informed trading, and strongly "toxic" order flow. Conversely, a low VPIN value indicates that the market is dominated by noise trading and the environment is relatively "safe."

The power of VPIN is that it has been shown to be an effective leading indicator of short-term market volatility, especially the "flash crashes" triggered by liquidity evaporation. When the VPIN value climbs steadily and exceeds a certain threshold, it usually foreshadows an imminent sharp price move. For a market maker, VPIN functions as a toxicity-warning indicator. When the VPIN reading is low, the market maker can offer narrow spreads and deep liquidity and earn a steady spread income. But when VPIN begins to surge, the system issues an alert, and the market maker's algorithms respond immediately: rapidly widening bid-ask spreads to compensate for the increased adverse-selection risk, cutting quoting depth to reduce risk exposure, and even, in extreme cases, withdrawing quotes entirely and waiting for market pressure to ease. The VPIN value is positively correlated with order-flow toxicity, and when VPIN enters a high range, market makers typically make more aggressive defensive quoting adjustments.

Nevertheless, VPIN as a measurement tool also has significant limitations. Its predictive power is contested in the academic literature, with some arguing that after controlling for volatility, VPIN's independent predictive value may drop substantially. In the special environment of crypto markets, the fake volume generated by wash trading may further weaken the accuracy of VPIN's buy-sell classification algorithm, because fake trades may artificially create a symmetric order flow that masks the true order-flow-imbalance signal. This also explains why, in practice, market makers do not rely on the single metric of VPIN alone. In the distinctive perpetual-futures market, a market maker's risk-monitoring system typically integrates more dimensions of targeted toxicity signals. One important class of signals stems from the mechanisms of perpetual futures themselves. For example, a rapid, large change in the funding rate often signals a sharp imbalance between long and short forces in the market, behind which strongly directional informed trading may well be hidden. Likewise, a swelling of the liquidation-queue size published by an exchange is a clear toxicity-warning signal, foreshadowing an imminent, high-impact "toxic flow" driven by forced liquidations. Astute market makers watch these indicators closely and cut their risk exposure before the liquidation wave arrives.

Another class of signals comes from a deeper analysis of order-book microstructure. For example, the frequency and size of large market orders is a key indicator. Informed

traders, to ensure immediacy of execution, tend to favor market orders. Thus, when the ratio of market-order volume to total volume per unit of time rises sharply, it usually means that order-flow toxicity is increasing. In addition, analyzing the "order-taking" behavior patterns of the order book—for example, whether a large order "sweeps" through multiple price levels or merely consumes the liquidity at the best price—can also provide clues for judging the intent behind it.

In practice, a cutting-edge market-maker risk-monitoring system is a complex decision matrix that fuses dozens or even hundreds of indicators. It includes not only core toxicity metrics such as VPIN, funding rates, liquidation data, and the market-order ratio, but may also integrate cross-exchange spread data, anomalies in on-chain whale addresses, social-media sentiment indices, and even the volatility indicators of related assets (such as U.S. equity-index futures). All these data streams are fed in real time into a machine-learning model, which dynamically assesses the current market's composite "toxicity score" and, according to preset risk parameters, adjusts the width, depth, and skew of quotes at the millisecond level. This is no longer a simple "quote-fill" loop but an ongoing process of information processing, data analysis, and risk control. The sensitivity, comprehensiveness, and response speed of the monitoring system directly determine a market maker's probability of survival in the asymmetric game against informed traders.

The analytical framework of this section has consistently taken classic "informed trading"—the order flow driven by information asymmetry that causes permanent price impact—as its core object. Yet the harmful order flow a perpetual-futures market maker faces in actual operation is far from this alone. The extractive behavior of arbitrageurs exploiting cross-platform spreads, the forced-liquidation instructions that pour in during a liquidation cascade, and MEV attacks in the on-chain environment all likewise inflict losses on market makers on the surface, but they differ fundamentally from informed trading in information content, the persistence of price impact, and the optimal response strategy. Lumping these risk sources of very different natures together under the category of "adverse selection" would lead to systematic bias in strategy design. Section 19.4 of this chapter systematically extends the discussion, proposing a "toxic-flow spectrum" classification framework that identifies and analyzes, beyond adverse selection, three other toxicity sources with different dominant causes.

19.3 Inventory risk

Adverse selection constitutes a market maker's core risk on the information dimension, whereas inventory risk is its persistent risk exposure on the position dimension. As a passive participant in the market, a market maker is by its very nature unable to choose its counterparties or trading direction. When buy orders outnumber sell orders, it is forced to accumulate a short position; when sell orders outnumber buy orders, it is forced to accumulate a long position. This net holding, passively formed in the course of trading, is the market maker's "inventory." Every unit of inventory is an exposure subject to the risk of market price fluctuations. Even if a market maker has no view whatsoever on the market's long-term direction, this holding, produced purely by business activity, itself constitutes one of its core sources of risk. Managing

inventory, keeping it within a controllable range of risk, and ultimately "clearing" it through subsequent offsetting order flow is the central task of a market maker's daily operations.

In any market, inventory management is a challenge. Yet the distinctive institutional design of perpetual futures raises the difficulty of this challenge to an entirely new magnitude. Market makers in traditional financial markets, though they also face inventory risk, usually have a reliable "safety anchor." For example, an equity market maker can hedge after the close through various cross-market instruments, while a futures market maker enjoys an ultimate certainty—the expiration and delivery date, at which all positions are forcibly closed, the basis converges to zero, and inventory risk is naturally lifted. A perpetual-futures market maker, by contrast, operates in a risk environment with no definite endpoint: the indefinite risk exposure brought by no expiration, the funding rate as a randomly fluctuating holding cost, and the highly nonlinear inventory tail risk under fat-tailed volatility together push inventory management to an entirely new level of difficulty. This section analyzes these three challenges in depth, revealing why and how, in the context of perpetual futures, inventory evolves from a routine business risk into a major source of risk for the market maker.

19.3.1 Adapting classical inventory models to perpetual futures

The academic study of market-maker inventory risk has a long history, and the most representative work is the model proposed by Ho and Stoll (1981) [6]. The core idea of the model is that, to manage its inventory level, a market maker dynamically adjusts its quotes. Specifically, a market maker's optimal quotes are not simply symmetric on either side of the market mid-price but are "skewed" according to its current inventory level. When a market maker has accumulated too much long inventory, it shifts its entire quote range downward: lowering the ask to attract buyers, while lowering the bid more sharply to suppress sellers, thereby incentivizing the arrival of offsetting order flow and helping to reduce its inventory. Conversely, when it has too much short inventory, it shifts its quote range upward. The degree of this quote skew depends on the market maker's inventory level, degree of risk aversion, and judgment about future price volatility.

The Ho-Stoll model provides a systematic theoretical framework for understanding inventory management, describing a market maker's behavior as an optimization process that seeks a balance between "earning the spread" and "controlling inventory risk." Building on this, Avellaneda and Stoikov (2008) formalized the problem of optimal bid and ask quotes under inventory aversion, proposing a standard framework in which quotes are set symmetrically around a "reservation price" and skewed by an amount proportional to the risk-aversion coefficient, the price variance, and the inventory level (i.e., $\delta \propto \gamma\sigma^2 Q$) [13]; this framework became the benchmark for subsequent high-frequency market-making quote optimization. However, when we apply this classic framework directly to perpetual-futures markets, we find that its underlying assumptions are significantly disconnected from market reality. This disconnect is reflected mainly in three respects, which force us to make key adaptations and extensions to the model.

First, the traditional model defines holding cost too simply. In the classic framework,

the holding cost of inventory comprises mainly two parts: the opportunity cost of capital (i.e., funding interest) and the risk exposure from price fluctuations. The former is relatively stable and predictable, and the latter is usually assumed to follow a normally distributed random walk. But in perpetual futures, a brand-new cost item with a significant impact on P&L is introduced: the funding rate. The net inventory a market maker holds must pay or receive a fee in every funding-settlement period. Unlike relatively stable interest, the funding rate is a highly uncertain random variable whose direction and magnitude may change sharply over short periods. This means that the inventory holding cost of a perpetual-futures market maker contains an additional, dynamic stochastic process, which greatly increases the complexity of cost estimation and risk management.

Second, the traditional model implicitly assumes that "inventory is disposable." Whether the end-of-day settlement of the equity market or the expiration and delivery of the futures market, each provides a market maker with a mechanism to manage or clear its inventory at a specific point in time. The "no expiration" feature of perpetual futures completely breaks this assumption. A market maker's inventory management no longer has any "natural endpoint" and must rely entirely on active, continuous hedging—for example, attracting offsetting flow through quote skewing within the same exchange, or building an offsetting position on another exchange or in the spot market. These active hedging actions themselves carry additional costs and risks, such as trading fees, cross-market basis risk, and execution-latency risk.

Third, the traditional model's assumption about price-volatility risk is too benign. The model usually assumes that price fluctuations follow a normal distribution or a similar "well-behaved" distribution. However, as repeatedly emphasized elsewhere in this book, crypto markets—especially high-leverage perpetual-futures markets—exhibit pronounced "fat-tailed" characteristics in their price distributions. The "jump" risk triggered by mechanisms such as liquidation cascades is far beyond what a normal distribution can describe. This means that the tail risk facing a market maker's inventory (i.e., the possibility of an extreme loss) is systematically underestimated. The relationship between inventory cost and volatility is also not linear but exhibits a convex relationship that grows rapidly in the high-volatility region.

Comparing the inventory-cost structures of a traditional futures market maker and a perpetual-futures market maker makes the urgency of these adaptation needs clearer. For a traditional delivery-futures market maker, the composition of the inventory holding cost is relatively simple: price-volatility risk dominates (usually accounting for the larger share of total cost), followed by the opportunity cost of capital, with the remainder being trading fees and operational costs. Because of the existence of an expiration date, the time boundary of these costs is definite, and the total can be estimated fairly accurately. For a perpetual-futures market maker, the cost structure is fundamentally reconstructed: although price-volatility risk remains important, its relative share of total cost falls with the introduction of institutional cost; funding-rate exposure becomes an independent cost item comparable in scale to price-volatility risk, and under extreme market conditions it may even exceed the latter; and the liquidation risk premium and mark-price deviation risk each contribute a non-negligible incremental cost. More critically, these institutional-cost items are not independent of one another: high volatility

simultaneously pushes up the extreme values of the funding rate and the probability of triggering liquidation, forming a positive correlation among the cost items that makes the total cost exhibit superlinear growth in extreme environments.

Therefore, an inventory-management model adapted to the perpetual-futures environment must incorporate these three new variables—the randomness of the funding rate, the indefinite risk exposure brought by no expiration, and the nonlinear risk under a fat-tailed distribution—into its core equation. If Section 19.1.3 defined "what" institutional cost is from the perspective of cost classification, then the task of this section is to answer, from the perspective of modeling, "how" institutional cost affects inventory-management decisions. In the classic Ho-Stoll and Avellaneda-Stoikov frameworks, a market maker's optimal quote-skew amount δ^* is determined mainly by the inventory level Q and the price volatility σ: $\delta^* = f(Q, \sigma, \gamma)$, where γ is the risk-aversion coefficient. In the perpetual-futures environment, this decision function must be extended to $\delta^* = f(Q, \sigma, \gamma, FR_t, T_{\text{eff}}, \kappa)$, where FR_t is the current funding rate (affecting the directional preference of inventory holding), T_{eff} is the effective time horizon (replacing the now-vanished expiration-date constraint—this is not an exogenously given parameter but a decision variable the market maker must optimize endogenously; in practice, many market makers use the 8-hour funding-settlement cycle as a natural time anchor, or adopt a multi-timescale nested risk-control framework to manage short- and medium-term inventory risk simultaneously), and κ is a tail-risk adjustment coefficient (capturing the nonlinear amplifying effect of fat-tailed events such as liquidation cascades on inventory cost). This extended decision function is analyzed term by term in the sections that follow.

19.3.2 The inventory dilemma of no expiration

For a market maker operating in a traditional delivery-futures market, the expiration date constitutes a key anchor for risk management. However tortuous the trading within a contract cycle, and however out of control the inventory level, the approach of the expiration date signals the return of certainty. As the delivery date arrives, the basis between the futures price and the spot price inevitably converges to zero. All long or short positions the market maker holds are delivered or cash-settled at the final settlement price, and inventory is "naturally" cleared to zero. This process provides a definite time boundary for a market maker's risk management. It can plan its inventory exposure within a definite time frame, and even if inventory sometimes deviates substantially, it knows that the term of its risk exposure is finite. The expiration date sets a definite ceiling on a market maker's risk exposure.

A perpetual-futures market maker does not have this institutional safeguard. One of the core designs of the perpetual future is precisely that it is "perpetual"—it has no expiration date. While convenient for traders, this feature creates enormous trouble for market makers: their inventory risk exposure is, in theory, indefinite. If the market exhibits a sustained one-directional trend—for example, in an uptrend lasting weeks—a market maker's sell orders are continuously hit by buyers, causing its passively accumulated short inventory to keep swelling. Because there is no expiration date as a "terminus," this short position does not close automatically; it remains exposed day

after day to the risk of continuously rising prices, and losses can expand without limit. The market maker's only way out is to conduct active, continuous hedging.

The cost of such active hedging is high and varied. The most direct approach is to attract offsetting order flow within the same exchange through "quote skewing" (the mechanism by which it adjusts quotes according to inventory direction is described in Section 19.3.1). But the price of this approach is that it may forgo opportunities to trade at more favorable prices, and if the market keeps rising, the act of high-priced hedging itself causes losses. Another approach is cross-market hedging—building an opposite position in the perpetual-futures market of another exchange or in the spot market. For example, if it has accumulated a short inventory of 100 BTC on exchange A, the market maker can buy 100 BTC on exchange B or in the spot market. Although this can neutralize the nominal risk exposure, it introduces new risks and costs: first, double trading fees; second, cross-exchange or cross-market basis risk, i.e., the spread between the two markets may move unfavorably; and finally, execution-latency risk, in that a delay in the hedging operation in a fast-moving market can cause huge slippage losses.

There is a fundamental difference in the inventory-evolution path between delivery futures and perpetual futures. Take a market maker that initially holds a short inventory of 100 BTC. In the delivery-futures market, even if this 100-BTC short position at one point expands to 200 or even 300 during the contract cycle, as the expiration date approaches, the basis-convergence effect causes the market to spontaneously generate convergent order flow, and the market maker's inventory level necessarily returns toward zero. The expiration date is like a "center of gravity": however far the inventory departs along the way, it will ultimately be pulled back to zero. This is a deterministic process embedded in the institutional design. In the perpetual-futures market, by contrast, the same 100-BTC short inventory has no such "center of gravity." If the market rises continuously for a month, this 100-lot short position not only fails to shrink automatically but may even swell further to 200 or 300 as the market maker keeps passively selling during the rise. The evolution of the inventory level depends entirely on the degree of the market's one-directional move and the efficiency of the market maker's active hedging. In the absence of an expiration-date constraint, a market maker's risk exposure has no natural convergence mechanism and can only rely on costlier active hedging to control risk. The practical implication of this path difference is significant: a delivery-futures market maker can treat an inventory deviation as "temporary," whereas a perpetual-futures market maker must treat every inventory deviation as "potentially permanent."

At a deeper level, "no expiration" also brings a cognitive dilemma—the "time horizon" problem of inventory management. Traditional market makers usually have clear risk-assessment cycles, such as end-of-day or end-of-week, at which they must ensure that inventory levels and risk exposures are within an acceptable range. The 24/7 uninterrupted trading feature of perpetual futures makes this kind of periodic assessment blurry. Risk is continuous, with no rest window. A market maker faces a thorny question: "On what timescale is my inventory level 'healthy'?" Is it one hour, eight hours, or a day? Too short a timescale may lead to overly frequent and expensive hedging ("overreaction"), while too long a timescale may leave risk exposed for too long

("underreaction"). This blurriness of the time horizon makes an optimal inventory-management strategy extraordinarily difficult to formulate and places heavy demands on a market maker's risk-control system.

19.3.3 Funding rates and inventory holding costs

If the no-expiration feature defines the "time" dimension of inventory risk, then the funding rate defines its "cost" dimension. The funding rate anchors to the spot price by periodically transferring fees between longs and shorts. For a market maker that passively holds inventory, however, this mechanism turns into a highly uncertain key variable that affects its profitability. The causal chain is clear: a market maker accumulates net inventory through passive fills → it is forced to participate in the transfer of fees between longs and shorts at every funding-settlement point → that inventory therefore incurs a payment or receipt determined by the direction of the rate. Whatever net inventory it holds, long or short, must participate in the payment or receipt of the funding rate in every settlement period. This makes the funding rate a new, dynamic, and hard-to-predict component of a market maker's inventory holding cost.

This mechanism can be precisely characterized with a minimal formalization: let Q_t be the market maker's signed net inventory at the settlement point (positive for long, negative for short), and FR_t be the funding rate at that point (a positive value means longs pay shorts); then the market maker's funding P&L in a single settlement period is

$$\text{Per-period funding P\&L} = -FR_t \cdot Q_t$$

When FR and Q have the same sign (e.g., holding long inventory under a positive rate), this term is negative—that is, the market maker pays; when FR and Q have opposite signs (e.g., holding short inventory under a positive rate), this term is positive—that is, the market maker receives. The following table is precisely the natural expansion of this expression across the four quadrants:

Funding rate	Market maker's net inventory	Sign of $-FR_t \cdot Q_t$	Effect on the market maker
Positive (longs pay)	Long (Q > 0)	Negative	Pays the fee; cost increases (implicit tax)
Positive (longs pay)	Short (Q < 0)	Positive	Receives the fee; cost decreases (implicit subsidy)
Negative (shorts pay)	Long (Q > 0)	Positive	Receives the fee; cost decreases (implicit subsidy)
Negative (shorts pay)	Short (Q < 0)	Negative	Pays the fee; cost increases (implicit tax)

Table 19-1. The correspondence between funding-rate direction and a market maker's inventory holding cost (Data source: derived by the author from the single-period funding P&L $-FR_t \cdot Q_t$)

The effect of the funding rate on a market maker's inventory is two-sided: it can be either an "implicit tax" or an "implicit subsidy." When the funding rate is positive,

market sentiment leans long, and longs must pay shorts. In this case, if the market maker holds net long inventory through passive fills, it must pay this funding fee, which directly increases the holding cost of its inventory. Conversely, if it holds net short inventory, it can receive this fee, and this income can partly or wholly offset the risk cost of its position, becoming a subsidy. When the funding rate is negative, the situation is exactly reversed. Table 19-1 summarizes this correspondence.

The volatility of the funding rate varies significantly across assets, further compounding the difficulty of inventory management for market makers. The following figure compares the 8-hour funding-rate distributions of four assets—BTC, ETH, SOL, and DOGE—whose cross-asset differences in standard deviation and share of positive rates are pronounced (BTC is the most concentrated, while SOL's volatility is nearly three times as high; see the statistics box in the figure). These differences directly affect market makers' pricing strategies and inventory management across pairs.

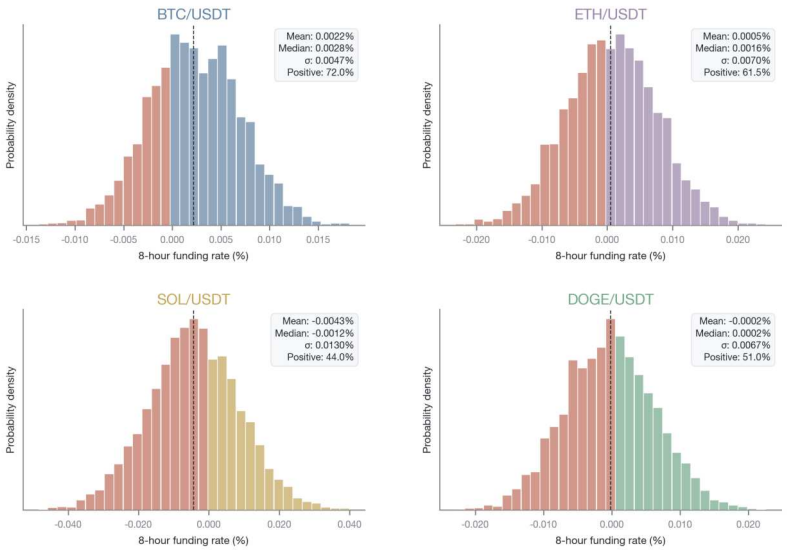


Figure 19-4. Comparison of 8-hour funding-rate distributions across assets (Data source: Binance perpetual-futures funding-rate history, December 2024–March 2025; the mean, median, standard deviation, and share of positive rates in the statistics box are authoritative values, while the histogram shape is an illustration reconstructed from the published means and standard deviations rather than a tick-by-tick measurement)

> Note on figure numbering: Figures in this chapter are numbered starting from Figure 19-1, in the sequence Figure 19-1, 19-2, 19-4, … 19-13. Figure 19-3 was reserved during layout and ultimately merged and omitted, so it does not appear in this chapter; to keep the figure numbers stable, the remaining numbers are not renumbered sequentially, and all original numbers are retained unchanged. The cross-sectional comparisons of the two paradigms and the three markets are instead carried by newly added tables (Table 19-3 and Table 19-4); see Sections 19.5.5 and 19.6.3.

The funding rate is itself a stochastic process, driven by the combined influence of market sentiment, leverage levels, arbitrage activity, and other factors, and it exhibits pronounced volatility and unpredictability. As shown in the following figure, a typical

time series switches frequently between positive and negative values, with occasional extreme highs or lows. For market makers, this means that the inventory holding cost embeds a random term that cannot be determined in advance—a stark contrast with the stable, predictable funding costs of traditional finance (such as the overnight rate).

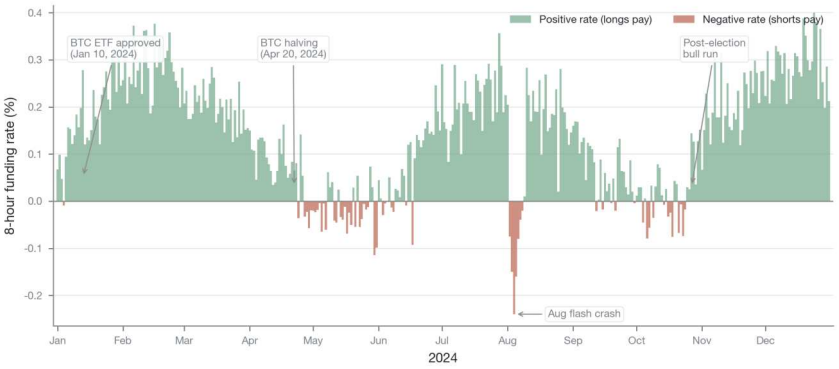


Figure 19-5. Time series of the 8-hour funding rate for BTC perpetual futures (Data source: CoinGlass Funding Rate History, January–December 2024; this figure is a representative, illustrative time series meant to convey the qualitative pattern of positive-negative switching and extreme values over the year rather than a period-by-period empirical sample, and the vertical-axis magnitude has been scaled to the realistic 8-hour funding-rate range of about $\pm 0.01\%$ to $\pm 0.375\%$)

Beyond market-internal sentiment, the overall level of the funding rate is also driven structurally by the global macro interest-rate environment. In a low-interest-rate environment (such as 2020–2021), the large spread between the low funding costs of traditional financial markets and the high leverage demand of crypto markets pushed the positive funding rate of perpetual futures steadily higher; whereas during a global rate-hiking cycle (such as 2022–2023), the rise in traditional funding costs was partly transmitted to crypto markets, compressing the carry space [9]. This means that a market maker's inventory holding cost embeds not only the microstructural randomness of the funding rate but also the slow-variable influence of the macro interest-rate cycle.

This randomness introduces new complexity into a market maker's strategy. On one hand, a market maker has an incentive to actively manage its inventory direction so that it aligns with the expected direction of the funding rate. For example, if the market broadly expects the funding rate to stay positive for some time (such as in a bullish atmosphere), a market maker will lean toward holding net short inventory, because this allows it to steadily "collect rent." It may adjust its quote-skewing strategy to make it easier to accumulate short positions. However, this introduces an entirely new risk dimension: a "directional bet" on the funding rate. If market sentiment suddenly reverses and the funding rate flips from positive to negative, then the short inventory that had been a source of subsidy instantly becomes a burden requiring high fee payments, causing a double loss from both "price fluctuations" and "the funding rate."

Furthermore, there is a subtle "reflexivity" between a market maker's behavior and the funding rate. We must first clarify the tension between this reflexivity and the earlier

point that "market makers passively hold inventory and cannot choose direction" (see Section 19.1.2 and the opening of Section 19.3): on the basis of the inventory formed by passive fills, a market maker can still, at the margin, actively lean toward a particular inventory direction through quote skewing; reflexivity operates precisely on this marginal active portion and does not negate its fundamental nature as a passive liquidity supplier. As the analysis of the funding-rate game in Chapter 10 of this book shows, the overall long-short position imbalance in the market is the key factor determining the direction and magnitude of the funding rate. As the most important liquidity providers in the market, market makers' marginal choice of inventory direction (i.e., whether they lean toward accumulating long or short inventory) affects, to some degree, the market's net-position balance. For example, if most market makers lean at the margin toward holding short inventory in anticipation of a positive rate, their collective behavior itself may intensify the market's short lean and thereby, in turn, push the funding rate down. This reflexive feedback loop is a theoretical inference drawn from Chapter 10; this chapter does not test it independently or empirically. This feedback loop between individual decisions and macro market indicators makes the dynamics of the funding rate even more complex, and it also means that any misjudgment of the rate direction by a market maker can turn the holding cost from a subsidy into a penalty, causing significant additional losses. The following figure shows the dynamic relationship among open interest, price, and the funding rate for BTC perpetual futures: rapid increases in open interest often accompany one-directional price movements, while the funding rate exhibits a pronounced clustering effect—positive and negative funding rates each form continuous "clusters" rather than a random distribution. This clustering is directionally consistent with the theoretical inference about funding-rate reflexivity and can be regarded as observational corroboration.



Figure 19-6. The dynamic relationship among open interest, price, and the funding rate for BTC perpetual futures (Data source: Binance BTCUSDT perpetual-futures open-interest and funding-rate data, February–March 2025; this figure is a representative, illustrative series in which the magnitudes of open interest, price, and funding rate follow public baselines rather than a period-by-period empirical sample)

19.3.4 The nonlinear relationship between volatility and inventory cost

In classical inventory-management models, the holding cost of inventory is usually assumed to be proportional to market volatility. This linear assumption is a reasonable and effective approximation in low-volatility, "normal" market environments. It means that if the market's annualized volatility doubles, the risk compensation a market maker requires per unit of inventory also roughly doubles. However, in perpetual-futures markets, where leverage, liquidation cascades, and information-transmission effects act together, this linear assumption deviates significantly from actual data. The relationship between a perpetual-futures market maker's inventory cost and volatility is nonlinear—specifically, convex.

A convex relationship means that as volatility increases, the inventory cost grows at an ever-accelerating rate. To make the magnitude of this convexity intuitive, consider the following illustrative numbers (a scenario constructed from market makers' empirical characteristics for explanatory purposes, not a measured-sample statistic): when the annualized volatility of BTC perpetual futures is in the low-volatility range below 30%, a market maker holding $1 million of net inventory incurs a daily risk cost roughly on the order of $200–500, broadly consistent with the linear assumption's prediction. But when volatility climbs to the medium-high range of 50–70%, the daily risk cost of the same-sized inventory may grow to $1,500–3,000, an increase far exceeding the growth in volatility itself. And in an extreme environment where volatility surges above 100% (such as during the global risk-asset sell-off of August 2024), the inventory cost may grow in a superlinear (convex) manner to more than $10,000 per day, on the order of 20 times the low-volatility level. The three tiers of figures above are meant to convey the qualitative conclusion that "cost rises at an accelerating pace with volatility," not to serve as precise, calibratable estimates. The traditional linear assumption and the convex reality of perpetual futures differ little in the low-volatility region but diverge rapidly in the high-volatility region. The ever-widening gap between the two is caused precisely by the distinctive risk-amplification mechanisms of perpetual futures (i.e., liquidation cascades, liquidity contraction, and the superposition of institutional cost).

The core reason for this convex relationship is that, in perpetual-futures markets, a high-volatility environment is not an isolated variable; it is often activated simultaneously with multiple negative-feedback mechanisms. First, high volatility is an important trigger of large-scale "liquidation cascades." When prices swing sharply, the margin of many high-leverage positions is punched through, triggering forced liquidations by the exchange. These forced-liquidation market orders further impact the market, driving prices further in the same direction and thereby triggering the liquidation of more accounts. In this positive feedback loop, if a market maker holds an opposing inventory, its losses are amplified disproportionately. It faces not a smooth random walk of prices but violent, cliff-like price "jumps."

Second, a high-volatility environment is always accompanied by a sharp contraction of liquidity. This mechanism is essentially identical to the "liquidity spiral" identified by Brunnermeier and Pedersen (2009) [14] in traditional financial markets: rising volatility $\rightarrow$ higher margin requirements $\rightarrow$ forced deleveraging $\rightarrow$ falling liquidity $\rightarrow$ amplified price impact $\rightarrow$ further rising volatility. In the high-leverage liquidation environment of

perpetual futures, this spiral is further accelerated, because the forced-liquidation mechanism turns "forced deleveraging" from a gradual process into a threshold-triggered discrete event. As discussed earlier in this chapter, when adverse-selection risk and inventory risk rise sharply, a market maker's rational choice is to widen bid-ask spreads, reduce quoting depth, and even withdraw from the market entirely to protect itself. This collective "risk-avoidance" behavior makes the order book extraordinarily thin during high-volatility periods. For market makers that remain in the market and hold inventory, this means that when they try to actively close or hedge their risk positions, they face enormous slippage costs. Inventory that could originally be managed with low-cost operations becomes "toxic" and hard to dispose of in a high-volatility environment, and its value at risk rises sharply.

Finally, a high-volatility environment also amplifies other institutional costs. For example, the deviation between the mark price and the latest traded price becomes larger, increasing the risk of needless liquidation; and the funding rate may become extreme over a short period, imposing a heavy penalty on market makers holding inventory in the unfavorable direction. All these factors together constitute the "extra cost" beyond the linear assumption. This extra cost is nearly invisible in a low-volatility environment but expands sharply in a high-volatility environment, becoming the major component of a market maker's inventory holding cost.

Using historical data for BTC perpetual futures, the following figure divides the 14-day rolling volatility into low, medium, and high regimes: in the high-volatility regime (annualized volatility above 40%), the mean intraday price range for BTC is about 1.6 times that of the low-volatility regime, and the dispersion of fluctuations increases markedly with extreme values appearing frequently. This nonlinear volatility-amplification effect is directionally consistent with the claim that market makers' inventory costs grow convexly.

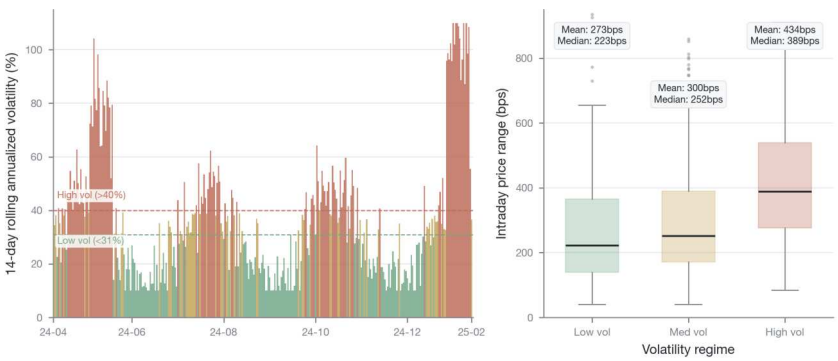


Figure 19-7. BTC volatility regimes and intraday price range (Data source: Binance BTCUSDT perpetual-futures daily candlestick data, March 2024–March 2025)

Going further, the following figure reveals the relationship between volatility and the funding rate: the time series on the left shows a temporal resonance between periods of surging volatility (such as August and December 2024) and extreme funding rates, while the scatter analysis on the right shows a weak linear correlation between the two

(r = 0.076). This must be interpreted with caution: a near-zero correlation coefficient of r = 0.076 can only refute the existence of a simple linear mean relationship between volatility and the funding rate; it does not by itself constitute heteroskedasticity evidence for "the variance of the funding rate being amplified under high volatility." The causal chain "higher volatility → greater funding-rate variance → surging uncertainty in inventory holding costs" should therefore be understood as a hypothesis awaiting testing rather than a conclusion already confirmed by Figure 19-8; rigorously testing this hypothesis would require further bucketing the funding rate by volatility and examining its standard deviation and extreme-value statistics, a formal estimation this chapter does not undertake.

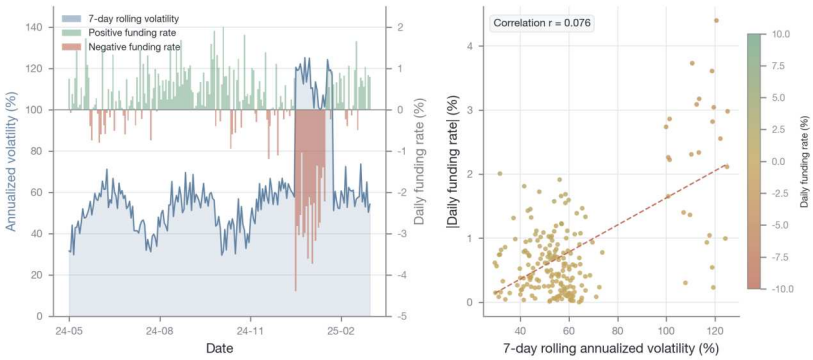


Figure 19-8. The dynamic relationship between BTC volatility and the funding rate (Data source: Binance BTCUSDT perpetual-futures daily candlestick and funding-rate data, March 2024–March 2025)

This nonlinear relationship directly explains market makers' behavioral patterns in high-volatility environments and their impact on market liquidity. It explains why, at the extreme moments when the market most needs liquidity (such as when prices crash or spike), liquidity instead contracts sharply. This is not because market makers lack willingness or harbor subjective malice, but because in that environment the inventory cost they face has grown to a level that normal spreads simply cannot cover. Continuing to provide liquidity would lead to unbearable losses. Withdrawing from the market is the only rational answer given by their profit equation. This also provides the most direct microeconomic-mechanism explanation for the theme we will examine in depth in the next chapter—why liquidity vanishes precisely when it is most needed. The nonlinearity of inventory risk is a key link in understanding the endogenous fragility of perpetual-futures markets.

19.4 The toxic-flow spectrum

In classical market-making theory, "toxic flow" (shorthand for the toxic order flow that inflicts systematic losses on market makers; the remainder of this chapter uses the term "toxic flow" uniformly) is a concept nearly interchangeable with "informed trading." It paints a clear but somewhat monotonous picture: some participants in the market hold nonpublic information—"the informed"—and they trade with market makers whose

information lags relatively behind, causing market makers to incur systematic losses, namely the adverse-selection cost. A market maker's way to survive is to compensate for potential losses by adjusting its quoting spread in a high-risk environment composed of informed trading. However, when this analytical framework is applied to the perpetual-futures market—high-leverage, around-the-clock, and spanning both centralized and decentralized platforms—the sources and impact mechanisms of toxic flow exhibit far greater complexity. Lumping all order flow unfavorable to market makers together as "informed trading" oversimplifies reality and can lead to serious strategic errors.

This section proposes and systematically develops one of this book's integrative classification frameworks: the toxic-flow spectrum. We must first clarify the nature and boundaries of this framework: it does not claim to have "discovered" several entirely new risk types but instead places the harmful order flow perpetual-futures market makers already face into a unified coordinate system composed of "information content" and "permanence of price impact," so as to systematically compare their mechanistic differences and design differentiated responses accordingly. Within this coordinate system, we identify, by "dominant cause," four typical types of toxic flow: fundamentals-driven "informed-trading toxicity," cross-platform-spread-driven "arbitrage-extraction toxicity," "liquidation-cascade toxicity" endogenously determined by the leverage mechanism, and "MEV-attack toxicity" unique to the decentralized environment. These four are a perspective divided by dominant cause, not a mutually exclusive and collectively exhaustive (MECE) set of independent categories: their boundaries can overlap on the coordinate system—"informed-trading toxicity" is essentially the naming, within this coordinate system, of the adverse-selection risk of Section 19.2, while the latency-arbitrage form of "arbitrage-extraction toxicity" on an on-chain CLOB overlaps mechanistically with the second "cross-market information-relay" type of informed trading in Section 19.2.2. The term-by-term analysis of the four types of toxicity below should therefore be understood as a characterization of four dominant regions on the same coordinate plane, not an enumeration of four mutually disjoint sets.

This coordinate system is built on two core dimensions. The first dimension is "information content": whether the order flow embeds new information about an asset's fundamental value. Informed trading has high information content; arbitrage extraction and liquidation cascades have lower information content (the former exploits an already-public cross-market spread, the latter is an endogenous product of the leverage mechanism); and MEV attacks contain no fundamental information. The second dimension is "permanence of price impact": whether the price change triggered by the order flow persists after the fact. The price impact caused by informed trading is almost entirely permanent; the impact of arbitrage extraction has medium permanence (pushing the price toward fair value); the impact of a liquidation cascade contains a large temporary-overshoot component; and the price impact of an MEV attack is usually absorbed within the same block. The following table maps the dominant characteristics of the four types of toxic flow onto this two-dimensional framework:

Toxicity type	Information content	Permanence of price impact	Temporal characteristics	Primary response strategy
Informed-trading toxicity	High	High (permanent)	Event-driven pulses	Widen spreads, reduce depth, avoid
Arbitrage-extraction toxicity	Low	Medium (toward fair value)	Continuous, positively correlated with volatility	Shorten oracle latency, dynamic fees
Liquidation-cascade toxicity	Low	Low (large temporary overshoot)	Sudden pulses, positive feedback	Real-time signal classification, controlled absorption
MEV-attack toxicity	None	Very low (absorbed within the block)	High-frequency, synchronized with blocks	Private transaction channels, MEV-resistant protocols

Table 19-2. The two-dimensional classification framework of the toxic-flow spectrum (Data source: constructed by the author)

This table characterizes the dominant features of the four types of toxicity under normal market conditions, not a mutually exclusive discrete classification. In extreme markets, the four types can superimpose and resonate, and individual types (such as arbitrage extraction) may even reverse direction during the price-recovery phase; the relevant dynamics are detailed in Section 19.4.5. Although the dominant regions of these four types of toxic flow each have their own emphasis, they often interweave and superimpose at the market's extreme moments, forming composite risk events that cause sharp liquidity contraction. Conflating them leads to a serious mismatch of response strategies: using a retreat strategy meant for informed trading to deal with liquidation flow would miss the rebound opportunity, whereas the reverse could incur major losses in a genuine price impact. Understanding this spectrum is the key to understanding the microdynamics of liquidity—and indeed the systemic risk—of perpetual-futures markets. The sections that follow analyze the mechanisms and characteristics of these four types of toxic flow in turn.

19.4.1 Informed-trading toxicity

Informed-trading toxicity is the most classic type in the toxic-flow spectrum—the naming, within this coordinate system, of the adverse-selection risk systematically analyzed in Section 19.2. As noted, incorporating it into the spectrum does not introduce a new risk but provides a common benchmark for the systematic comparison of the four types of toxicity. This section does not repeat its theoretical mechanism (the applicability of the Glosten-Milgrom model (Glosten and Milgrom, 1985 [4]), the four sources of informed trading, the amplifying effects of leverage and 24/7 trading, and quantitative tools such as VPIN have all been elaborated in detail in Section 19.2); instead, it focuses on placing its core characteristics into the comparative framework of the toxic-flow spectrum, so as to form a systematic contrast with the three types of toxicity that

follow.

Informed-trading toxicity has three core defining characteristics. First, its price impact is permanent. The buying and selling of informed traders reflect fundamental information not yet priced by the market, so the price change they drive does not revert after the trade is completed, and the market maker's loss is deterministic and irreversible. Second, its temporal characteristic is an event-driven pulse distribution, erupting in the brief windows before and after the release of major information. Third, its optimal response strategy is "detect and avoid": widening spreads to raise the informed trader's cost, reducing depth to limit maximum loss, and skewing quotes to actively manage inventory direction. These three characteristics—permanent impact, pulse-like temporal distribution, and avoidance-type response—recur repeatedly in the contrast with the three types of toxicity that follow, and the differences among them are precisely the core value of the toxic-flow classification.

19.4.2 Arbitrage-extraction toxicity

If informed-trading toxicity is a game between a market maker and those with an information advantage, then arbitrage-extraction toxicity manifests more as a structural friction between different market mechanisms. This toxicity is especially pronounced in the automated market maker (AMM) or liquidity-vault models of DEXs; it stems from the inherent lag of DEX pricing mechanisms relative to CEXs. A CEX has a high-frequency order-book matching engine, whose internal processing can reach the microsecond level and whose external price updates are at the millisecond level. A DEX, by contrast—especially a protocol that relies on oracle price feeds—has an inherent lag in its price updates, which can range from seconds to minutes. This time difference creates a risk-free arbitrage opportunity.

When the market price swings sharply, the CEX price has already reflected the change quickly, while the oracle price on the DEX has not yet updated. At this point, arbitrageurs step in swiftly. They buy at the latest low price on the CEX and then sell to the liquidity provider (LP) on the DEX at the not-yet-updated, relatively high "stale price"; or they sell at the latest high price on the CEX and then buy from the DEX's LP at the lower stale price. For the DEX's LP, they are forced, in every round of price fluctuation, to trade at a belated, wrong price with the fastest-reacting arbitrageurs in the market. This systematic loss is termed loss-versus-rebalancing (LVR) in the academic literature (Milionis et al., 2022) [15]; it constitutes a structural cost facing DEX LPs and forms a continuous value-extraction effect.

Unlike informed-trading toxicity, the price impact caused by arbitrage-extraction toxicity is not necessarily permanent. The behavior of arbitrageurs is aimed precisely at eliminating the spread, and their trades push the DEX price toward the CEX's fair price. But for the LP, the harm is already done. Their loss comes from the moment of trading at an unfavorable price, and this loss is amplified as market volatility increases, because higher volatility means more frequent and more violent CEX-DEX spreads. Arbitrage-extraction toxicity is therefore not an event-driven pulse but a persistent, structural toxicity positively correlated with market volatility. As long as the informational and mechanistic lag between the CEX and the DEX exists, this value extraction

will not stop.

Faced with this toxicity, the responses of on-chain market makers lie more at the level of protocol design. For example, a protocol can strive to shorten the oracle's update lag and adopt more manipulation-resistant price-feed mechanisms; it can introduce dynamic fees that automatically raise trading fees when market volatility intensifies, thereby increasing the arbitrageurs' cost; or it can weaken arbitrageurs' time advantage through mechanisms such as trading-speed limits and batch auctions. For on-chain market makers, LVR is a structural cost that must be paid to participate in DEX market making, and its return (trading-fee income) must adequately compensate for this foreseeable loss.

The concept of LVR strictly applies to liquidity providers based on AMMs or oracle pricing (such as GMX's GLP model), because its core mechanism is the systematic deviation between the "stale price" and the "latest price." For professional market makers operating on an on-chain CLOB (such as Hyperliquid or dYdX), the form of arbitrage extraction they face differs from LVR and is closer to cross-platform latency arbitrage between CEXs—essentially identical to the second "cross-market information-relay" type of informed trading discussed in Section 19.2.2, except that the higher latency of the on-chain environment amplifies this risk. This is precisely one piece of evidence for the earlier statement that "the boundaries of the four types of toxicity can overlap": the same harmful order flow can, depending on its on-chain or off-chain carrier, fall simultaneously at the boundary of the two dominant regions of "arbitrage-extraction toxicity" and "informed-trading toxicity." The "arbitrage-extraction toxicity" in Table 19-2 should therefore be understood as an umbrella concept encompassing two concrete manifestations: LVR in the AMM environment and latency arbitrage in the order-book environment.

19.4.3 Liquidation-cascade toxicity

Relative to the other three types of toxic flow, liquidation-cascade toxicity has higher instantaneous impact intensity and a stronger positive-feedback character in perpetual-futures markets. Its mechanistic root lies in the positive feedback loop between high leverage and forced liquidation. When a sharp price move triggers the forced liquidation of opposing high-leverage positions, these market-order closings further impact the price, thereby triggering a new round of forced liquidations and forming a dangerous positive feedback loop that generates a huge one-directional order flow in an extremely short time—the liquidation cascade. The empirical analysis by the OECD (2023) of three major DeFi lending protocols—Aave, Compound, and Maker—provides quantitative evidence for the positive feedback loop between liquidations and price volatility [16]; although its subjects are DeFi lending protocols rather than CEX perpetual futures, the liquidation positive-feedback mechanism it reveals is essentially common to both.

This toxicity differs fundamentally from the other three types. The initiators of liquidation orders (the traders being liquidated) are not "informed." Their motive for trading is not that they possess future information but that they are forcibly liquidated due to insufficient margin. However, liquidation flow has characteristics similar to informed-

trading flow in its price-impact effect, and because of its price-insensitive market-order nature, its impact intensity per unit of time is often higher.

The key difference, however, is that the price impact driven by informed trading reflects a permanent change in fundamentals, whereas the price impact driven by a liquidation cascade contains a large "temporary" or "overshoot" component. Liquidation flow is price-insensitive; it merely aims to close out risk exposure as fast as possible, and therefore often pushes the price to an extreme level far beyond its fair value. After the cascade ends, this excessive price deviation is usually partly or even fully corrected, and the price shows a pronounced V-shaped reversal.

The market event of August 5, 2024, can serve as a single illustrative case of the practical value of the toxic-flow spectrum, rather than as a hypothesis test of the framework. A prior caveat is in order: the following quantitative description of the event is synthesized from public exchange historical candlesticks and third-party liquidation data (such as CoinGlass Liquidation Data), with sample time points noted for readers to verify, but a single case alone is insufficient to confirm or refute general claims such as "toxicity superposition." This event has pedagogical value because it clearly demonstrates the process by which multiple types of toxic flow are activated in causal sequence. The initial shock did not originate within the crypto market but was a typical macro-information-driven event: an unexpected rate hike by the Bank of Japan triggered a large-scale unwinding of the global yen carry trade, setting off a chain sell-off of risk assets from Japanese equities to U.S. equities and then to crypto markets. In the first phase of the event, the price decline in crypto markets was driven primarily by macro-information-driven informed-trading flow: traders who were the first to read the signal of a global repricing of risk assets built short positions in the perpetual-futures market—this belongs to the informed-trading toxicity in this section's coordinate system. In the second phase, the continued price decline began to trigger the forced liquidation of many high-leverage long positions, and liquidation-cascade toxicity took over: according to public historical data, the BTC perpetual-futures price fell sharply from about \$62,000 to about \$49,000 within a few hours (a decline of about 21%), and according to third-party liquidation data, over \$1 billion of long positions were forcibly liquidated. During the roughly 30-minute window of most intense liquidation, the bid-side order-book depth of the Binance BTCUSDT perpetual futures was heavily consumed, and the price briefly touched a low of about \$49,100. For market makers, identifying the transition point from the first phase (information-driven, with the price reflecting a permanent fundamental repricing) to the second phase (liquidation-driven, containing substantial temporary overshoot) is the key to signal classification. After the flood of liquidations subsided, the price rebounded to about \$56,000 over the following roughly 24 hours, recovering about 54% of the decline as calculated by $(56{,}000 - 49{,}000)/(62{,}000 - 49{,}000)$. This V-shaped movement is consistent with the core difference between liquidation-cascade toxicity and informed-trading toxicity in the permanence of price impact. For market makers that accurately identified the "liquidation-dominant" signal in the second phase and engaged in controlled absorption, this constituted a conditional positive-expected-return opportunity (whose positive expectation depends on the real-time accuracy of estimating the upper bound of the cascade's scale, as discussed below).



Figure 19-9. Distribution of liquidation volumes for major liquidation events in 2024 (Data source: CoinGlass Liquidation Data, January–December 2024)

This poses a key signal-identification problem for market makers. In the extreme environment of an erupting liquidation cascade, a market maker must judge within seconds or minutes: is this enormous selling pressure an informed-trading flow reflecting some major bad news, or merely a temporary liquidation flow triggered by the implosion of the leverage system? If it is judged to be the former, the market maker should immediately pull its quotes and stay away from the market, or it will suffer a huge permanent loss. But if it is judged to be the latter, this may instead constitute a positive-expected-value trading opportunity: by actively absorbing these price-insensitive liquidation orders, the market maker can profit in the subsequent price rebound. The asymmetric payoff structure of this decision demands high-precision, real-time signal classification.

A market maker's response strategy therefore becomes highly dependent on the ability to parse real-time data. It closely monitors indicators such as the liquidation data, open-interest changes, and funding rates published by major exchanges, trying to identify the true source of the current order flow. If the data confirm that the selling pressure comes mainly from the liquidation engine, a mature market maker may adopt a "controlled absorption" strategy: it provides liquidity moderately, absorbing part of the liquidation orders, while strictly controlling its own risk exposure through algorithms to avoid exhausting all available capital as the cascade continues to expand. The positive expected value of this strategy is conditional, not unconditional: it holds only if the market maker can make a sufficiently accurate real-time estimate of the cascade's remaining scale; once that estimate fails, the act of absorption itself turns into a major loss. In other words, "controlled absorption" has both success stories and many failures, and one must beware of survivorship bias when describing it as a hallmark capability of top market makers. This strategy faces multiple constraints in practice. First, the total scale and duration of a liquidation cascade are unknowable in advance, and absorbing too early or on too large a scale may risk the price falling a further tens of percent. In the August 5, 2024, event, the time span from the initial decline to the final low exceeded 6 hours, with multiple false rebounds along the way. Second, the exchange's liquidation engine enjoys priority of execution, and a market maker's buy orders usually rank behind the liquidation sell orders in the matching queue, which

means the market maker may not be filled at the expected price. Third, implementing an absorption strategy simultaneously on multiple exchanges may cause large amounts of capital to be frozen on exchanges experiencing extreme conditions, losing the flexibility to make necessary reallocations in other markets. Controlled absorption is therefore by no means simple contrarian buying but requires extremely high technical execution, precise risk-budget management, and the ability to assess the cascade scale in real time; whether it can produce a positive return depends heavily on whether the above conditions are met simultaneously.

19.4.4 MEV-attack toxicity

MEV-attack toxicity is a distinctive risk born of the blockchain environment, especially DEXs. Its root lies in the transparency of transactions on public chains. When a user or market maker submits a transaction, it is not executed immediately but enters a public "mempool," waiting for a miner or validator to package it into the next block. This waiting time provides an attack window for "MEV searchers" who monitor the mempool and perform precise arbitrage on high-value transactions. In their systematic study of front-running behavior in decentralized exchanges and miner extractable value, Daian et al. (2020) were the first to provide an ontological definition and empirical characterization of MEV and priority gas auctions (PGA) [17], laying the foundation for understanding the extractable value of on-chain transaction ordering.

For market makers providing liquidity on on-chain order-book DEXs, the MEV attacks they face take mainly two forms: front-running and sandwich attacks. In a front-running attack, an MEV searcher detects that a market maker has submitted a profitable limit order (for example, a buy order below the current market price) and immediately submits its own buy order, paying a higher gas fee so that its transaction is executed before the market maker's, thereby intercepting the fill opportunity that should have belonged to the market maker. In the more destructive sandwich attack, when a searcher observes a large transaction likely to have a significant price impact, it acts immediately: first, it pays a high gas fee to submit its own buy order, executing ahead of the large transaction to push the price up (front-running); then, after the large transaction is filled at the elevated price, it immediately submits a sell order to sell off the position it bought earlier for a profit (back-running). The market maker is sandwiched in the middle, completing two trades at unfavorable prices, and its value is systematically extracted by the MEV searcher.

The temporal characteristic of this toxicity is highly synchronized with block production; it is a high-frequency, structural toxicity. In every block, an on-chain arbitrage attack may occur. Every quote a market maker posts and every position adjustment it makes are exposed to the monitoring of all MEV searchers, and its strategic intent is publicly interpreted and exploited. This makes active, high-frequency market-making strategies extraordinarily difficult and dangerous to run on DEXs.

The strategies for countering MEV-attack toxicity are also highly technical. The most direct method is to bypass the public mempool and send transactions directly to block builders through private channels, avoiding having transactions "seen" in the public market. In addition, market makers can establish cooperative relationships

with block builders, exchanging payment for priority processing of their transactions and protection from attack. From the perspective of protocol design, some new DEX architectures—such as exchanges that adopt batch auctions or commit-reveal schemes—aggregate orders within a specific time window and then clear them at a uniform price, eliminating the value of transaction ordering and thereby simultaneously weakening both LVR extraction and sandwich attacks at the mechanism level. The research of Canidio and Fritsch (2023) [18] provides theoretical and simulation evidence that a function-maximizing batch-auction AMM can make LP returns slightly higher than the empirical level of Uniswap v3 while sharply compressing arbitrageurs' profits and the space for sandwich attacks. For DEX market makers, choosing a trading platform with an MEV-resistant design is the first line of defense for survival.

19.4.5 Toxicity superposition and synergistic dynamics

This section does not introduce a fifth type of toxicity; it analyzes the cross-cutting superposition dynamics of the preceding four types within the same coordinate system. The previous four sections characterized the typical regions of the four types of toxicity by dominant cause; this section focuses on how they are activated simultaneously and reinforce one another at the market's extreme moments, thereby forming a composite risk that no single type of toxicity alone can explain. A market crash or violent swing is never a linear process triggered by a single factor but a complex-system phenomenon in which multiple risks resonate and positive feedback loops are activated.

We can envision a typical market-crisis scenario. First, an unexpected major piece of bad news (for example, the depegging of a large stablecoin) appears, triggering large-scale selling by informed traders—the manifestation of "informed-trading toxicity." The sharp price decline causes the CEX-DEX spread to widen far beyond normal levels, attracting a swarm of arbitrage bots to pour into the DEX and extract value—the further intensification of "arbitrage-extraction toxicity." Immediately after, the continued price decline triggers the forced liquidation of the first batch of high-leverage long positions, setting off "liquidation-cascade toxicity," as huge market sell orders flood into the market and push the price down further. And on-chain, as panic spreads, trading volume surges, gas fees soar, and the mempool becomes a battleground of fierce competition among MEV searchers, with "MEV-attack toxicity" reaching its peak; any attempt to conduct a rescue or adjust a position on-chain may be sandwiched or front-run.

At this moment, the four types of toxic flow reach their peaks simultaneously. A market maker finds that the total level of toxicity it faces may be far from the simple arithmetic sum of the four toxicity intensities, but instead exhibits a superlinear amplification. We call this proposition the "toxicity-resonance hypothesis." To make this hypothesis distinguishable by data rather than always true, we offer a falsifiable criterion: if, in an event where multiple types of toxicity occur concurrently, the total price impact the market bears does not exceed the sum of the impacts produced by each type of toxicity acting alone (at comparable scale), then "superlinear superposition" does not hold in that case—that is, the resonance hypothesis is supported only when the concurrent impact is systematically greater than the sum of the individual impacts. This criterion

could, in principle, be tested by comparing the price-impact elasticity of a "multi-toxicity concurrent window" and a "single-toxicity-dominant window" at the same order size; this chapter does not undertake such a formal estimation, so it is proposed as a hypothesis awaiting testing.

The core transmission mechanism of this superlinear superposition (when it holds) lies in the positive-feedback path of "depth consumption → impact amplification": the first wave of toxic flow (informed trading) consumes the first layer of order-book depth, so that the subsequent arbitrage flow and liquidation flow face a thinner order book; an order of the same size produces a significantly larger price impact on a thinner order book, and this amplified impact in turn triggers more liquidations and more space for MEV extraction. Each wave of toxic flow creates more extreme impact conditions for the next. Every buy order a market maker posts may be attacked simultaneously from three directions—by the informed, by arbitrageurs, and by the liquidation engine—while an on-chain LP must also endure the fourth blow of MEV. In this extreme environment, the expected profit of market making quickly turns deeply negative, and any rational market-making algorithm points to the same conclusion: immediately withdraw from the market, or widen the spread to a level at which it is almost impossible to fill. This is not because market makers are "irresponsible" or "lacking in commitment," but is the inevitable choice driven by their economic rationality. When the cost of providing liquidity soars far beyond any possible return, the supply of liquidity stops.

As a contrast to the resonance scenario above, consider a "mild pullback in which toxicity does not resonate": when a price pullback is driven by a relatively single, mild factor (such as a neutral-to-bearish macroeconomic data release), and the first layer of order-book depth, once consumed, can be replenished in time by off-book market makers' supplementary quotes, then subsequent orders do not face a continuously thinning order book, no depth-consuming positive feedback forms among the various types of toxicity, and the total price impact roughly equals the linear superposition of the individual factors' impacts. The resonance hypothesis predicts that these two scenarios should differ systematically in the ratio of "concurrent impact to the sum of individual impacts"—in the mild-pullback scenario the ratio is close to or below 1, whereas in a genuine resonance scenario the ratio is significantly greater than 1. The significance of introducing this contrast scenario is to make "toxicity resonance" a conditionally holding claim distinguishable by data, rather than an ex post narrative that can justify itself however the market behaves.

This is the microlevel explanation of the core question of "why liquidity contracts sharply when market pressure is greatest." Liquidity is not an exogenous resource that can be supplied without limit but a commodity endogenously produced by market makers after precisely calculating costs and returns. When the supply cost rises sharply because of the resonance of toxic flow, the supply activity quickly stops. The four types of toxic flow do not superimpose in the same direction at all phases. In the price-recovery phase after a liquidation cascade ends, the direction of arbitrage-extraction toxicity may reverse: arbitrageurs now relay price information from the underpriced market toward fair value, objectively aiding the price's return. This phase difference means that the superposition effect of toxic flow is concentrated mainly in the initial phase of a crisis's eruption, while in the recovery phase a partial hedge may emerge.

Understanding the toxic-flow spectrum and the dynamics of its superposition and hedging is not only a survival guide for market makers but also the key to understanding the fragility of the entire market ecosystem. It foreshadows the theme we will examine in depth in the next chapter: the endogeneity of liquidity and the roots of its fragility in a crisis.

19.5 Market making on CEX CLOBs

In perpetual-futures markets, the CLOB of centralized exchanges has always occupied a core position. Although the wave of decentralized finance has brought various innovative paradigms such as on-chain order books, in terms of the absolute scale and depth of liquidity, the CEX CLOB still dominates to this day. According to estimates in CoinGlass's industry annual report, the average daily volume of the CEX derivatives market in 2025 was on the order of $260 billion (institutional data source, sample period 2025) [19], of which perpetual futures accounted for the overwhelming majority. The liquidity of this market does not form naturally but is systematically built and maintained by highly specialized market-making institutions. Leveraging their enormous advantages in technology, capital, and information, these institutions have built the core skeleton of perpetual-futures market liquidity, while also forming a complex game structure around risk and profit. This section analyzes in depth the ecosystem, strategic framework, and risk management of CEX CLOB market making, as well as the symbiotic relationship between market makers and exchanges, and ultimately reveals, through comparison, the distinctiveness and extreme difficulty of the perpetual-futures market-making business.

A disclosure about evidence grade is in order at the outset: the specific figures involved—market-maker concentration, order-book depth share, and fee rebates—come mainly from the in-house metrics of data vendors, market makers, and exchanges (industry research reports and official announcements), and have not been independently verified by academia or regulators. In interpreting these figures, readers should treat them as order-of-magnitude references provided by interested parties rather than as precise measurements confirmed by peer review; where such figures are cited below, the nature of their source is noted at each point, and independent academic or regulatory cross-sources are added where possible.

19.5.1 The market-maker ecosystem

The market-making arena for CEX perpetual futures is a classic oligopolistic market. The participants are primarily a small number of well-capitalized, technologically leading professional institutions, which provide the ecosystem with its most fundamental and important element: liquidity. This ecosystem can be roughly divided into three layers.

At the top of the pyramid are the global Tier 1 crypto market makers, such as Wintermute, Jump Crypto, Amber Group, GSR Markets, B2C2, and Cumberland/DRW. These institutions typically originate from traditional high-frequency trading firms and possess deep quantitative-trading backgrounds and powerful technical infrastructure.

Their business spans the major CEXs worldwide, and they provide liquidity for hundreds of crypto assets simultaneously. Their advantages are all-encompassing: ample capital allows them to withstand enormous inventory risk, cutting-edge low-latency systems let them capture arbitrage opportunities at the microsecond level, and large data teams support complex pricing models and toxic-flow-identification algorithms. These top institutions are the core providers of market liquidity. According to industry reports from firms such as Kaiko and Wintermute, the top three to five market makers together contribute on the order of 70% of the order-book depth in mainstream perpetual-futures markets such as those for Bitcoin and Ethereum. The evidence grade of this figure must be treated with caution: the relevant estimate comes from in-house metrics, its calculation method is not public, and Wintermute is itself a top market maker, so an interested party estimating the share of the very group to which it belongs entails an inherent conflict of interest; this figure is therefore used not as a single precise number bearing a systematic claim but only as an order-of-magnitude reference. As cross-corroboration, Aramonte et al. (2021), in a study for the Bank for International Settlements (BIS), also point out that liquidity provision in both decentralized and centralized crypto markets shows a pronounced "decentralization illusion" in which supply is highly concentrated among a few participants [20]; this independent observation from an authoritative institution supports, in direction, the judgment that "market makers are highly concentrated," although it does not provide a metric fully comparable to the "70% order of magnitude" above. Whatever the specific number, this high degree of concentration brings, on one hand, economies of scale that let the market enjoy narrower bid-ask spreads and greater depth; on the other hand, it plants the seeds of systemic risk: if these top institutions simultaneously withdraw because of an extreme market event, a technical failure, or regulatory pressure, market liquidity will face a sharp collapse.

The business model of top crypto market makers is far more diversified than "pure market making." They typically play multiple roles at once—liquidity provider, cross-market arbitrageur, and over-the-counter (OTC) counterparty—and market-making income is only part of their composite profit matrix. This business fusion has important analytical implications: it may explain why certain market makers are willing to keep quoting even when their spread revenue on a particular pair is negative—some top institutions may obtain compensating revenue from cross-market arbitrage opportunities associated with their market-making inventory. The scale of this compensation is hard to separate out in public data, so the explanation above should be regarded as an explanatory hypothesis that helps make sense of market makers' quoting and retreat behavior, rather than an established fact.

Below Tier 1 are numerous regional Tier 2 market makers or those focused on specific asset classes. They may have a competitive advantage on only one or two exchanges, or focus on market-making the tokens of a particular "ecosystem." Their strategies are more flexible, and they may offer more competitive quotes than Tier 1 institutions in certain niche markets. However, in terms of capital scale and technological coverage, they lag noticeably behind the top institutions.

The final layer consists of the proprietary market-making desks of exchanges themselves. Some CEXs run in-house market-making teams to provide basic liquidity for certain

pairs on their platforms (especially newly listed or thinly traded pairs). Although this practice can ensure the basic functioning of the market to some degree, it also raises serious concerns about conflicts of interest. As the maker and enforcer of market rules, an exchange that simultaneously participates directly in market trading gains an unfair advantage over external market makers in information access (such as seeing all users' order flow) and rule enforcement (such as the liquidation mechanism). From a regulatory perspective, the severity of this problem is rising: in traditional finance, exchange proprietary market making is subject to strict information-barrier requirements and independent audit oversight, whereas the widespread absence of these institutional protections in crypto markets is not only a governance challenge but may also translate into direct legal risk as global crypto-regulatory frameworks tighten. We will explore this governance challenge in greater depth in Chapter 21.

Overall, the market-making ecosystem for CEX perpetual futures exhibits a clear pyramid structure: the three to five Tier 1 institutions at the top control on the order of 70% of the order-book depth on mainstream pairs (an institutional estimate, with the evidence grade as noted above); the dozens of Tier 2 market makers in the middle seek differentiated survival space in niche markets and emerging assets; and at the bottom, exchanges' own proprietary market-making desks fill the gaps in the thinnest corners of liquidity. This highly concentrated pattern is the product of both a technology race and capital barriers: in an industry whose profit space is continuously compressed, only top participants with economies of scale can sustain a positive profit equation. Yet this concentration also means that the entire market's liquidity supply depends heavily on the health and operational continuity of a few institutions—a structural fragility discussed further in Section 19.7.2.

19.5.2 The quoting strategy framework

A professional market maker's operation on a CEX CLOB is not simple quoting but a precise, dynamically adjusted strategic framework. The core of this framework can be reduced to the real-time optimization of three key parameters: the width, depth, and skew of quotes. Together, these three parameters determine a market maker's risk exposure and profit potential, and they are continuously adjusted dynamically in a fast-changing market environment. Their mathematical skeleton is precisely the optimal market-making framework of Avellaneda and Stoikov (2008) cited earlier in Section 19.3.1 [13]: a market maker sets bid and ask quotes symmetrically around the reservation price and adjusts the width and skew of quotes in real time according to the inventory level, price variance, and degree of risk aversion ($\delta \propto \gamma\sigma^2 Q$); the three-parameter optimization discussed in this section can be viewed as an engineering implementation of that framework in the perpetual-futures context.

Width (the bid-ask spread) is a market maker's core source of profit as well as its foremost tool for risk compensation. Setting the spread is, in essence, a market maker's judgment and pricing of the current market's "toxicity" level. When the market is calm, the market maker faces low adverse-selection risk and tends to narrow the spread to compete for order flow. Conversely, when the market swings sharply or toxic flow increases, the market maker rapidly widens the spread, raising the cost for traders

to obtain liquidity so as to compensate for the losses it may suffer from information asymmetry. Observing the dynamic changes in a pair's bid-ask spread is therefore an effective indicator of its current liquidity quality and market risk level.

Depth reflects the total amount of inventory risk a market maker is willing to bear at a given moment. Providing greater market depth means the market maker is willing to absorb larger orders when the price moves slightly, which usually attracts large traders; but greater depth also means that, once the price moves unfavorably, the market maker accumulates more risk exposure. The setting of depth is therefore a trade-off between risk and return. The following figure shows a real-time depth profile of the BTC perpetual-futures order book—that is, how cumulative depth on the bid and ask sides varies with distance from the mid-price: cumulative depth climbs rapidly as price levels extend outward, and the distributions on the two sides are not perfectly symmetric (see the figure for specific values). This asymmetry reflects market makers' implicit view of market direction and their current inventory state.



Figure 19-10. Depth profile of the BTC perpetual-futures order book: cumulative depth distribution on the bid and ask sides (Data source: real-time order-book data for Binance BTCUSDT perpetual futures, March 2025; this figure is an instantaneous snapshot profile of the inner order book, covering roughly ±9 bps around the mid-price)

In a high-volatility or high-uncertainty environment, a market maker "withdraws depth"—reducing the resting order size at each price level to control potential losses. This sharp withdrawal of depth means that a market order of the same size will cause a price impact several times greater than in normal times, further intensifying market volatility. Depth withdrawal is the rational product of a market maker's profit equation under extreme conditions, but its collective effect—liquidity vanishing precisely when it is most needed—constitutes a core fragility of market microstructure.

Skew describes the asymmetry of a market maker's quotes on the bid and ask sides. A perfectly neutral market maker would post quotes symmetrically around the market mid-price. In reality, however, a market maker's quotes are almost always skewed. Skew serves two main purposes: inventory management and directional betting. When a market maker has accumulated net long (or net short) inventory through successive fills, it skews its quotes to incentivize the market to trade in the opposite direction, thereby returning inventory to neutral. For example, if it holds long inventory, the market maker lowers its bid (dampening buy-side fills so that it accumulates less) and

lowers its ask (making its sell orders fill more readily so that it offloads more), attracting buyers to "take away" its inventory. In addition, some more aggressive "statistical market-making" strategies actively skew quotes based on predictions of short-term price movements, trying to capture a small directional gain while providing liquidity. For example, if a model predicts a high probability of a short-term price rise, the market maker will quote more aggressively on the bid side and relatively conservatively on the ask side.

The dynamic adjustment of these three core parameters is driven by a series of complex triggers. These triggers include, but are not limited to: sharp changes in market volatility, jumps in toxic flow detected by indicators such as VPIN, the market maker's own inventory level exceeding a preset threshold, the approach of a funding-settlement window, abnormal spreads relative to other exchanges, and large liquidations about to be triggered as detected by on-chain data monitoring. The distinctive mechanisms of perpetual futures, such as the funding rate, add extra dimensions to this adjustment framework. For example, when a high positive funding rate is expected, the market maker has an incentive to reduce its net long position, or even build a net short position, before settlement, in order to collect the funding rate as extra income. This kind of strategic adjustment based on institutional rules is a complex game that does not exist in traditional market making.

19.5.3 Cross-exchange market making and position management

For top professional market makers, the scope of business is by no means confined to a single exchange. To maximize capital efficiency and capture broader order flow, cross-exchange market making is an inevitable choice. Yet this operation, while magnifying opportunities, also brings higher-dimensional complexity and challenges, especially in position management. Market makers must precisely manage risk and return across multiple platforms simultaneously.

First, multi-exchange market making faces a severe "margin fragmentation" problem, a concept in line with the capital-efficiency constraints discussed in Chapter 17. A market maker must deposit sufficient margin on every exchange where it is active to support its market-making activity and potential risk exposure. This means that an institution making markets on N exchanges has a total capital requirement far greater than N times that of making markets on a single exchange, because the margin pools of the various exchanges are mutually isolated and cannot be shared in real time. This fragmentation of capital severely tests a market maker's capital-management capabilities.

The more central challenge lies in the cross-exchange hedging of inventory. In theory, if a market maker accumulates a net long of 100 BTC on exchange A through fills while accumulating a net short of 100 BTC on exchange B, its total position appears neutral. In practice, however, this is a dangerous "pseudo-neutrality," because these two positions are each exposed to the counterparty risk, platform risk, and liquidation risk of two different exchanges. Once an extreme move on exchange A causes its long position to be forcibly liquidated while the price on exchange B does not move in sync, the market maker instantly shifts from a seemingly balanced state to bearing enormous one-sided risk. A market maker must therefore build a global risk-management system

that monitors and aggregates net positions across all exchanges in real time and actively hedges, rather than relying merely on the "natural" offset of positions across different exchanges.

Cross-exchange market making also faces a risk dimension often overlooked in theoretical analysis but central in practice: counterparty risk. The margin a market maker deposits at each CEX essentially constitutes an unsecured credit exposure to that exchange. The collapse of FTX in November 2022 already proved, in the most brutal way, that even a top-three global exchange can go bankrupt within days, leaving a market maker unable to recover all its deposited funds. For a market maker operating on five to ten exchanges simultaneously, counterparty-risk management (including exposure caps for a single exchange, a dynamic capital-allocation strategy based on exchanges' financial-health indicators, and an early-warning capital-withdrawal mechanism triggered by abnormal withdrawal delays) constitutes an indispensable part of its risk-management system. A market maker must continuously seek a balance between "concentrating more capital on the most liquid exchanges to maximize capital efficiency" and "dispersing capital across multiple exchanges to reduce single-counterparty risk."

The active-hedging process itself is also fraught with risk. When a market maker's inventory on one exchange exceeds a threshold, it must execute an offsetting trade on another highly liquid exchange (or in the spot market) to balance risk. This process involves a critical "hedging latency window." Under normal market conditions, the round-trip cross-exchange, cross-region hedging latency (even for top market makers that have already deployed servers nearby) is usually on the order of 10–50 milliseconds, and this latency is itself manageable. The 10–50 milliseconds here refers to the end-to-end round-trip latency across exchanges and regions, not the internal latency of co-location in an exchange's data center—the latter is usually at the microsecond to sub-millisecond level. However, in extreme conditions, cross-exchange hedging latency may expand sharply to the seconds level because of the target exchange's application programming interface (API) rate limiting and system congestion. More critically, in extreme conditions the true bottleneck for hedging failure is often not network latency itself but the simultaneous drying-up of liquidity on the target exchange: even if the hedging order arrives instantly, it may incur slippage costs far beyond expectations because the counterparty market's order book has likewise thinned. The value of low-latency technology therefore lies not only in compressing the hedging window under normal conditions but, more importantly, in completing the hedge ahead of competitors in the first few seconds of an extreme move, when the counterparty market's depth has not yet been exhausted by other market makers' hedging demand.

In practice, the boundary between cross-exchange market making and the cross-exchange arbitrage analyzed in Chapter 16 is not clear. Many top institutions in fact play the dual roles of market maker and arbitrageur simultaneously. Their strategy can be described as "hedging the inventory generated by passive market making on one exchange through active trading on another exchange where the price is more favorable." This hedging trade, from the perspective of inventory management, is part of "market making," but from the perspective of exploiting the cross-exchange spread, it also constitutes a typical cross-exchange arbitrage. This fusion of market making and arbitrage allows large institutions to build a more complex profit matrix, and their

integrated return and risk-management capabilities far exceed those of participants able to execute only a single strategy.

19.5.4 The symbiotic relationship between market makers and exchanges

Market makers and centralized exchanges share a deep and complex symbiotic relationship. Exchanges need market makers to provide liquidity, which is the core element for attracting and retaining traders; and market makers need the exchange platform to implement their strategies and earn profits. This interdependence gives rise to a complex set of mechanisms built around incentives and games, the most influential of which are the institutional arrangements exchanges provide to market makers, including fee rebates, API priority, information advantages, and margin benefits.

Among the various incentives, market-maker fee rebates have the broadest applicability and the most direct impact on the profit model. In most exchanges' fee structures, a taker who actively takes liquidity pays a fee, while a maker who provides liquidity pays a lower fee or even receives a negative fee, i.e., a rebate. This means that each time a market maker successfully provides liquidity, it not only pays nothing but also receives income from the exchange. This rebate is an important component of a market maker's profit, especially on mainstream pairs with extremely narrow spreads, where rebate income may even exceed the spread itself. Exchanges design tiered rebate structures to incentivize market makers to provide more liquidity. For example, Binance's USD(S)-margined futures market maker program rates market makers according to indicators such as their weekly market-making volume share and quote quality; the higher the tier, the higher the rebate ratio the market maker enjoys. According to the fee arrangements of Binance's official market maker program, the maker rebate at the best tier is usually in the range of -0.5 to -2.5 bps (i.e., -0.005% to -0.025%), and during specific limited-time promotions the maker fee applicable to top market makers can go even lower (in some promotional periods the maker fee even drops to 0.0000% or offers a higher rebate) [21]. Even at the regular-tier rebate level, for top market makers with daily volume on the order of billions of dollars, this still constitutes an extremely substantial source of income. The rebate figures above come from the exchange's own announcements, and the specific values for promotional and regular tiers change as the exchange adjusts its policies, so readers should treat them as order-of-magnitude references.

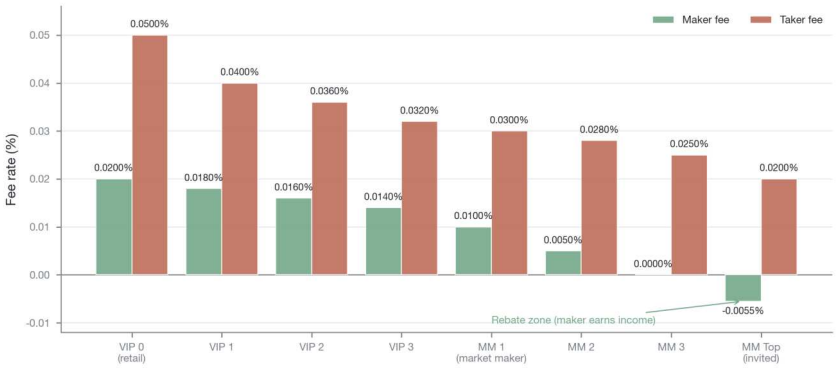


Figure 19-11. Binance USD(S)-margined futures maker/taker fee structure (Data source: Binance official market maker program announcement, March 2025; the figure shows the regular best tier, while during limited-time promotions the maker rebate can go deeper, on the order of -2.5 bps, which is not shown in the figure)

The figure above shows the tiered fee structure of Binance's market maker program: from the basic VIP0 tier to the highest MM top (invitation-only) tier, the maker fee decreases step by step to a negative value (i.e., a rebate), and the taker fee also declines as the tier rises. The economic meaning of this tiered design is that the exchange reserves the most favorable rebates for the market makers with the highest volume and quote quality, forming a positive-incentive "flywheel effect": higher rebates attract deeper liquidity, deeper liquidity attracts more traders, and greater trading volume in turn further consolidates the market maker's rating and rebate level. However, this also intensifies the competitive asymmetry between large market makers and small participants.

Broadening the view to the whole industry, the following figure compares the fee structures of six major perpetual-futures trading platforms (including CEXs and two on-chain DEXs). At the base tier (VIP0), fee differences across platforms are relatively small, with maker fees generally around 0.02% (i.e., 2 bps); but at the best tier of the market maker programs, the differences widen substantially: most mainstream platforms offer negative-fee (rebate) incentives to top market makers, with the most favorable maker rebate reaching the order of -2.5 bps (see the figure for each platform's base and best-tier fees). This fee competition directly affects market makers' platform selection and capital-allocation strategies.

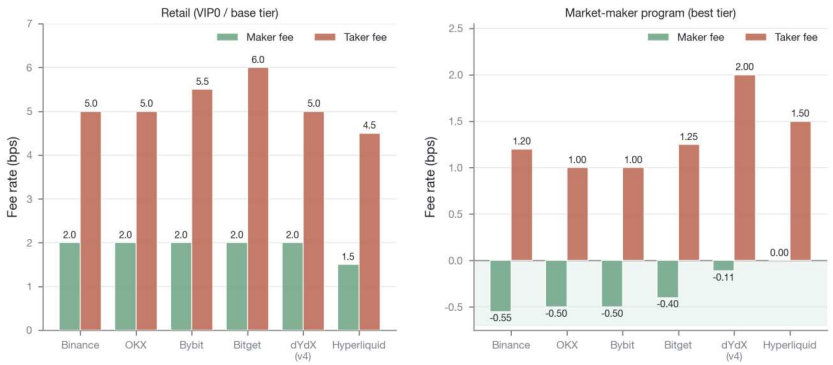


Figure 19-12. Comparison of fee structures across major perpetual-futures trading platforms (including CEXs and on-chain DEXs): base tier vs. market-maker best tier (Data source: each exchange's official fee pages and market maker program announcements, March 2025; the six subjects in the figure include two on-chain DEXs, dYdX and Hyperliquid; the market-maker best maker rate shown in the figure stops at -0.55 bps, while the -2.5 bps order of magnitude mentioned in the text is the limited-time-promotion end and is not shown in the figure)

Beyond direct fee incentives, priority technical access is likewise an important source of competitive advantage. Exchanges provide top market makers with priority service on the API. This may include faster network connections, higher API request-rate limits, and lower data-push latency. In the world of high-frequency trading, where every fraction of a second counts, even a lead of a few milliseconds can determine the life or death of a strategy. Obtaining API priority means a market maker can receive market data faster than ordinary traders and submit and cancel orders faster, thereby managing risk and capturing opportunities more effectively.

Information advantage is a more hidden but potentially more valuable "privilege." Some exchanges may provide their core market makers with finer-grained order-flow data, such as anonymized statistical information about large orders or the behavior of specific types of traders. This information can help market makers judge market sentiment more accurately and identify potentially toxic order flow, thereby optimizing their pricing and risk models. However, this information advantage is also precisely the most controversial point in the symbiotic relationship between exchanges and market makers, because it not only touches the bottom line of market fairness but may also constitute a legal risk. In several jurisdictions, providing nonpublic order-flow data to specific market participants may be deemed market abuse. As the EU's Markets in Crypto-Assets (MiCA) Regulation and the crypto-asset regulatory frameworks of other major economies are gradually implemented, the compliance-scrutiny risk facing such information-sharing arrangements is rising systematically.

Finally, margin benefits are also a common incentive. Exchanges may offer reputable, well-capitalized market makers lower margin requirements or higher leverage caps. This allows market makers to operate at higher capital efficiency, supporting larger-scale market-making activity with the same amount of capital.

However, this symbiotic relationship is not always harmonious; it is full of ongoing games and tension. An exchange's goal is to maximize the platform's trading volume

and liquidity, so it always wants market makers to provide the narrowest possible spreads and the deepest possible market depth. A market maker's goal is to maximize its own profit, so it must ensure that its risk is controllable and its business profitable while providing liquidity. There is an inherent contradiction between these two. An exchange's rebate and incentive policies are, in essence, a search for an equilibrium in this game: attracting market makers to provide sufficient liquidity without letting the exchange's own operating costs run too high. When the market environment deteriorates and market makers' risk rises sharply, they tend to widen spreads and withdraw depth to protect themselves—which is precisely what the exchange least wants to see. This fundamental conflict of interest is the key to understanding why liquidity contracts sharply in moments of crisis.

19.5.5 How perpetual-futures market making differs from traditional market making

To understand just how demanding perpetual-futures market making on a CEX CLOB is, the best approach is to compare it directly with more traditional spot market making and delivery-futures market making. Although all three follow the basic logic of "earning the spread by bearing risk," the distinctive institutional design of perpetual futures makes it fundamentally different from the other two in risk dimensions and operational complexity. We can dissect the distinctiveness of this business along six core dimensions. The following table first summarizes the differences among the three types of market-making businesses across these six dimensions, and each is then elaborated in turn.

Dimension of comparison	Spot market making	Delivery-futures market making	Perpetual-futures market making
Inventory time horizon	Can be held indefinitely, but unleveraged with linear risk	Has an expiration date; the basis necessarily converges and inventory is cleared to zero at maturity	No expiration date; inventory risk is indefinite and must be actively hedged
Certainty of holding cost	Fiat funding interest, slow and predictable	Day-by-day basis decay, a largely determined path	Dominated by the randomly fluctuating funding rate; direction can reverse within hours
Liquidation risk	No leverage, no liquidation risk	Leveraged, but with an expiration constraint plus institutional dominance, relatively controllable	High leverage plus no expiration plus many retail traders; liquidation-cascade risk is extremely high
Source of extreme toxic flow	Primarily traditional informed trading	Primarily traditional informed trading	Beyond informed trading, uniquely subject to liquidation cascades—a positive-feedback toxicity
Operational continuity	Usually has market-closure windows for calibration/maintenance	Usually has market-closure windows for calibration/maintenance	24/7 without interruption; systems and teams must operate around the clock
Complexity of the information environment	Stricter disclosure regulation in mature jurisdictions, relatively transparent	Stricter disclosure regulation in mature jurisdictions, relatively transparent	Inconsistent disclosure standards, higher manipulation/insider risk

Table 19-3. A six-dimensional comparison of spot, delivery-futures, and perpetual-futures market making (Data source: compiled by the author)

First, the inventory time horizon. The inventory a spot market maker holds can, in theory, be held indefinitely, but because there is no leverage, its risk exposure is linear. A delivery-futures market maker enjoys an important institutional safeguard: the expiration date. However far the inventory deviates along the way, the basis necessarily converges to zero at expiration, all positions are forcibly delivered or settled, and risk is confined to a finite time window. A perpetual-futures market maker, by contrast, faces the most severe combination: it must, like a spot market maker, confront indefinite inventory-holding risk, while also, like a futures market maker, bear the risk-amplifying effect of high leverage. The absence of an expiration date as a natural settlement mechanism means that inventory risk can only be continuously managed through active hedging.

Second, the certainty of holding cost. The holding cost of spot market making is mainly fiat funding interest, which changes slowly and predictably. The holding cost of delivery-futures market making manifests as the day-by-day decay of the basis, whose path is also largely determined. However, the inventory holding cost of a perpetual-futures

market maker is dominated by a highly random and volatile variable, the funding rate (whose randomness and magnitude were analyzed in detail in Section 19.3). This random "implicit tax" or "implicit subsidy" brings enormous uncertainty to a market maker's risk pricing and inventory management.

Third, liquidation risk. Spot trading, having no leverage, has no liquidation risk. Delivery futures, though leveraged, have relatively controllable large-scale liquidation risk because of the expiration-date constraint and the participation of relatively more mature institutions. The perpetual-futures market, with its high leverage, absence of an expiration date, and large numbers of retail traders with weaker risk-management ability, together constitutes a forced-liquidation environment extremely prone to triggering chain reactions, pushing liquidation risk to the extreme.

Fourth, the source of extreme toxic flow. The toxic flow faced by spot and delivery-futures markets comes mainly from traditional informed trading. The perpetual-futures market, on top of this, adds a uniquely destructive form of "extreme toxic flow": the liquidation cascade. This positive feedback loop, triggered by the market's own decline, can generate overwhelming one-directional order flow several times the normal level in a short time, posing a significant threat to the safety of a market maker's positions.

Fifth, operational continuity. Most traditional financial markets have closing and holiday hours, which provide market makers with necessary adjustment windows for recalibrating models, managing risk, and maintaining systems. The 24/7 uninterrupted trading feature of crypto markets requires that perpetual-futures market makers' systems and teams also operate around the clock, which imposes significant additional constraints on operating costs, technical reliability, and human-resource allocation.

Sixth, the complexity of the information environment. Spot and futures markets, especially in mature jurisdictions, are subject to relatively strict information-disclosure regulation, and their information environment is relatively transparent. Crypto markets, especially the perpetual-futures space, have inconsistent disclosure standards and higher risks of market manipulation and insider trading, which intensifies the adverse-selection problem market makers face.

In summary, if we rate the difficulty of these three types of market-making businesses, spot market making is at a higher difficulty level, delivery-futures market making is at an expert level, and perpetual-futures market making is at an extreme difficulty level. It pushes the risk of the market-making business to the limit along virtually every dimension. In theory, because of the additional dimension of institutional cost, the bid-ask spread of perpetual futures (measured in basis points) tends to be wider than that of spot and delivery futures under comparable liquidity conditions; this extra spread is precisely the risk premium the market pays for market makers bearing these three superimposed (adverse-selection, inventory, and institutional) ultra-high costs. Rigorously verifying the judgment that "perpetual spreads are systematically wider than spot/delivery spreads" would require paired typical-spread data for the same asset across the three markets, which this chapter does not provide; the judgment is therefore proposed as a directional theoretical prediction rather than an established conclusion.

19.6 Liquidity provision on DEXs

From centralized to decentralized exchanges, the paradigm of liquidity provision for perpetual futures shifts fundamentally. This book takes order-book-based perpetual futures as its core object of study, and the corresponding forms in the DEX domain are the on-chain central limit order book paradigm and the hybrid market-making vault paradigm derived from it. The market dynamics of decentralized exchange adoption itself have been studied specifically (Capponi & Jia, 2021 [22]); this section instead focuses on the different answers these two perpetual-market-making paradigms give to three core questions—"who bears the risk," "how liquidity is priced," and "how toxic order flow is handled"—and constructs a unified comparative framework.

19.6.1 The on-chain CLOB paradigm

The on-chain CLOB paradigm takes a direct path: replicating the mature central limit order book model of CEXs onto the blockchain. Protocols represented by dYdX v4 and Hyperliquid achieve high-performance on-chain order matching and settlement by building dedicated application chains, making it possible for professional market makers to operate on-chain.

On dYdX v4's Cosmos application chain or Hyperliquid's proprietary high-performance HyperBFT chain, market makers can, just as on Binance or Bybit, connect via API, deploy complex market-making strategies, submit and cancel limit orders in real time, and actively manage their inventory risk. This model returns the responsibility for liquidity provision to professionals, who provide dense liquidity on both sides of the order book and earn the bid-ask spread by virtue of their capital, technology, and strategy advantages. This fundamentally changes the risk-bearer: risk is no longer borne collectively by a passive group of LPs but by individual professional market makers pursuing excess returns.

However, deploying a CLOB on-chain is not a simple technical migration; the inherent properties of the blockchain bring distinctive challenges and opportunities. First, transparency has a dual effect. The full state of the on-chain order book is publicly visible to all, which significantly increases the market's information symmetry but also makes a market maker's strategy easier to "reverse-engineer" and raises the risk of MEV attacks. For example, an MEV bot can monitor a market maker's order-cancellation behavior and use the information advantage to front-run and extract value. Second, latency remains an issue. Although Hyperliquid's block time has been compressed to the 200-millisecond level (and its optimistic confirmation mechanism may make the user-experienced latency lower than the block finalization time), this still lags significantly behind a CEX matching engine's internal processing latency (microseconds) and external/end-to-end price-update latency (milliseconds). Longer latency means a market maker's quotes are exposed to adverse-selection risk for a longer time, placing higher demands on its risk-management capabilities.

Despite the above challenges, the on-chain CLOB paradigm still demonstrates strong competitiveness thanks to its friendliness to professional market makers. Hyperliquid is a case in point: by virtue of its efficient application chain and market-maker incen-

tives, it rose rapidly in 2024. Here it is necessary to clarify several share metrics that are often confused. First, the single-protocol metric: during the peak of the fourth quarter of 2024, the single protocol Hyperliquid at one point held about 55% of the DEX perpetual-futures market, and over the same period its daily volume and open interest reached roughly 10% and 15–20%, respectively, of the Binance perpetual-futures market. During this period, Hyperliquid's points-incentive campaign may have positively distorted its volume figures to some degree, and the organic-volume share after excluding points incentives may be lower than the 55% peak figure above. Second, the paradigm-aggregate metric: as of March 2025, all protocols using the CLOB paradigm (including Hyperliquid, dYdX, and others) together accounted for about 93% of DEX perpetual-futures volume. Third, the absolute-volume metric: over the same period, Hyperliquid led DEX perpetual protocols by a wide margin with weekly volume of about $53 billion. The "55%" and "93%" above are not measurements of the same object at the same point in time—the former is the share of a single protocol at its Q4 2024 peak, distorted by points incentives, while the latter is the aggregate share of the CLOB paradigm in March 2025; the two differ in both metric and time point and cannot be directly compared or substituted for each other. Combining these three metrics yields a robust judgment: in the specific domain of perpetual futures, the CLOB paradigm dominated by professional market makers has taken a leading position in the competition on efficiency; but this judgment applies mainly to high-liquidity mainstream-asset scenarios, while in emerging markets or low-barrier scenarios, other paradigms such as market-making vaults still retain their niche (see Section 19.6.3).

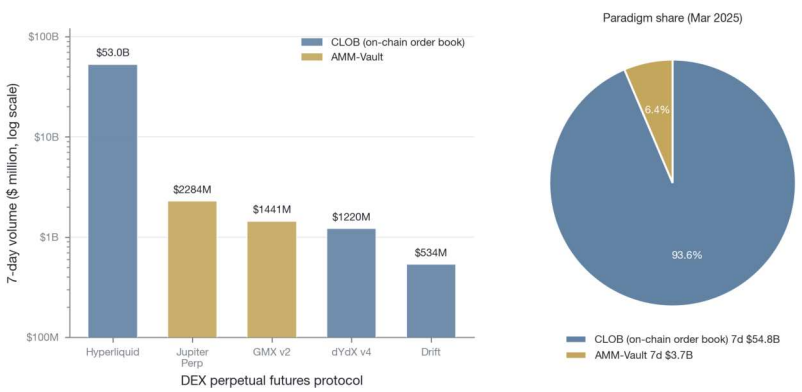


Figure 19-13. Comparison of DEX perpetual-futures protocol trading volumes and paradigm shares (Data source: DeFiLlama Derivatives Protocol Data, March 2025)

19.6.2 The market-making vault paradigm

As a point of comparison with the on-chain CLOB, the market-making vault paradigm represents a different path for liquidity provision. Represented by Hyperliquid's HLP vault, the core feature of this paradigm is "the protocol is the market maker": ordinary users deposit funds into a vault managed by the protocol, and the funds no longer passively serve as counterparty (as in GMX's GLP model) but are used by the protocol's algorithm to execute active market-making strategies on the CLOB, including dynamically adjusting quotes, managing positions, and widening spreads as volatility

intensifies. This grants the LP community the risk-management initiative that previously only professional market makers possessed.

The core trade-off of this model is that it offers a lower barrier to entry at the cost of limited LP autonomy and a reliance on trust in the algorithm and governance. An LP's returns depend heavily on the quality of the protocol's market-making algorithm: once the algorithm has a flaw or is exploited by a large holder, the vault may suffer major losses. For example, in March 2025, the HLP vault recorded a significant drawdown due to a squeeze-style operation via a single large position (the so-called JELLY incident), highlighting the fragility of algorithmic market making in the face of targeted manipulation. In addition, the ownership of the decision-making authority over the market-making strategy's key parameters (spread, depth, and risk-control thresholds) also constitutes a new governance challenge.

19.6.3 Comparing the two paradigms

Placing the two DEX liquidity-provision paradigms above within a unified framework, we can clearly see their different trade-offs across core dimensions. This is not a simple judgment of which is superior but a design choice about how decentralized markets construct their liquidity foundation. The following table first summarizes the trade-offs of the two paradigms across five core dimensions, and the discussion then unfolds.

Dimension of comparison	On-chain CLOB paradigm	Market-making vault paradigm
Barrier to entry	High (requires technology, capital, and strategy)	Low (ordinary users need only deposit funds)
LP autonomy	Fully active (independent quoting and risk control)	Limited (executed on their behalf by the protocol's algorithm)
Toxicity exposure	Borne and managed directly by individual professional market makers	Centrally managed by the protocol's algorithm; individual LPs do not respond directly
Capital efficiency	Highest	Medium, dependent on algorithm quality
Trust dependency	Mainly on one's own strategy and the exchange's matching engine	Additionally on the protocol's algorithm and governance (algorithm risk plus governance risk)

Table 19-4. A comparison of the on-chain CLOB paradigm and the market-making vault paradigm across core dimensions (Data source: compiled by the author)

From the core trade-off dimensions—"barrier to entry," "LP autonomy," and "toxicity exposure"—the two paradigms occupy different niches. The on-chain CLOB paradigm offers professional market makers with technology, capital, and strategy full autonomy and the highest capital efficiency, but its barrier to entry is also the highest. The market-making vault paradigm is a compromise: it provides LPs with a degree of risk-management capability through a protocol-managed algorithm, and its barrier to entry is lower, but LP autonomy is limited, and it introduces a trust dependency on the algorithm and governance.

This trade-off is also directly reflected in the cost structures of the different paradigms. For professional market makers on an on-chain CLOB, the cost structure is highly similar to that of CEX market makers: adverse-selection cost and inventory-risk cost dominate, MEV-attack toxicity constitutes an additional on-chain-specific cost item, and they must also bear costly technical infrastructure (low-latency nodes, strategy engines) and ongoing operating costs. This cost structure means that only professional institutions with ample capital and technical capability can profit sustainably. For the market-making vault paradigm, by contrast, the cost structure LPs face is more balanced: adverse-selection cost and inventory risk are centrally managed by the protocol's algorithm, and individual LPs need not respond directly; but the core risk LPs bear shifts to algorithm risk (systematic losses caused by flawed strategy design or improper parameter settings) and governance risk (the opacity of the decision-making authority over strategy adjustments). An LP's returns therefore depend heavily on a variable it cannot directly control: the quality of the algorithm.

Ultimately, the DEX perpetual-futures market will settle into an ecosystem in which the two paradigms—on-chain CLOB and market-making vault—coexist and evolve dynamically. High-liquidity mainstream-asset markets will be dominated by the most capital-efficient on-chain CLOB professional market makers, while emerging markets may rely more on market-making vaults to bootstrap initial liquidity. This still-ongoing paradigm contest will determine the trading experience and costs of future DEXs and leave a lasting mark on the form of on-chain financial infrastructure.

19.7 Competitive landscape and evolution

Liquidity provision in perpetual-futures markets is not a static picture but a dynamic ecosystem of competition and iteration. The preceding sections analyzed the cost structure, risk sources, and core strategies market makers face on CEXs and DEXs, but this concerns only the behavioral logic of individual market participants. This section extends the analytical dimension to the interactions among market participants and between participants and the institutional environment, examining how these interactions jointly shape the macro landscape of liquidity provision and its direction of evolution. This evolution unfolds around a complex game of technology, capital, information, incentives, and risk, exhibiting three major trends—concentration, algorithmic intelligence, and protocolization—which, while improving market efficiency, also breed new systemic risks.

19.7.1 The three-dimensional competition in technology, capital, and information

Competition among perpetual-futures market makers is a continually escalating technology race. The core driver of this competition stems from the fundamental source of market-making profit: capturing tiny spreads in the crevice between information asymmetry and inventory risk. Any advantage that can reduce latency, optimize models, or expand capital scale translates directly into higher profit margins and stronger survivability. This competition unfolds mainly along three dimensions: technology, capital,

and information.

The technology race is most directly expressed in the extreme pursuit of "speed." In market microstructure, a speed advantage means being able to react earlier to market signals—whether updating quotes ahead of competitors or being the first to complete an arbitrage when a cross-market spread appears. This gives rise to the classic "latency race" of high-frequency trading. Market makers invest heavily in server colocation, physically deploying their trading servers in the same data center as the exchange's matching engine to reduce network latency to the microsecond level. However, as the technology of top market makers converges, the excess profit obtained from a pure speed advantage is quickly compressed. The focus of competition then shifts to a more complex arena: algorithms and models. A superior risk model can quantify adverse selection more precisely, a smarter inventory-management algorithm can minimize risk exposure while maximizing volume, and a more sophisticated trading-signal-generation system tries to mine faint predictive information from massive data.

The capital race is expressed as the powerful barrier of economies of scale. A market maker with more ample capital can withstand greater inventory risk, thereby providing deeper liquidity on the order book and capturing more trading flow. At the same time, it can make markets on dozens or even hundreds of pairs across multiple exchanges simultaneously, smoothing overall P&L through diversification. More importantly, ample capital provides the necessary risk buffer for a market maker to keep operating in extreme conditions. When the market swings violently and a liquidation cascade erupts, undercapitalized market makers are forced to liquidate their positions and exit the market, while well-capitalized market makers may find opportunities for high returns in this "liquidity vacuum." This cumulative-advantage effect makes the market-making industry structurally prone to concentration.

The information race is harder to observe directly than the previous two dimensions, but its impact on long-term competitive advantage is equally significant. Beyond public market data, market makers seek a unique "information advantage." This can come from a deep parsing of on-chain data—for example, anticipating market sentiment by monitoring anomalies in whale addresses or the fund flows of DeFi protocols. It can also come from the symbiotic relationship established with exchanges, obtaining more favorable trading fees, higher API rate limits, and, in some cases, finer-grained order-flow data through a market maker program. The accumulation of this information advantage further consolidates the competitive barriers of top institutions.

Competition along these three dimensions leads to one inevitable result: profit compression and market concentration. As competition intensifies, the bid-ask spreads of mainstream pairs are continuously pushed to the limit, and the profit space of market making becomes ever thinner. Market makers slightly behind in technology, capital, or information see their profits quickly eroded and are ultimately forced to exit the market. This causes the share of liquidity provision to concentrate increasingly in a few global top market-making institutions. Meanwhile, a new competitive dimension is emerging: as crypto ETFs are approved in several major markets and regulatory frameworks gradually clarify, traditional-finance market-making giants such as Citadel Securities and Jane Street are progressively expanding their footprint in crypto market making.

The capital scale, cross-market experience, and compliance capabilities these institutions bring far exceed those of crypto-native market makers. The scale and pace of traditional giants' entry into crypto market making fluctuate with the regulatory environment (for example, some institutions' crypto exposure has at times contracted amid regulatory uncertainty), but the overall direction is that their entry may fundamentally change the competitive landscape: future competition will no longer be merely a contest among crypto-native institutions but a new-dimensional game between crypto-native market makers and traditional-finance giants.

By virtue of their technology and capital advantages, top market makers can continuously offer narrower bid-ask spreads than other competitors, and this gap widens over time. At the same time, as profit space is compressed, large numbers of small and mid-sized market makers are squeezed out of the market, the total number of active market makers declines year by year, and the market share occupied by top institutions correspondingly climbs steadily. This concentration trend constitutes the root of the systemic risk discussed in the next section.

19.7.2 Concentration and systemic risk

The concentration of the market-making industry has a dual effect. On the positive side, it brings significant efficiency gains. Top market makers, leveraging their economies of scale, can provide the market with more stable, deeper, and lower-cost liquidity. For ordinary traders, this means smaller trading slippage and better execution prices. From this angle, concentration is an optimization result under market competition and natural selection, handing the specialized work of liquidity provision to the most efficient producers.

However, this efficiency gain comes at a steep cost: the accumulation of systemic risk. When the overwhelming majority of a market's liquidity depends on a few highly homogeneous participants, the entire market becomes extraordinarily fragile. This concern is not unique to crypto markets; traditional market-microstructure research long ago provided an empirical foundation for it: Comerton-Forde et al. (2010), using inventory and P&L data of specialist market makers on the New York Stock Exchange, demonstrate that market makers' inventory levels and cumulative returns can significantly explain the time variation in market liquidity—when market makers' inventory is under stress or they suffer losses, bid-ask spreads widen nonlinearly and liquidity contracts accordingly [23]. This transmission mechanism from "market makers' financial condition to liquidity provision" provides a direct theoretical and empirical anchor for this section's analysis of the fragility of a concentrated market. In the crypto perpetual-futures market specifically, this risk is manifested mainly at two levels: single-point-of-failure risk and behavioral-correlation risk.

Single-point-of-failure risk is the danger that the sudden "going offline" of any top market maker can inflict a disproportionately large shock on the market. Such an offline event may stem from various factors: a technical failure, an operational error (such as a mistaken risk-parameter setting), a financial crisis (such as huge losses on its own assets), or even a regulatory raid. If a market maker holding a substantial share of liquidity (such as one of the top three to five institutions mentioned earlier) suddenly

withdraws all its quotes, the order book will thin instantly, bid-ask spreads will widen sharply, and the market's price-discovery capacity will be severely impaired. In extreme cases, this could trigger a chain reaction, leading to broader liquidity evaporation and a price flash crash. An even more extreme scenario is that the market maker itself may be forcibly liquidated in a violent move: under a unified margin system, a huge loss on one pair may consume the market maker's available margin on other pairs, triggering a chain of forced liquidations across pairs. This passive, involuntary "going offline" is more destructive than an active retreat, because it both removes liquidity supply and directly injects additional selling pressure into the market.

Behavioral-correlation risk is more hidden and dangerous. Although top market makers are competitors in day-to-day business, their technology stacks, risk models, and decision logic are largely convergent. They all rely on similar data sources, follow similar microstructure theories, and are constrained by the same market institutions (such as the funding rate and liquidation mechanism). This means that when the market encounters extreme stress, their reactions are likely to be highly correlated. For example, in a sharp price decline, several market makers' risk models may simultaneously hit their maximum inventory limits or maximum loss thresholds, so they collectively decide to contract quotes or withdraw from the market. This "collective retreat" behavior is precisely the core microlevel mechanism of the "sharp liquidity contraction" phenomenon we will examine in depth in the next chapter. Rational decisions at the individual level produce a systemic negative externality at the collective level: at the very moments when the market's demand for liquidity is highest, the supply of liquidity contracts in sync.

However, the "collective retreat" narrative above needs an important qualification: market makers are not entirely homogeneous actors, and different types of participants behave differently in a crisis. This requires an observable demarcation criterion for the two scenarios—"collective-retreat-dominant" and "competitive-backfill-dominant"—so as to avoid ex post self-justifying explanations. We propose the number of heterogeneous participants and their capital buffers as the demarcation variables: when the number of heterogeneous participants "willing and able to provide liquidity" is ample and their risk-control models and data sources are mutually independent, the backfill effect dominates, and liquidity can be partly replenished during a crisis; conversely, when such heterogeneous participants are scarce, or their risk-control models trigger in sync because of high homogeneity, the retreat effect dominates, and liquidity dries up in unison. The specific thresholds corresponding to this criterion await empirical calibration, so this section proposes it as a theoretical inference. Under this criterion, several structural regularities can be observed: Tier 1 institutions, because of their thicker capital buffers and more mature risk-control systems, usually trigger their retreat thresholds later than Tier 2 market makers, but once they retreat, their impact on market depth is also greater. Strategy heterogeneity matters equally: institutions running pure market-making strategies tend to contract comprehensively in extreme conditions, whereas institutions running a mix of market-making and arbitrage strategies may increase cross-market arbitrage even as they withdraw from market-making activity, and the latter's behavior objectively provides the market with some resilience. In mild volatility, the number of heterogeneous participants is usually ample, and the

backfill effect can effectively mitigate concentration risk; but in extreme tail events, when all institutions' risk-control models trigger in sync because they use similar risk measures and the same data sources, heterogeneity plunges, the backfill effect fails sharply, and the destructive power of behavioral correlation is fully revealed. The true critical point of systemic risk therefore lies not in the number of market makers in daily operation but in whether heterogeneous participants "willing and able to provide liquidity" still exist in extreme conditions.

The systemic risk brought by this concentration has precedents in traditional financial markets and has drawn close attention from regulators. But in the cryptocurrency market, this problem is especially acute, because the regulatory framework here is still immature and lacks a crisis-response mechanism analogous to a "market maker of last resort" or central-bank liquidity support. Understanding and monitoring the concentration of liquidity provision, assessing the health of top market makers, and designing mechanisms that incentivize market makers to remain in a crisis therefore become one of the core challenges of maintaining market stability.

19.7.3 Incentive design and market-maker behavior

Market makers and exchanges are locked in an ongoing game. Exchanges want market makers to provide the narrowest possible spreads and the deepest possible liquidity to attract more traders and improve market quality. Market makers want the highest possible return while bearing the lowest possible risk. The equilibrium of this game is determined by the incentive mechanisms exchanges provide. These incentive designs directly affect market-maker behavior and are ultimately transmitted to the liquidity condition of the entire market.

On centralized exchanges, incentive mechanisms revolve mainly around the "market maker program." The basic components of this program—fee rebates, API priority, and margin benefits—were analyzed in detail in Section 19.5.4 and are not repeated here in structure; instead, we focus on how these incentives shape market makers' behavioral patterns. A core behavioral effect is that higher rebates incentivize market makers to quote narrower spreads, because even if the spread itself is unprofitable, they can still profit from the rebate. However, this incentive can also lead to the emergence of "phantom liquidity," in which market makers post large numbers of orders to obtain rebates but quickly cancel them when the market truly needs liquidity. This "phantom liquidity" problem reveals the inherent limitation of a static rebate mechanism: it incentivizes depth supply under normal conditions but cannot constrain the collective retreat in extreme conditions.

On decentralized exchanges, incentive design becomes more varied and complex. For protocols adopting the on-chain CLOB or market-making vault model, incentives for liquidity providers come mainly from a share of trading fees and protocol-token incentives, with the design goal of attracting professional market makers to provide deep liquidity continuously.

For DEXs adopting an on-chain order book, incentive design is closer to that of CEXs, likewise including trading-fee rebates and volume-based rewards. But because of the transparency of the on-chain environment, incentive design faces new challenges. For

example, a protocol must design sophisticated algorithms to identify "effective" market-making behavior and prevent market makers from cheating for incentives through fake fills. Hyperliquid's points-reward system is one example; it tries to reward market makers that genuinely contribute liquidity more fairly by comprehensively assessing multiple dimensions such as resting-order time, distance from the mid-price, and volume.

Furthermore, some protocols are beginning to explore "protocol-controlled liquidity" or keeping ownership of liquidity in their own hands through instruments such as bonds. This model attempts to fundamentally solve the problem of market makers withdrawing in a crisis. However, it also fully transfers the complexity of risk and incentive design to the protocol itself. The protocol must play the role of a "super market maker," whose performance depends entirely on the algorithm design and risk-management capabilities of its internal team.

From a game-theoretic perspective, the optimal incentive design should achieve "incentive compatibility"—aligning market makers' pursuit of their own maximum self-interest with the protocol's goal of improving market liquidity and robustness. This requires protocol designers to deeply understand market makers' cost structure (the three costs) and the risks they face (the four types of toxic flow). A good incentive mechanism should encourage competition and compress spreads when the market is calm, while providing sufficient risk compensation for those market makers that choose to stay and continue providing liquidity when the market is turbulent. For example, one can design a dynamic rebate mechanism that automatically raises the rebate ratio for makers when market volatility rises sharply, thereby partly offsetting market makers' increased inventory risk and adverse-selection cost. Implementing such a dynamic incentive mechanism is a potential advantage of DeFi protocols relative to CEXs and a key direction for future liquidity-governance exploration.

19.7.4 Future directions of evolution

The evolution of perpetual-futures liquidity provision does not stop at the current competitive landscape. Starting from the analytical framework constructed in this chapter—the three-pillar cost structure, the toxic-flow spectrum, and the market-making competitive landscape—we can identify two evolutionary directions directly relevant to the core argument. This is only a forward-looking, outline-level discussion, meant to point out these directions' connection to this chapter's framework rather than to develop them into independent new frameworks.

First, artificial intelligence is reshaping the risk-management and toxicity-detection capabilities of market making. Traditional market-making algorithms are based on a parameterized rule set—for example, "when inventory exceeds X, skew the quotes by Y basis points"—and their ability to adapt to changes in market state is limited. Reinforcement-learning-based market-making agents, by contrast, can explore optimal strategies in a multidimensional market-state space (volatility, order-flow imbalance, funding rate, etc.). This technological advance intersects with this chapter's framework at two direct points: first, at the level of toxic-flow identification, AI systems have the potential to distinguish in real time, more precisely, the four types of toxic flow defined

in Section 19.4 (traditional indicators such as VPIN mainly capture informed-trading toxicity, whereas AI models have the potential to simultaneously identify liquidation-cascade signals, arbitrage-extraction patterns, and MEV-attack characteristics); second, at the level of inventory management, AI can dynamically optimize the parameters in the extended decision function $\delta^* = f(Q, \sigma, \gamma, FR_t, T_{\text{eff}}, \kappa)$ discussed in Section 19.3.1. A caution, however: when many market makers adopt similar AI architectures and training data, their strategies may produce highly correlated response patterns, withdrawing liquidity in sync during extreme moves and thereby intensifying the behavioral-correlation risk analyzed in Section 19.7.2. In addition, AI market making introduces an operational-risk dimension that traditional algorithmic market making does not have: a model may produce unpredictable anomalous behavior in market states not covered by its training data (the 2012 case in which Knight Capital lost about $460 million in 45 minutes due to an out-of-control algorithm is a cautionary precedent from traditional finance), and the black-box nature of AI makes both post hoc diagnosis and real-time intervention more difficult.

Second, dynamic incentive mechanisms may change market makers' behavioral patterns in a crisis. The core contradiction repeatedly argued in this chapter—that liquidity provision is scarcest precisely when demand is greatest—is rooted in the fact that a market maker's profit equation points to withdrawal in extreme conditions. Dynamic incentive mechanisms attempt to modify the parameters of this equation. For example, one can design a volatility-responsive rebate mechanism that automatically raises the rebate ratio for makers when market volatility exceeds a specific threshold, directly increasing the "spread revenue" term in a market maker's profit equation and partly offsetting surging institutional cost and adverse-selection cost. Such a mechanism essentially transforms crisis-time liquidity provision from a purely market behavior into a behavior backed by institutional incentives, changing a market maker's optimal strategic choice under extreme conditions. On-chain programmable contracts make the implementation of such dynamic mechanisms possible, which is also a potential institutional advantage of the on-chain CLOB paradigm relative to traditional CEXs.

In addition, intent-based architectures and cross-chain liquidity integration are also gradually changing the infrastructural form of liquidity provision: the former shifts execution complexity from the user side to the professional-market-maker side, while the latter works to solve the margin-fragmentation problem. These two directions intersect only indirectly with this chapter's core analytical framework and belong more to the evolution of market infrastructure, so they are not elaborated here.

Overall, AI-driven intelligence and the dynamization of incentive design are the two evolutionary directions most directly relevant to this chapter's three-pillar cost framework and the toxic-flow spectrum. They have the potential to mitigate the inherent fragility of liquidity provision at the technical and institutional levels, respectively, but they also introduce new risk dimensions such as strategy resonance and incentive arbitrage, which market participants and protocol designers must continue to monitor.

19.8 Chapter summary

This chapter analyzed a central question: in perpetual-futures markets, where does liquidity come from? The analysis shows that liquidity is not a naturally existing resource but a commodity produced by participants in different roles at the cost of bearing specific risks and costs. It offered three interrelated frameworks for understanding this production process.

At the level of cost structure, this chapter argued that the twin pillars of classic market-making theory—adverse-selection cost and inventory-risk cost—are insufficient to explain market-making behavior in perpetual-futures markets, and that the institutional design of perpetual futures (the funding rate, the liquidation mechanism, no expiration date, and 24/7 trading) introduces a "third pillar": institutional cost. This "third pillar" is an integrative consolidation and unified naming of several already-known cost sources in the perpetual-futures context, not a newly discovered cost type. It extends a market maker's profit equation to Total profit = spread revenue − (adverse-selection cost + inventory-risk cost + institutional cost). Institutional cost structurally raises the threshold of perpetual-futures market making and explains why its liquidity cost is usually higher than that of spot or traditional futures markets under comparable conditions. From BTC to altcoins, the sharp rise in the share of adverse-selection cost further reveals the microeconomic root of the liquidity inefficiency of small-cap tokens. Building on this, this chapter constructed the "toxic-flow spectrum" classification framework, placing the toxicity risk market makers face into a unified coordinate system composed of "information content" and "permanence of price impact," and accordingly identifying four toxicity regions with different dominant mechanisms but overlapping (rather than mutually exclusive and collectively exhaustive) boundaries: informed-trading toxicity, arbitrage-extraction toxicity, liquidation-cascade toxicity, and MEV-attack toxicity. These four types of toxicity differ systematically in information content, permanence of price impact, temporal characteristics, and response strategy, and confusing them leads to a strategy mismatch: using a retreat strategy meant for informed trading to deal with liquidation flow would miss the rebound opportunity, whereas the reverse could incur major losses in a genuine price impact. At the level of market structure, this chapter proposed a unified framework for comparing CEX and DEX liquidity-provision paradigms. The on-chain CLOB paradigm and the market-making vault paradigm make different trade-offs across dimensions such as the risk-bearing party, pricing mechanism, LP autonomy, capital efficiency, and barrier to entry; there is no all-around dominant solution, and the two represent different positions in the trade-off among decentralization, efficiency, and risk control.

Together, these three frameworks answer this chapter's core question: the liquidity of perpetual futures is dynamically produced by professional market makers on CEXs and by liquidity providers in on-chain CLOBs and market-making vaults on DEXs, within the game of coping with the three costs and the four types of toxic flow. The cost structure and risk characteristics of this production process directly determine the quantity, quality, and stability of the liquidity observable in the market. However, this production system contains a structural contradiction: the quantity of liquidity supplied is negatively correlated with market demand. When the market is calm, market-making

risk is low and liquidity supply is ample, manifested as narrow spreads and a deep order book; when the market swings violently, the four types of toxic flow erupt simultaneously, inventory risk and liquidation risk rise sharply, and a market maker's profit equation points to the only rational choice: widen spreads, cut depth, and even withdraw from the market. Liquidity supply is most ample when demand is smallest and scarcest when demand is greatest. This paradox does not stem from a behavioral flaw of market makers but is the inevitable result of the profit equation under extreme conditions—a fragility endogenous to market microstructure. The next chapter focuses on the dynamic unfolding of this core paradox: the endogeneity and reflexivity of liquidity, and how it leads to the recurring market phenomenon of liquidity contracting sharply precisely when it is most needed.

References

[1] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, *135*(2), 293–319. https://doi.org/10.1016/j.jfineco.2019.07.001

[2] Ackerer, D., Hugonnier, J., & Jermann, U. (2025). Perpetual futures pricing. *Mathematical Finance*, *35*(1), 3–31. https://doi.org/10.1111/mafi.70018

[3] Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. https://doi.org/10.2307/1879431

[4] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100. https://doi.org/10.1016/0304-405X(85)90044-3

[5] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

[6] Ho, T., & Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, *9*(1), 47–73. https://doi.org/10.1016/0304-405X(81)90020-9

[7] Amihud, Y., & Mendelson, H. (1980). Dealership market: Market-making with inventory. *Journal of Financial Economics*, *8*(1), 31–53. https://doi.org/10.1016/0304-405X(80)90020-3

[8] He, S., Manela, A., Ross, O., & von Wachter, V. (2022). Fundamentals of perpetual futures. *arXiv preprint arXiv:2212.06888*. https://doi.org/10.48550/arXiv.2212.06888

[9] Schmeling, M., Schrimpf, A., & Todorov, K. (2023). Crypto carry. *BIS Working Papers*, No. 1087. https://doi.org/10.2139/ssrn.4268371

[10] Galati, L. (2024). Exchange market share, market makers, and murky behavior: The impact of no-fee trading on cryptocurrency market quality. *Journal of Banking & Finance*, *165*, 107222. https://doi.org/10.1016/j.jbankfin.2024.107222

[11] Easley, D., O'Hara, M., Yang, S., & Zhang, Z. (2024). Microstructure and market dynamics in crypto markets. *SSRN Working Paper*. https://doi.org/10.2139/ssrn.481

4346

[12] Easley, D., López de Prado, M. M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *The Review of Financial Studies*, *25*(5), 1457–1493. https://doi.org/10.1093/rfs/hhs053

[13] Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217–224. https://doi.org/10.1080/14697680701381228

[14] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[15] Milionis, J., Moallemi, C. C., Roughgarden, T., & Zhang, A. L. (2022). Automated market making and loss-versus-rebalancing. *arXiv preprint arXiv:2208.06046*. https://doi.org/10.48550/arXiv.2208.06046

[16] OECD. (2023). DeFi liquidations: Volatility and liquidity. *OECD Working Papers on Finance, Insurance and Private Pensions*, No. 48. https://doi.org/10.1787/0524faaf-en

[17] Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., & Juels, A. (2020). Flash Boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. In *2020 IEEE Symposium on Security and Privacy (SP)* (pp. 910–927). IEEE. https://doi.org/10.1109/SP40000.2020.00040

[18] Canidio, A., & Fritsch, R. (2023). Arbitrageurs' profits, LVR, and sandwich attacks: Batch trading as an AMM design response. In *Proceedings of the 5th Conference on Advances in Financial Technologies (AFT 2023)*, LIPIcs Vol. 282, pp. 24:1–24:17. https://doi.org/10.4230/LIPIcs.AFT.2023.24

[19] CoinGlass. (2025). *2025 cryptocurrency derivatives market report*. https://www.coinglass.com/learn/2025-annual-report-en

[20] Aramonte, S., Huang, W., & Schrimpf, A. (2021). DeFi risks and the decentralisation illusion. *BIS Quarterly Review*, December 2021, 21–36. https://www.bis.org/publ/qtrpdf/r_qt2112b.htm

[21] Binance. (2025). *USD(S)-margined futures market maker program: Fee schedule and program rules*. https://www.binance.com/en/support/faq/futures-market-maker-program

[22] Capponi, A., & Jia, R. (2021). The adoption of blockchain-based decentralized exchanges. *arXiv preprint arXiv:2103.08842*. https://doi.org/10.48550/arXiv.2103.08842

[23] Comerton-Forde, C., Hendershott, T., Jones, C. M., Moulton, P. C., & Seasholes, M. S. (2010). Time variation in liquidity: The role of market-maker inventories and revenues. *The Journal of Finance*, *65*(1), 295–331. https://doi.org/10.1111/j.1540-6261.2009.01530.x

The Endogeneity and Fragility of Liquidity

In early November 2022, the bid-ask spread on Bitcoin perpetual futures at major exchanges held at roughly 1 basis point, and aggregate order-book depth within 1% of the mid-price across the entire market totaled about $580 million [1]. Even a $10 million market order could be filled with minimal slippage. Yet after the FTX crisis erupted in full on November 8, liquidity deteriorated abruptly. Over the following week, bid-ask spreads at major exchanges widened to several times their normal level, and aggregate order-book depth within 1% of the mid-price collapsed from about $580 million to roughly $230 million, a contraction of more than 60%. Kaiko termed this structural collapse in liquidity the "Alameda Gap" [2]. At the same time, liquidity on FTX's own platform fell to nearly zero, and some trading pairs saw order-book depth evaporate to an extreme degree during the most violent hours of the collapse [3]. This was not a gradual process of depletion; it was market makers actively withdrawing their supply of liquidity within an extremely short window. The phenomenon raises the central question of this chapter: why can a liquidity system that appears amply supplied collapse within days or even hours, and why does that collapse always occur at the very moment the market's demand for liquidity is greatest?

> **Note on the level of evidence.** Several of the specific figures in this chapter's crisis cases—such as the opening $580 million→$230 million depth decline and 60% contraction, as well as the later characterizations of spreads, depth, and drawdowns for events including the May 19, 2021 crash, the May 2022 Terra-LUNA collapse, and the November 2022 FTX collapse—are drawn in part from data-vendor research notes and from exchange and project-team blogs (Kaiko, Deribit, Perpetual Protocol, and others). These are firsthand but independently unverified gray-literature sources. Where this chapter cites such figures, it does so primarily to convey qualitative trends rather than to present them as precise measured conclusions. The formalizations of theoretical mechanisms (Sections 20.2.5, 20.3.3, and 20.8) are all illustrative specifications rather than empirically calibrated models, and the relevant qualifications are flagged explicitly at each point.

The sudden disappearance of liquidity is not an incidental system failure but a structural tendency of the market maker's profit equation under extreme conditions. When a sharp price decline triggers a rise in volatility, market makers' inventory costs grow convexly, forcing them to widen spreads and reduce depth. This ebbing of liquidity in turn amplifies price impact, forming a self-reinforcing loop between volatility and

liquidity. At the same time, falling prices cause the system's effective leverage to climb passively, triggering forced-liquidation mechanisms. Liquidation orders enter the market as perfectly inelastic demand for liquidity: they not only directly consume the already-thin order-book depth, but the high order-flow toxicity they carry also drives out the remaining market makers, causing available liquidity to contract at a multiplicative rate. In addition, abrupt swings in the funding rate under extreme conditions force large volumes of arbitrage positions to close in concert, and the resulting one-directional order flow breaks the symmetry of liquidity on the two sides of the book. Running through all three of these positive-feedback loops is a split between apparent liquidity and true liquidity. The depth displayed in a calm market is merely a conditional commitment by market makers; once stress exceeds their tolerance band, those commitments quickly lapse.

This chapter first establishes the theoretical foundations of liquidity endogeneity (Section 20.1), then analyzes in turn the three positive-feedback loops of the volatility spiral, the liquidation cascade, and funding-rate distortion (Sections 20.2–20.4). It then examines the split between apparent and true liquidity (Section 20.5), a five-stage framework of liquidity evaporation (Section 20.6), and the cross-dimensional pathways of contagion (Section 20.7).

20.1 Theoretical foundations of liquidity endogeneity

Traditional market-microstructure theory tends to treat liquidity as an exogenously given background condition. In the institutional environment of perpetual futures, however, liquidity exhibits a high degree of endogeneity, coupled deeply with price, volatility, and leverage through positive-feedback loops. This section establishes the theoretical foundations of this endogenous perspective and reveals the deeper mechanisms that cause liquidity to evaporate suddenly under stress.

20.1.1 Exogenous versus endogenous liquidity

In the classical framework of financial-market analysis, liquidity is typically assumed to be an exogenous variable, determined by market makers' capital reserves and risk preferences and treated as a slowly changing constant over the short term. In perpetual futures markets, however, the supply of liquidity is the dynamic optimal solution that market makers arrive at after continuously assessing toxicity risk, inventory costs, and institutional constraints. The parameters in the profit equation depend, directly or indirectly, on the current market state; changes in the market state adjust the supply of liquidity through the profit equation; and that adjustment in turn reshapes the market state, forming an inseparable closed loop. This paradigm shift from exogenous to endogenous liquidity means that the sudden disappearance of liquidity no longer requires an external shock to explain it, but is instead a foreseeable outcome once the system's internal positive-feedback loops are activated. A somewhat larger-than-usual routine trade, or an ordinary liquidation event, is sufficient to set off a self-reinforcing collapse spiral so long as it happens to cross some trigger band. This perspective draws heavily on the theory of market-liquidity and funding-liquidity spirals proposed

by Brunnermeier and Pedersen (2009) [4], but within the specific institutional context of perpetual futures, this endogenous coupling exhibits more complex features.

20.1.2 The reflexivity triangle of liquidity

Within the framework of endogenous liquidity, we can construct a core theoretical model that explains the fragility of liquidity in perpetual futures markets: the reflexivity triangle of liquidity. In this model, the three vertices of the triangle represent the supply of liquidity, the price level, and the position structure, while the three edges connecting them represent three positive-feedback loops: the volatility-driven mechanism, the leverage-liquidation mechanism, and the funding-rate mechanism. Deterioration along any one dimension can trigger a chain reaction along the others, ultimately converging into a systemic evaporation of liquidity.

The concept of reflexivity has a specific scholarly lineage in finance. Soros (1987) [5] defined reflexivity as a two-way feedback relationship between market participants' perceptions and market reality. Brunnermeier and Pedersen (2009) [4] formalized this feedback mechanism through the two-way spiral between market liquidity and funding liquidity. The reflexivity triangle proposed in this chapter builds on the Brunnermeier-Pedersen framework by introducing the funding-rate dimension that is unique to perpetual futures, extending the two-sided spiral into a three-dimensional positive-feedback network. This extension reflects the fact that perpetual futures, as a synthetic instrument with no expiry and high leverage, have liquidity dynamics with additional sources of fragility that traditional derivatives markets do not possess.

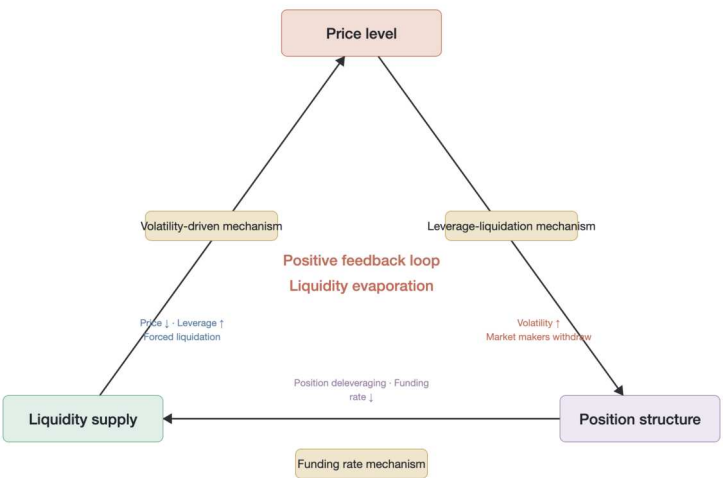


Figure 20-1. The reflexivity triangle of liquidity: the positive-feedback loops among price, volatility, and leverage

The first edge of the reflexivity triangle connects price, volatility, and liquidity. When the market suffers an initial price-decline shock, the realized volatility of the asset rises

markedly. As our earlier analysis of market makers' inventory costs showed, there is a strong convex relationship between the carrying cost of inventory risk and volatility. A multiplicative increase in volatility causes market-making costs to surge convexly (quadratically). Faced with a rapidly deteriorating profit equation, rational market makers adopt defensive strategies: first widening the bid-ask spread to compensate for adverse-selection risk, and then cutting order-book depth to limit their potential inventory exposure. When volatility breaches a market maker's tolerance band, the maker withdraws from the market entirely. This substantive decay in liquidity means that subsequent trades of the same size produce larger price impact, which further raises volatility, forming a self-reinforcing vicious cycle.

The second edge of the triangle describes the interaction among price, leverage, and liquidity. In perpetual futures markets, a sharp decline in price rapidly erodes the margin balance of long positions, causing the system's effective leverage to climb passively. When margin falls below the maintenance threshold, the liquidation engine is triggered. Liquidation orders, as a form of mandatory demand for liquidity, pour into the market as market orders and directly consume what little depth remains on the order book. At the same time, the extremely high toxicity carried by the liquidation flow forces market makers to retreat faster. This twofold drain on liquidity pushes the price down further, dragging still more previously safe positions into the liquidation zone and triggering a cascading collapse.

The third edge of the triangle reveals the intrinsic link among the funding rate, position structure, and the directional distribution of liquidity. In extreme market conditions, the funding rate often tilts sharply in one direction. For example, in a scenario of plunging prices accompanied by extreme negative rates, short holders face high carrying costs and are forced to close by buying. This concentrated buy-side demand for liquidity comes into sharp conflict with the reduced buy-side supply that market makers offer because of inventory skew, producing a severe directional distortion in liquidity. An abrupt swing in the funding rate can also trigger the concentrated unwinding of enormous cash-and-carry arbitrage positions, and this structural mass migration of positions places unbearable one-directional pressure on an already fragile liquidity pool.

These three edges do not operate in isolation; they exhibit a high degree of nonlinearity. In real crisis scenarios, they are often activated simultaneously and produce cross-amplification effects. Surging volatility accelerates the retreat of market makers, so that liquidation orders confront thinner depth; the shock of liquidation in turn drives volatility higher and worsens the distortion of the funding rate. This multidimensional resonance of risk means that the total destructive force on liquidity typically exceeds the simple sum of the individual factors' effects. The claim of "exceeding the simple sum" here is a testable proposition: if, in a given collapse, the total shock when multiple loops act concurrently does not exceed the sum of the shocks from each loop acting alone, then the cross-amplification hypothesis does not hold in that case. This chapter treats it as a working hypothesis awaiting empirical testing rather than a proven conclusion.

The foregoing analysis carries an important simplifying assumption: that the market's major market makers are highly homogeneous in their risk-control models, risk thresholds, and information sets. This assumption is what allows the arguments about

coordination games and synchronized retreat to hold. In actual markets, some market makers may adopt differentiated strategies—for example, providing liquidity counter-directionally during a crisis to capture extreme spreads—but empirical evidence shows that market makers' behavior in extreme conditions is far more synchronized than in normal times (Anand and Venkataraman, 2016 [6]), supporting the reasonableness of the homogeneity assumption as a first-order approximation.

The reflexivity-triangle model also has an important boundary of applicability: it describes endogenous liquidity crises at the level of market microstructure, but it does not cover liquidity evaporation driven by counterparty credit risk. In the FTX collapse, the disappearance of liquidity was largely driven not by a reflexive loop of volatility, leverage, or funding rates, but by a collapse of market participants' confidence in the exchange's own solvency. Even when volatility and leverage are at normal levels, once traders doubt that an exchange can honor their margin and profits, liquidity evaporates instantly, and this is a contagion channel independent of the reflexivity triangle. This chapter's analytical framework focuses on endogenous mechanisms at the market-microstructure level; a systemic analysis of counterparty credit risk is developed in a later chapter.

20.1.3 The four liquidity amplifiers of perpetual futures

If the reflexivity triangle describes the general dynamics of a liquidity collapse, then the institutional design specific to perpetual futures constitutes four structural dimensions that amplify that mechanism. These features, regarded in normal times as core advantages of the product, become sources of heightened liquidity fragility under stress.

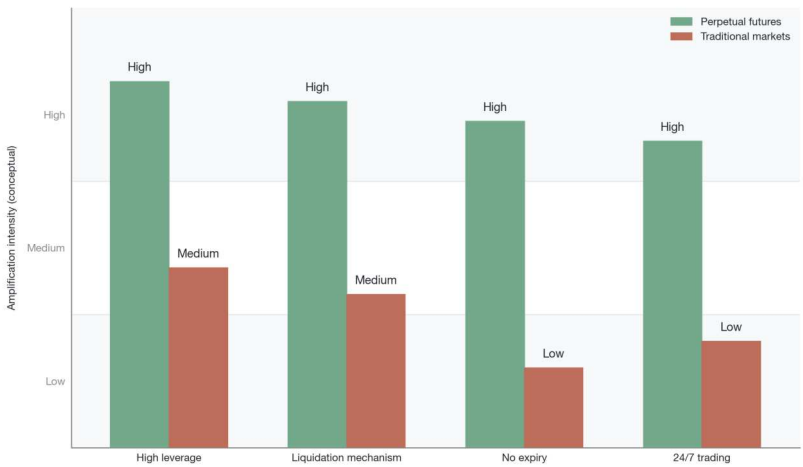


Figure 20-2. The four liquidity amplifiers of perpetual futures (conceptual illustration, not measured values)

The first amplifier is extremely high leverage. Unlike traditional financial markets, which impose strict regulatory limits on leverage, perpetual futures markets commonly permit trading at very high multiples of leverage. This means that a given amount

of initial capital can establish a notional position far exceeding its own equity. When the market reverses, these enormous notional positions convert into equally enormous liquidation volume, exerting far greater impact pressure on market liquidity than in traditional markets. High leverage not only amplifies returns; it also structurally amplifies the system's potential claim on liquidity.

The second amplifier is a mandatory and immediate liquidation mechanism. In traditional margin trading, when account equity falls below the requirement, a trader typically receives a margin call and is given some buffer of time to add funds or close positions voluntarily. Liquidation in perpetual futures, however, is strictly enforced by code: it is non-negotiable and offers no time buffer. Once a threshold is breached, the system automatically takes over and sells off the assets in the form of market orders or aggressive limit orders. Although major modern exchanges have introduced multi-tiered liquidation processes to mitigate this shock (the tiered execution and its negative dependence on depth impact are detailed in Section 20.3.3), the effectiveness of these buffers declines markedly when the liquidation queue backs up rapidly in extreme conditions, and the shock still approaches immediate full execution. This mechanism converts potential risk into an actual run on liquidity within an extremely short time and can readily drain the order book in short order.

The third amplifier is the absence of an expiry date. Traditional deliverable futures contracts have a definite expiry; at expiry, the contract price naturally converges to the spot price and all open positions are forcibly settled. This mechanism provides the market with a natural stabilizing anchor. Perpetual futures instead maintain their price anchoring through the funding-rate mechanism, which allows imbalanced positions to persist indefinitely. In extreme conditions, market makers cannot count on holding to expiry to work off a temporary paper loss, which pushes them toward immediate stop-losses and retreat when confronting risk, thereby reducing the system's resilience in absorbing shocks.

The fourth amplifier is a round-the-clock trading environment. Perpetual futures markets operate continuously, seven days a week, twenty-four hours a day, and lack the circuit breakers and price limits that serve as shock absorbers in traditional markets. This means a liquidity shock can erupt at any moment, including the pre-dawn hours when market-maker participation is lowest and order-book depth is thinnest. During these liquidity troughs, even a relatively small shock is enough to trigger violent price swings and set off the collapse spiral of the reflexivity triangle. The fragility of round-the-clock trading also includes an often-overlooked operational-risk dimension: the technical infrastructure of crypto exchanges has a poor track record under extreme load. During the crash of May 19, 2021, several major exchanges experienced surging matching-engine latency, suspended deposit and withdrawal services, and even complete outages, leaving many traders unable to place stop-loss orders or add margin in time. Technical failures not only impede risk-management operations; they also force market makers to withdraw liquidity preemptively out of distrust in the stability of an exchange's systems, making infrastructure failure an independent amplifier of liquidity evaporation.

20.1.4 The wedge between apparent and true liquidity

The endogenous evolution and amplification of liquidity also give rise to a key concept: the wedge between apparent liquidity and true liquidity. The bid-ask spread and depth displayed on the order book constitute apparent liquidity, but these metrics reflect only market makers' conditional commitments given the current market state. Once volatility breaches their tolerance band or one-directional order flow accumulates, those commitments are withdrawn within milliseconds. Because major market makers use similar risk-control models, their retreat is highly synchronized, so the disappearance of order-book depth manifests as an abrupt collective evaporation rather than a gradual decay. The mechanism by which the wedge between apparent and true liquidity forms, the microstructure of ghost liquidity, and the resulting failure of liquidity metrics are analyzed systematically in Section 20.5.

20.2 The volatility-liquidity spiral

The first edge of the reflexivity triangle of liquidity reveals a self-reinforcing positive-feedback loop between volatility and liquidity. At the core of this loop is the fact that a rise in volatility is not merely a symptom of market risk; it is a direct driver that weakens the supply of liquidity. When market makers confront a violently volatile market, their rational defensive behavior—widening spreads and pulling orders—in turn amplifies the market's price impact and thereby drives volatility higher. This mechanism can be observed across multiple crashes in crypto-asset markets.

20.2.1 The market maker's volatility-response function

The effect of volatility on liquidity can be explained from the market maker's profit equation. In Chapter 19, we decomposed the market maker's three costs in detail: adverse-selection cost, inventory cost, and institutional cost. The convex relationship between inventory cost and volatility is central to understanding market makers' volatility response. Ho and Stoll (1981) [7], in their classic dealer-inventory model, first derived that the optimal bid-ask spread has a nonlinear positive relationship with inventory risk (a function of volatility). Avellaneda and Stoikov (2008) [8] further formalized this conclusion within an optimal-control framework for high-frequency market making, showing that the optimal spread is proportional to the square of volatility (optimal spread $\propto \gamma\sigma^2$), where γ is the risk-aversion coefficient (here γ follows the notation of Avellaneda and Stoikov's original paper and is synonymous with the α used to denote the risk-aversion coefficient in Section 20.2.5 of this chapter). This means that when market volatility doubles, the inventory risk borne by the market maker does not increase linearly but climbs at four times the rate. This nonlinear cost structure dictates that a market maker's response to rising volatility is necessarily sharp and abrupt.

Rising volatility also weakens liquidity through a second channel: it intensifies adverse-selection risk. The sequential-trade model of Glosten and Milgrom (1985) [9] shows that the equilibrium level of the bid-ask spread under information asymmetry is determined by the proportion of informed traders and the size of their informational advantage. When market volatility rises sharply, it is often accompanied by intensifying informa-

tion asymmetry, and the market maker cannot tell whether incoming orders reflect informed trading based on genuine information or mere panic selling. To protect itself against adverse selection by informed traders, the market maker must widen the bid-ask spread to add a buffer. In perpetual futures markets, the dual overlay of inventory-cost convexity and adverse-selection risk makes this defensive mechanism especially sharp. Because of the absence of an expiry constraint and the amplifying effect of high leverage, both costs are sharply magnified simultaneously in extreme conditions.

In high-frequency data on Bitcoin perpetual futures over 2020 to 2023, realized volatility and the bid-ask spread show a significant positive relationship (the sample period and field definitions remain to be specified; the thresholds below are illustrative characterizations). In calm market conditions (daily volatility below 3%), the bid-ask spread on Bitcoin perpetual futures at top exchanges typically holds between 0.5 and 1 basis point. When volatility breaches a certain threshold (for example, daily volatility above 8%), however, the spread widens rapidly to 10 basis points or more (as shown in Figure 20-3). This widening is not gradual; after the threshold is breached, it exhibits a sharp convex surge. This is the direct microstructural manifestation of inventory-cost convexity: when expected profit can no longer cover the sharply rising cost of risk, the rational choice for the market maker is to contract the supply of liquidity immediately. The spread data above reflect the gap between the best bid and best ask, but in extreme volatility the resting depth at the best quotes may be minimal. For a trader who needs to execute a particular size, the realized spread they face—the deviation of the volume-weighted execution price from the mid-price—will be far larger than the best-quote spread, so that the actual deterioration in liquidity is more severe than the quote data suggest.

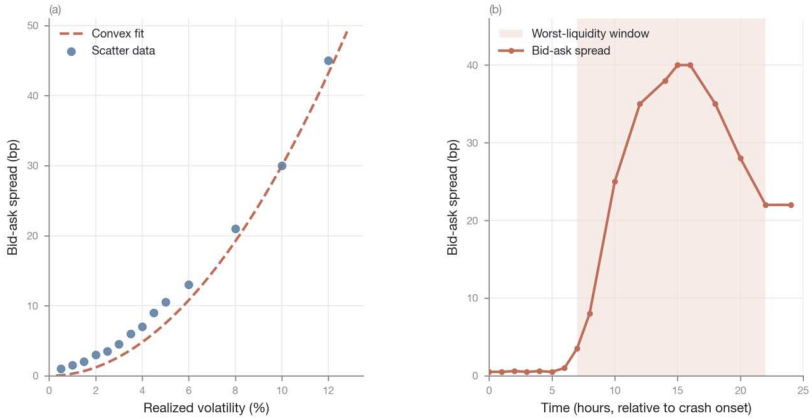


Figure 20-3. The nonlinear relationship between realized volatility and the bid-ask spread (illustrative characterization)

The retreat of market makers shows up not only in widening spreads but, more so, in the rapid evaporation of order-book depth. In the early phase of rising volatility, market makers may merely widen their quotes; but if volatility keeps climbing and reaches the circuit-breaker threshold in their internal risk-control models, they rapidly cancel all limit orders. At that point, the seemingly abundant "apparent liquidity"

on the order book vanishes in an instant, leaving an extremely fragile market vacuum. This behavior stems not from panic or irrationality but from a precise calculation of the profit function: when inventory-cost convexity makes the expected return to continued market making negative, retreat becomes the only optimal choice.

20.2.2 The transmission mechanism of spread widening

Market makers' widening of spreads and withdrawal of depth is merely the starting point of the volatility-liquidity spiral. This microstructural contraction of liquidity then acts back on the macro market's price-discovery process through a clear transmission chain, ultimately driving volatility still higher.

First, the widening of spreads directly causes transaction costs to rise sharply. For an ordinary market-order trader, this means not only greater slippage but also an order book whose depth has been drastically reduced. In such an environment, even a medium-sized order cannot be fully absorbed near the best bid and ask, and instead inevitably "breaks through" several price levels. This phenomenon is the amplification of price impact.

The amplification of price impact means that a given trading volume produces a far larger price movement in a liquidity-starved market than it would under normal conditions. The price-impact coefficient proposed by Kyle (1985) [10] can quantify this effect. Under normal market conditions, a $10 million Bitcoin market order might cause only a 0.05% price move; but in the extreme conditions of a liquidity retreat, the same order could cause the price to plunge 1% or more in an instant. This violent price swing caused by liquidity depletion shows up statistically as a direct surge in realized volatility.

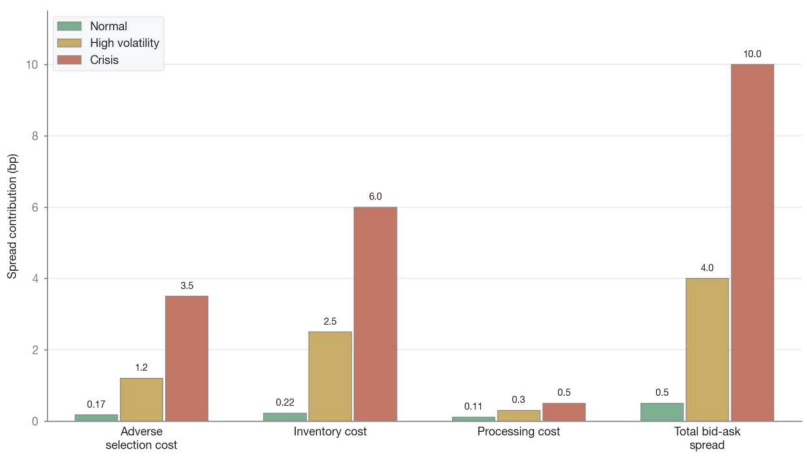


Figure 20-4. The Glosten-Milgrom decomposition of the bid-ask spread (normal, high-volatility, and crisis regimes; illustrative values)

At this stage, the rise in volatility has decoupled from any external fundamental information and has become a purely liquidity-driven phenomenon. No new negative news has arrived; simply for want of a bid, the price falls rapidly under the impact of a small

sell order. The emergence of this "non-informational volatility" marks the formal closing of the positive-feedback loop. High volatility drives liquidity to retreat, the retreat of liquidity amplifies price impact, and price impact drives volatility higher still. Once this spiral is set in motion, it accelerates continuously on its own endogenous positive feedback until some external force—an exchange outage, a circuit breaker, or the entry of large new capital—forcibly interrupts the process.

20.2.3 Trigger conditions and the critical point of the spiral

If the volatility-liquidity spiral is so destructive, why does it not occur every day? The answer is that the spiral's onset requires specific trigger conditions and involves a highly nonlinear critical band.

From a game-theoretic perspective, market makers providing liquidity face a classic "coordination game." Under normal conditions, if all market makers remain in the market, overall liquidity is ample, price impact is small, and the risk each maker bears is within a controllable range. This is a Pareto-optimal "good equilibrium." Once the market suffers some initial shock that raises volatility, however, each market maker begins to guess how its competitors will respond. If market maker A expects market maker B to pull its orders out of fear of risk, then A's optimal strategy is to pull its own orders ahead of B, so as not to become the sole liquidity provider in the market and bear all the adverse-selection risk.

This expectation-based, self-fulfilling mechanism gives the collapse of liquidity its suddenness and synchronization. When volatility approaches a certain critical band, market expectations reverse in an instant. At that moment, the strategic risk of maintaining quotes rises sharply, and collective retreat becomes the only Nash equilibrium. This mechanism has a structure similar to the bank-run model described by Diamond and Dybvig (1983) [11]: when depositors lose confidence in a bank's liquidity, the run itself causes the bank to fail. In perpetual futures markets, once market makers' confidence in market liquidity is shaken, their collective retreat instantly manufactures the very liquidity crisis they fear.

This analogy, however, contains an important structural difference. The strategic complementarity in the Diamond-Dybvig model arises from the first-come, first-served constraint of the deposit contract: depositors at the back of the line face the risk that the bank's assets have already been emptied, so each depositor has an incentive to withdraw first. The source of strategic complementarity in market-maker retreat is entirely different: it is not a queuing effect but an informational externality. When other market makers retreat, those who remain must face all of the toxic order flow alone, and their adverse-selection risk climbs sharply. The global-games framework of Morris and Shin (1998) [12] may capture this mechanism more precisely: each market maker makes a binary decision (stay or retreat) based on its private signal about volatility, order-flow toxicity, and peer behavior, and when the signal crosses a certain threshold, the state flip from staying to retreating happens almost simultaneously. This distinction carries important implications for the design of governance solutions: a bank run can be prevented through deposit insurance, whereas market-maker retreat driven by an informational externality requires different intervention tools, such as enhancing mar-

ket transparency to reduce signal noise, or introducing mandatory minimum-quoting obligations to remove the dominant strategy of retreat.

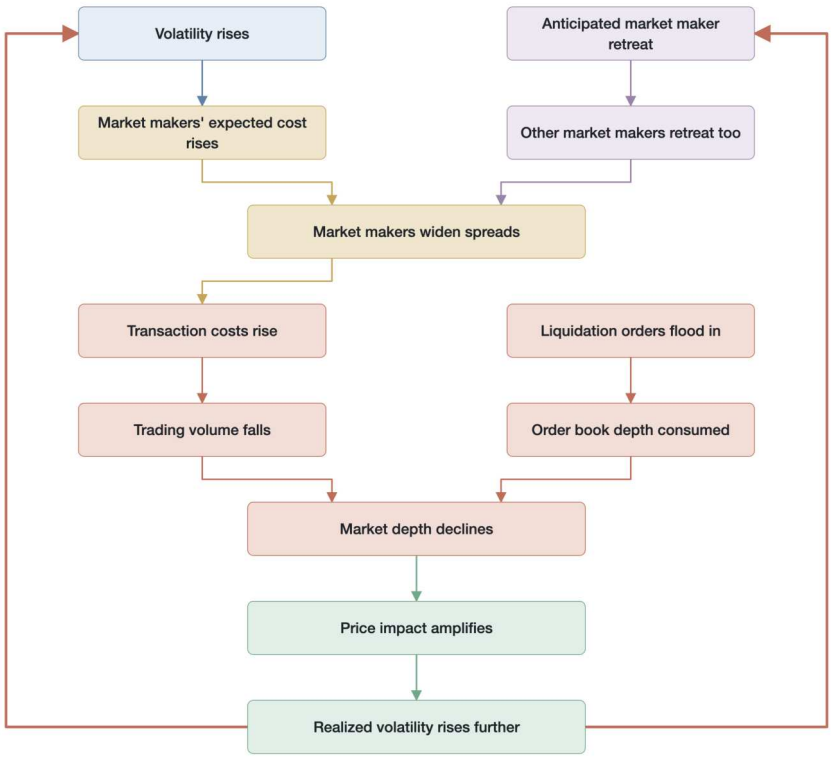


Figure 20-5. The transmission chain and critical point of the volatility-liquidity spiral

The Bitcoin flash crash of May 19, 2021 provides a typical case. That day, under the blow of multiple macro headwinds, the Bitcoin price began to fall. At first this was merely a normal market correction. But as the price broke through key support levels, large numbers of high-leverage long positions were pushed into the liquidation zone. Forced-liquidation market sell orders poured into the market, causing volatility to surge. Within about six hours, the bid-ask spread at top exchanges widened from a normal 0.5 basis point to 40 basis points, and order-book depth evaporated by more than 80%. Under the extreme load, Binance even experienced a gap of as long as 40 minutes in its trading data. In the phase of complete liquidity vacuum, the Bitcoin price fell at one point from about $43,000 to about $30,000, a drop of more than 30%. The core figures on the BTC side in this event—the BTC drawdown, depth evaporation, spread, and duration—are compiled from exchanges' historical candlestick and liquidation data (the exchange-side price-protection mechanism is detailed in Section 20.3.4 and the Deribit report cited there); they are anchored here once and referenced back in later passages, while the ETH-side drawdown figures are first given in Section 20.3.4. This is precisely what the volatility-liquidity spiral looks like once volatility crosses its critical point and spins out of control.

20.2.4 Cross-asset spillover effects

The impact of the volatility-liquidity spiral is not confined within a single asset; it is also strongly contagious across assets and across markets. When market makers suffer inventory losses on a core asset or face extreme volatility, portfolio-level risk-control directives force them to withdraw their supply of liquidity on other assets simultaneously. This contagion unfolds along four dimensions—cross-asset, cross-exchange, cross-layer, and cross-product—amplifying a local volatility shock into a systemic liquidity crisis. Section 20.7 provides a detailed analysis of these four contagion dimensions.

20.2.5 A mathematical characterization of the spiral

The mechanism of the volatility-liquidity spiral can be formalized through the mathematical properties of the market maker's profit function. The formalizations in this section and throughout the chapter are illustrative specifications that characterize the qualitative structure of a mechanism; their parameters are not calibrated to real data, and the associated numbers should be read as illustrations of the mechanism rather than empirical estimates. The market maker's profit function can be simplified into a key trade-off. Let the market maker's profit at time t be:

$$\Pi_t = \text{spread} \times \text{volume} - \text{inventory cost} - \text{adverse-selection cost}$$

Here the relationship between inventory cost and volatility is markedly convex. Let inventory cost be $C_{inv} = \alpha\sigma^2 Q^2$, where σ is realized volatility, Q is the inventory quantity, and α is a dimensional risk-aversion coefficient (used to normalize $\sigma^2 Q^2$ into monetary units). Here α, σ, and Q are all illustrative specifications, not calibrated to real volatility and inventory data. The implication of this functional form is profound: when volatility rises from σ_0 to $2\sigma_0$, inventory cost rises from $\alpha\sigma_0^2 Q^2$ to $4\alpha\sigma_0^2 Q^2$, a fourfold increase. This factor of four reflects the quadratic convexity of σ^2, not an exponential relationship. This nonlinear cost structure means that when volatility doubles, the market maker must maintain profit by sharply raising spreads or sharply reducing inventory.

In the Bitcoin flash crash of May 19, 2021, this mechanism was especially evident: that day, realized volatility surged from a daily 3% to more than 25% in just six hours, and spreads and depth accordingly went through a three-stage illustrative deterioration (its staged figures are the May 19 definitions anchored in Section 20.2.3; see also Figures 20-3 and 20-4), with most market makers withdrawing from the market entirely. This staged deterioration is precisely the rational response of market makers to steadily rising inventory costs.

More important, this process exhibits a clear "critical point" phenomenon. As the spread widened from 10 bps to 15 bps, trading volume did not decline noticeably. But once the spread breached 20 bps, volume began to shrink sharply, which caused market makers' total profit to fall instead, further incentivizing them to withdraw liquidity. This "abrupt" behavioral shift is the direct manifestation of the profit equation turning from positive to negative somewhere within a certain volatility band. Mathematically, this corresponds to a rapid (non-smooth) transition of the market maker's profit equation from positive to negative over a certain volatility range. Because the expression

for Π_t above is not given an explicit differentiable functional form, this section makes no formal claim about whether that transition is a discontinuous jump in the strict sense.

20.3 The leverage-liquidation-liquidity cascade

Of the three positive-feedback loops, the liquidation cascade produces the most direct shock. Liquidation not only consumes liquidity directly, by devouring order-book depth, but also destroys liquidity indirectly, by driving out market makers. The superposition of these two effects creates a positive-feedback loop far more violent than the volatility-liquidity spiral, leaving the market highly prone to a nonlinear collapse when shocked. Unlike the volatility-liquidity spiral discussed above, the destructiveness of the liquidation cascade lies in its compulsory and irreversible nature: once set in motion, it is hard to stop through traders' voluntary adjustments. The liquidation cascade is also distinctive in that it converts microscopic individual risk into macroscopic systemic risk, so that the liquidation of a single trader can trigger a chain reaction that ultimately drains liquidity from the entire market.

20.3.1 The endogeneity of leverage

Traditional market-microstructure theory tends to treat a trader's leverage level as an exogenous variable—a fixed multiple chosen at the time a position is opened. In perpetual futures markets, however, the system's effective leverage is highly endogenous: it is a dynamic function of the price level, open interest, margin balance, and liquidation threshold. This endogeneity makes the market increasingly fragile as prices fall, forming a hidden risk-amplification mechanism.

The endogeneity of effective leverage shows up in its procyclical character. When prices rise, the unrealized profit on long positions increases, margin balances become ample, and the system's average effective leverage falls, so the market grows safer. Conversely, when prices fall, the margin on long positions is eroded and effective leverage climbs passively. This mechanism means that as prices fall, the number of positions drawing ever closer to the liquidation trigger line increases at an accelerating rate.

Concretely, suppose a trader goes long 1 BTC of perpetual futures at \$50,000 with an initial margin of \$10,000 (initial leverage of 5x) and a maintenance-margin rate of 5%. When the price rises to \$55,000, an unrealized profit of \$5,000 raises account equity to \$15,000, and effective leverage falls from 5x to 3.67x (\$55,000/\$15,000), markedly improving the margin cushion. But when the price falls to \$45,000, an unrealized loss of \$5,000 erodes half the initial margin, account equity shrinks to \$5,000, and effective leverage climbs passively to 9x (\$45,000/\$5,000). If the price falls further to \$42,000, account equity is only \$2,000 while the maintenance-margin requirement is \$2,100 (\$42,000 $\times$ 5%), and the position will be forcibly liquidated. This passive rise in leverage is not the trader's active choice but the algebraic result of the market state.

Figure 20-6 shows how, under a fixed initial margin (\$10,000) and 5x initial leverage, a price decline causes effective leverage to rise nonlinearly and approach the liquidation threshold.

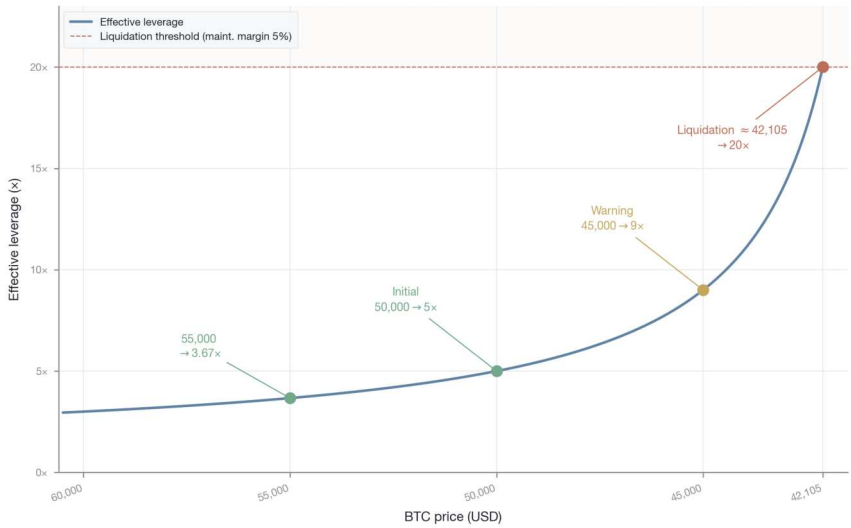


Figure 20-6. The procyclicality of leverage and the nonlinear rise in liquidation risk (Data source: derivation from a theoretical model)

This procyclicality of leverage explains why a seemingly mild initial price shock can evolve into a catastrophic market collapse. In the early phase of a price decline, the rise in system leverage mainly absorbs traders' buffer capital; but once the price crosses a certain critical point, large numbers of high-leverage positions hit their liquidation lines simultaneously, converting potential demand for liquidity into an actual market shock. This procyclical dynamic of "accumulating leverage in the bull market and releasing it all at once in the bear market" is precisely the core proposition characterized by Geanakoplos's (2010) [13] theory of the leverage cycle: in a rising asset-price phase, collateral constraints loosen and leverage climbs endogenously; in a falling phase, shrinking collateral values force deleveraging to occur all at once, amplifying price swings. Brunnermeier and Pedersen (2009) [4], in their classic liquidity-spiral model, further point out that the procyclicality of margin requirements is one of the core mechanisms that cause market liquidity to dry up suddenly. Their model shows that when funding costs rise (that is, when margin requirements increase), market makers' funding constraints tighten, which reduces their supply of liquidity, forming a self-reinforcing vicious cycle. In perpetual futures markets, this mechanism is amplified through the nonlinear character of the liquidation trigger line.

The endogeneity of leverage also means that high-leverage positions accumulated in a bull market become a concentrated source of systemic-risk release in a bear market. Many traders steadily increase their leverage multiple during a sustained price rise, because their margin cushion keeps improving and their risk perception is dulled. But once the price reverses, these positions go from "safe" to "dangerous," and from "profitable" to "loss-making," in a very short time. The speed of this state change often exceeds traders' ability to react, resulting in passive liquidation.

20.3.2 Inelastic demand for liquidity

In a normal trading environment, demand for liquidity is elastic: traders can adjust order size or defer trading according to market depth and execution costs. The liquidation mechanism, however, converts this elastic demand into perfectly inelastic, mandatory demand for liquidity, thereby fundamentally shocking the market structure. This transformation is one of the most fundamental differences between perpetual futures markets and traditional markets.

The special nature of liquidation orders stems from their execution mechanism. When a position's margin rate falls below the maintenance-margin requirement, the liquidation engine takes over the position and pushes it into the market as a market order or an aggressive limit order. This mode of execution strips the trader of any freedom to choose timing or price; the order must be filled immediately at whatever price is currently available. As the price falls rapidly, liquidation volume grows at an accelerating rate because of leverage endogeneity, causing more and more inelastic sell orders to pour into the order book.

The concentrated eruption of inelastic demand for liquidity directly causes a nonlinear amplification of price impact. Kyle's (1985) [10] price-impact model shows that the sensitivity of the market price to order flow—the price-impact coefficient λ—is inversely proportional to market depth. When enormous liquidation orders pour in over an extremely short interval, they not only consume the existing liquidity reserve but also force the market to clear at an extreme discount, thereby creating the conditions for the next round of liquidation. Kyle's linear price-impact model assumes that market makers cannot distinguish informed from uninformed order flow; as Section 20.3.3 discusses, however, liquidation orders have identifiable mechanical features. This means the actual price impact in a liquidation cascade may exhibit nonlinear (convex) features that exceed the linear model's predictions.

Compared with traditional futures markets, the liquidation mechanism of perpetual futures is more aggressive: as described in Section 20.1.3 and Table 20-3, traditional exchanges grant a margin-call buffer period when margin is insufficient, whereas perpetual liquidation executes the moment it is triggered, leaving no time to react. This difference makes liquidation flow in perpetual markets markedly higher than in traditional markets during stress periods, exerting greater shock on liquidity.

The compulsory nature of liquidation also means that the arrival of liquidation orders is unpredictable. Market makers cannot know in advance when a large volume of liquidation orders will pour in, so they must prepare for the worst case. This uncertainty further raises the risk cost of market making, making market makers more inclined to shrink their supply of liquidity during high-volatility periods.

20.3.3 The dual-shock mechanism

The damage that a liquidation cascade does to market liquidity is not a single-dimensional consumption of depth; it is a dual shock composed of depth devouring and market-maker expulsion. The interaction of these two effects explains why, under extreme conditions, liquidity disappears far faster than a linear model predicts.

The first shock is the direct devouring of depth. The market orders generated by the liquidation engine consume the limit orders on the book layer by layer in price-priority order. Because liquidation orders are typically far larger than ordinary trades, they can easily break through the first and second layers of liquidity, producing significant slippage. This physical consumption of depth not only reduces the market's capacity to absorb subsequent shocks but also directly depresses the mark price, thereby pushing still more previously safe positions into the liquidation zone.

The second shock is the active retreat of market makers. As noted earlier, market makers face enormous adverse-selection risk when confronting extreme volatility and one-directional order flow. The toxicity of liquidation orders differs from the informational toxicity that informed traders bring in traditional market microstructure: liquidation orders themselves carry no private information about the asset's fundamentals; their toxicity stems from the extreme adverse-selection risk that their compulsory and unpredictable nature imposes on market makers. When a market maker's algorithm detects a large influx of liquidation orders, it cannot distinguish whether these orders are uninformed forced closures or informed directional selling. To protect their own capital, market makers respond rationally by widening the bid-ask spread or pulling their quotes outright. This behavior causes available liquidity to evaporate actively at the very moment it is most needed. This "liquidation → selling → further price decline → more liquidation" fire-sale spiral is isomorphic to the fire-sale mechanism characterized by Shleifer and Vishny (1992) [14]: when forced sellers clear positions in concert while the natural buyers (peers with the valuation capacity and balance sheet to absorb them) are equally constrained, assets trade at liquidation prices far below fundamentals, and liquidation value and debt capacity depress each other, forming positive feedback.

To quantify the compounded effect of the dual shock, Figure 20-7 contrasts the predictions of a single-factor model that considers only the depth-devouring effect with those of a two-factor model that also incorporates the market-maker-retreat effect on liquidity consumption.

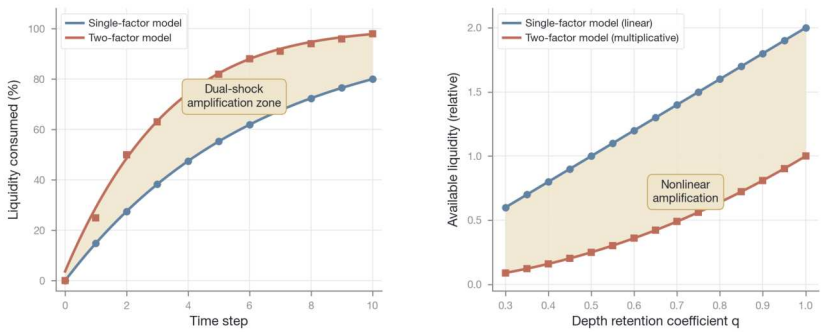


Figure 20-7. The dual-shock mechanism of liquidity consumption: the multiplicative effect of depth devouring and market-maker retreat (Data source: derivation from a theoretical model)

The superposition of the dual shock produces a marked multiplicative effect. On one hand, liquidation orders physically consume order-book depth; on the other, the retreat of market makers reduces the replenishment of new liquidity. This two-way deteriora-

tion of supply and demand causes the market's effective liquidity to fall sharply (superlinearly), ultimately forming a price gap. From a mathematical standpoint, let initial available liquidity be L_0, let the depth-consumption factor $d \in [0, 1]$ denote the proportion of order-book depth remaining after liquidation orders are consumed, and let the market-maker-retention factor $m \in [0, 1]$ denote the proportion of market makers still providing quotes. If an analyst considers only the larger of the two single effects, they might use $\min(d, m)$ as the benchmark for assessing liquidity. The actual effect of the dual shock, however, is multiplicative: $L = L_0 \cdot d \cdot m$. For example, **under the assumption that d and m are mutually independent**, when depth consumption is 40% ($d = 0.6$, illustrative example) and market-maker retreat is 40% ($m = 0.6$, illustrative example), taking only a single effect would overestimate remaining liquidity at 60% of the initial level, whereas the multiplicative model gives 36%. This 36% is the **worst-case lower bound under the independence assumption (the theoretical upper bound on losses)**, not a measured value. In fact, real exchanges' liquidation engines implement a variety of buffering mechanisms, including tiered liquidation (first attempting to fill via limit orders and converting to market orders only on failure) and smart order-splitting strategies, so that a negative dependence exists between the two factors: more refined liquidation execution reduces the depth impact of any single liquidation, which in turn slows the pace of market-maker retreat. Therefore, the multiplicative model describes a worst-case theoretical lower bound under the assumption that d and m are independent, while the actual effect (milder, because d and m are negatively dependent) depends on the specific implementation of the liquidation engine. The earlier references to "underestimation" and "underestimated by nearly half" should be understood as the gap between the upper bound under this independence assumption and the single-factor estimate, not as a measurement of the true liquidity loss.

Another consequence of market-maker retreat is that it destroys the market's price-discovery function. Under normal conditions, market makers help the market discover the equilibrium price by providing liquidity on both sides of the book. In a liquidation cascade, however, the retreat of market makers causes a severe asymmetry in liquidity on the two sides, and the market cannot conduct effective price discovery. In this situation, price movements no longer reflect fundamental information but are purely liquidity-driven.

20.3.4 The nonlinear cascade and state transitions

The most salient feature of the liquidation cascade is its state-transition-like nonlinearity. Faced with shocks of different magnitudes, the market exhibits sharply different microstructural responses. This nonlinearity explains why there is a marked state transition between "just short of triggering a cascade" and "just barely triggering a cascade." This chapter here borrows the term "phase transition" from physics as an analogical description of this transition, and gives a falsifiable criterion for its core assertion rather than treating it as a proven mathematical property.

When liquidation volume is below a certain critical threshold, the market's existing liquidity is sufficient to absorb the forced sell orders, price impact is confined to a

local range, and the system can quickly return to equilibrium. Once liquidation volume exceeds that threshold, however, the existing depth is completely broken through, the price is pushed to the next liquidation line, and new liquidation orders are triggered. At this point the positive-feedback loop is activated, and the system transitions rapidly from a controllable state to an uncontrollable one. The existence of this transition means that market risk is not smoothly distributed but concentrated near certain critical bands. This "phase-transition" hypothesis can be falsified as follows: if, as liquidation volume increases continuously, both price impact and available depth change continuously and smoothly, without any abrupt change (jump-like deterioration) in depth and price impact appearing near a particular liquidation-volume level, then for that event a phase transition in the sense of this section does not hold—it degenerates into a steep but continuous response curve.

The crypto-market flash crash of May 19, 2021 is a typical case of this nonlinear cascade. Over the full span of less than 12 hours, the Ethereum price plunged more than 46% and Bitcoin fell about 32% cumulatively (this is the cumulative drawdown over the full span; the \$43,000→\$30,000 move and the drop of more than 30% recorded in Section 20.2.3 refer to the interval drawdown within the roughly six-hour window in which the collapse in spreads and depth was most violent, and the two definitions are mutually compatible). According to Deribit's market-analysis report [15], extreme network-wide liquidation volume caused severe liquidity depletion, prompting exchanges to trigger price-protection and rate-limiting mechanisms. In this event, Deribit's price-protection mechanism was triggered when the ETH index price moved rapidly [15]. Unlike the circuit breakers of traditional markets, Deribit's protective measures did not halt trading entirely but instead slowed the response speed of the liquidation engine by limiting the rate of change of the mark price, buying market participants a window of time to reassess. Even with such a protective mechanism, however, the speed of liquidity evaporation still exceeded the expectations of most market participants. The BTC drawdown, depth evaporation, and duration in this passage are the reference-backs to the figures anchored in Section 20.2.3, while the ETH drawdown is first given here. This event shows that when liquidation volume breaches the critical band, the speed and scale of liquidity evaporation largely decouple from fundamental logic.

Figure 20-8, taking the dynamic relationship between liquidation volume and available liquidity as its through-line, shows the full path by which a liquidation cascade accelerates nonlinearly after breaching the critical threshold and ultimately produces a liquidity vacuum.

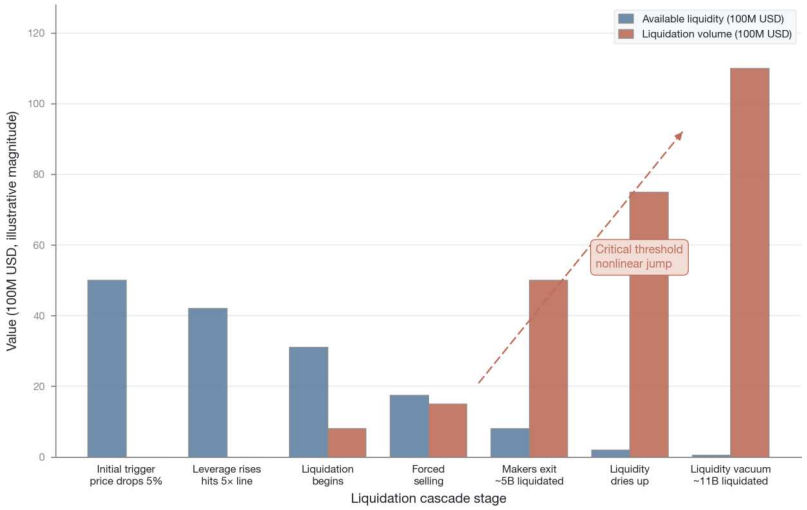


Figure 20-8. The nonlinear development of the liquidation cascade and the formation of a liquidity vacuum (Data source: derivation from a theoretical model)

To mitigate such extreme shocks, exchanges have introduced insurance funds and auto-deleveraging mechanisms. The insurance fund is designed to absorb bankruptcy losses near the liquidation threshold and prevent the spread of systemic risk. The capital scale of the insurance fund, however, is often severely inadequate relative to potential liquidation losses. Binance, for example, had an insurance fund of about $1 billion in 2023, whereas single-day network-wide liquidation volume on May 19, 2021 once exceeded $8 billion; if one takes as reference the extreme single-day network-wide liquidation demand on the same basis, a single exchange's insurance-fund coverage ratio is under 20% (here the numerator is a single exchange's fund and the denominator is network-wide liquidation volume, so the dimensions are not strictly aligned; this is used only to illustrate the vast disparity between fund size and systemic liquidation demand, and the two figures come respectively from exchange disclosures and third-party liquidation statistics, at different points in time). More important, the mechanism by which the insurance fund accumulates is itself procyclical: it builds up gradually through liquidation surpluses in calm markets and is consumed all at once in a crisis, exactly opposite to the direction of demand. In an extreme liquidation cascade, when enormous losses exhaust the insurance fund, the exchange is forced to activate the auto-deleveraging mechanism, forcibly closing profitable positions to cover the losses. Although this mechanism guarantees the system's solvency, by forcibly stripping profitable traders of their positions it further destroys the stability of market expectations and intensifies the systemic retreat of liquidity.

The activation of the auto-deleveraging mechanism marks the market's entry into the deep phase of the liquidity vacuum. By this point the market's microstructure has completely collapsed, the price-discovery mechanism has failed, and a vast price gap has opened on the order book. What traders face is no longer a matter of liquidity cost but a fundamental difficulty of execution. This phase is characterized by the fact that

even a trader willing to clear a position at an extreme price may be unable to find a counterparty. This is the essential definition of a liquidity crisis.

20.4 Funding-rate distortion and position squeezes

Extreme funding rates cause a structural imbalance in liquidity by forcibly altering the position structure. This is the third positive-feedback loop unique to perpetual futures, in which the funding-rate mechanism—originally a tool for price anchoring—is transformed into a catalyst that destroys the supply of liquidity under extreme market conditions. Under normal conditions, the funding rate maintains the convergence of the perpetual and spot prices through a mild arbitrage incentive; but under stress, extreme rate levels trigger large-scale one-directional demand to close positions while simultaneously driving out market makers, forming a self-reinforcing liquidity crisis.

20.4.1 The distortion of the rate signal

Under normal conditions, the funding rate reflects the relative strength of the long and short forces in the market, but under extreme market conditions it turns into a structural factor that distorts the directional distribution of liquidity. The settlement frequency of the funding rate differs markedly across exchanges and is undergoing rapid evolution. Early perpetual futures commonly used a discrete mechanism settling once every 8 hours, but Binance, Bybit, and OKX have successively launched product lines settling every 4 hours or even every 1 hour, while on-chain perpetual futures platforms use near-continuous, high-frequency settlement. Higher settlement frequency in theory lowers the absolute rate value of any single settlement, thereby lessening the concentrated closing shock at the moment of settlement. This improvement, however, does not necessarily eliminate the rate-driven fragility of liquidity: higher-frequency settlement means the rate responds more sensitively to the market state, and in an extreme one-sided market this can cause the rate to remain persistently at an extreme level rather than delivering an intermittent shock. The following analysis proceeds on the basis of a discrete settlement mechanism, but its core logic—directional liquidity imbalance driven by extreme rates—holds under different settlement frequencies. Under normal conditions, market makers provide roughly symmetric liquidity on both sides of the order book, keeping the trading environment friendly to both longs and shorts. When the market enters a one-sided extreme regime, a violent surge in the funding rate breaks this symmetry and triggers a severe mismatch between the supply of and demand for liquidity.

During periods of extreme positive rates, holders of long positions face high carrying costs. This cost pressure prompts longs to close in concert, which means the sell-side demand for liquidity in the market rises sharply. At the same time, however, market makers face markedly intensified inventory skew and adverse-selection risk. To avoid accumulating an excessive long inventory by absorbing large volumes of long-closing flow, market makers rationally reduce or even withdraw their sell-side supply of liquidity. This severe mismatch between the willingness to buy and to sell causes liquidity to become directionally distorted: sell-side depth dries up while buy-side depth may be relatively ample. Conversely, during periods of extreme negative rates, holders

of short positions are forced to pay high rates, causing shorts to close in concert by buying, and buy-side demand for liquidity surges. The same logic drives market makers to shrink their buy-side supply of liquidity, causing buy-side depth to fall sharply.

There is a negative relationship between the funding rate and available liquidity, and this deterioration typically accelerates in the high-rate range. Specifically, when the absolute value of the funding rate is in a mild range, market makers can absorb the risk by adjusting spreads, and liquidity declines relatively gently; but when the rate enters a high range that portends an imminent wave of one-directional liquidation, market makers' profit-equation expectations turn negative, and they choose to pull their orders outright rather than merely widen spreads, so the deterioration in liquidity noticeably accelerates. Whether an estimable critical rate threshold exists, and how the deterioration rate compares on either side of that threshold, is an empirical question that can be tested with bucketed data; what this chapter offers is a qualitative direction rather than a calibrated threshold. This retreat is not only one-directional but is often carried out by multiple market makers simultaneously, instantly manufacturing a liquidity vacuum in a particular direction.

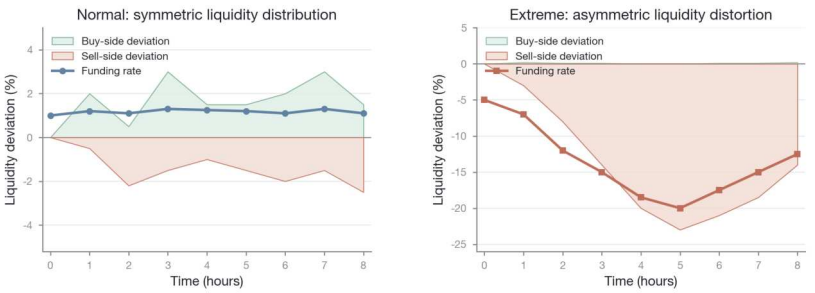


Figure 20-9. The relationship between the funding rate and the directional distribution of liquidity

The left panel of Figure 20-9 shows the symmetric two-sided quoting under normal conditions, and the right panel shows the asymmetric distortion under extreme negative rates, with buy-side liquidity shrinking sharply and sell-side orders piling up passively. This distortion means that closing a short will face far higher slippage costs than normal, whereas opening a long is unusually easy, further intensifying the one-sided imbalance in the market.

20.4.2 Position squeezes

A position squeeze is the direct consequence of an extreme funding-rate environment, inflicting severe shock on market liquidity through compulsory directional trading demand. Position squeezes fall into long squeezes and short squeezes, which are markedly asymmetric in their trigger mechanisms and market impact. This asymmetry stems from at least three levels. First, the physical asymmetry of liquidation direction: long liquidation manifests as selling, and the price can approach zero; short liquidation manifests as buying, and the price has no theoretical upper bound, so a short squeeze has greater price-impact potential in the extreme case. Second, the asymmetry in the structure of market participants: retail investors in crypto markets are generally biased

toward going long, so long crowding is systematically higher than short crowding, which makes a long squeeze affect a broader base of participants and propagate panic faster. Third, the asymmetry of the funding-rate formula: most exchanges impose upper and lower bounds on the rate, and some exchanges' premium-index algorithms depend on order-book depth, so that when depth is thin the extreme values of the premium index may be amplified, thereby magnifying the distortion of the rate signal (the specific strength of this effect depends on each exchange's premium-index implementation). On balance, a long squeeze is typically accompanied by a rapid price collapse, whereas a short squeeze manifests as a violent upward price surge.

Short squeezes are especially prominent in perpetual futures markets, and their core mechanism lies in the persistent loss effect of negative funding rates on short positions. After the market has experienced a sharp decline, large numbers of speculative shorts pour in and push the perpetual price down, causing the funding rate to turn deeply negative. This extreme negative rate not only increases the carrying cost of shorts but also attracts arbitrageurs to buy the perpetual and sell the spot. When the price stabilizes or rebounds slightly, some high-leverage shorts are forced to close by buying, and this compulsory buy-side demand for liquidity, in an environment lacking market-maker sell-side supply, rapidly drives the price up, thereby triggering more shorts' stop-losses and liquidations and forming a chain of short squeezes.

The Bitcoin flash crash of February 22 to 23, 2021 provides a typical squeeze case. At the time, the Bitcoin price plunged from about \$58,000 to about \$45,000 over roughly two days, a drop of more than 22% (according to the post-mortem by the Perpetual Protocol team [16], which dates the flash crash to February 21; this chapter records it as February 22 to 23 by the event's duration). This extreme move stemmed not from fundamental news but from the liquidation of a few large high-leverage positions; here the attribution to "a few large positions triggering it" is the post-mortem attribution in [16], while the price range is public market data, and the two come from different sources. As the price fell, the funding rate changed sharply, and market liquidity became extremely scarce because of market-maker retreat. When the price bottomed, extreme negative funding rates and an oversold condition triggered a violent short-covering rally and an influx of arbitrage flow. Because the sell-side depth on the order book had already been cleared, the compulsory buy-to-close orders could only be filled at extremely high slippage, causing the price to "squeeze" upward at great speed. This squeeze destroyed not only directional traders but also inflicted enormous adverse-selection damage on market makers attempting to provide liquidity.

20.4.3 The stabilization limit of arbitrageurs

Funding-rate arbitrageurs are usually regarded as the natural stabilizers of perpetual futures markets. Through basis trades—establishing offsetting positions between the spot market and the perpetual futures market—they earn funding-rate income and, in doing so, objectively promote the convergence of the perpetual price toward the spot price. This stabilizing role, however, is not unlimited; when market stress exceeds their capacity limit, arbitrageurs not only lose their stabilizing function but may even become a factor that intensifies liquidity stress.

The stabilizing force of arbitrageurs stems from their sensitive response to the funding rate. When the funding rate is markedly positive, arbitrageurs buy the spot and short the perpetual, which provides sell-side liquidity to the perpetual market and suppresses excessive premiums; when the rate turns negative, they close or reverse, providing buy-side liquidity. This mechanism works well under normal conditions and effectively dampens price deviations. The capacity of arbitrage capital, however, is finite. When the market suffers an extreme shock and one-directional speculative demand far exceeds the absorptive capacity of arbitrage capital, the funding rate breaks through the arbitrageurs' dampening band and enters an extreme state.

The critical condition for arbitrageurs to retreat is typically determined jointly by a sharp reversal of the funding rate and the depletion of spot-market liquidity. When many basis traders hold "long spot plus short perpetual" positions to collect a positive funding rate, if the market suddenly plunges and drives the rate deeply negative, these originally positive-yielding positions instantly turn into persistently loss-making ones. To stop losses, arbitrageurs are forced to simultaneously buy to close their shorts in the perpetual market and sell to close their longs in the spot market. This large-scale influx of same-direction orders not only creates enormous buy-side demand for liquidity in the perpetual market but also forms heavy selling pressure in the spot market. When all arbitrageurs try to retreat through a narrow liquidity exit at the same time, a concentrated-exit effect inevitably occurs, and the originally stabilizing mechanism turns into an amplifier of volatility.

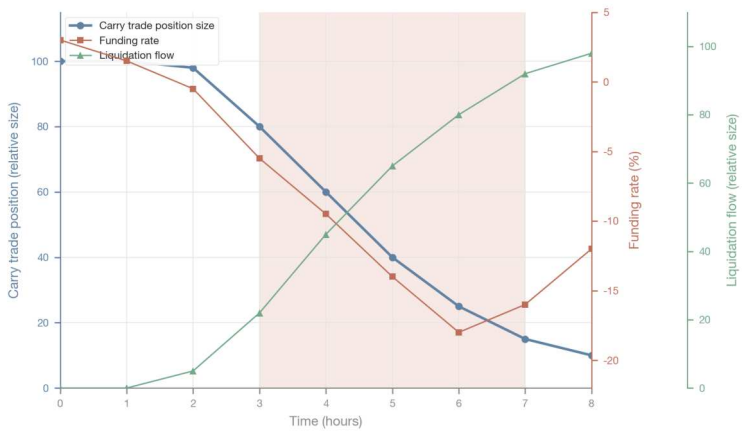


Figure 20-10. The time series of basis-arbitrage position squeezes (illustrative characterization)

As shown in Figure 20-10, as the funding rate moves from positive to an extreme negative value, the scale of basis-arbitrage positions falls sharply, accompanied by a synchronous rise in liquidation flow. This dynamic process clearly shows how the concentrated exit of arbitrageurs turns a stabilizing force into a destructive one: as the rate reverses, originally dispersed closing demand converges within a short time into a concentrated one-directional liquidity shock, accelerating the market's further decline.

In addition, the discrete settlement mechanism of the funding rate can also create a liquidity gap at particular moments. This settlement-related liquidity gap is seen

mainly in low-frequency contracts that settle discretely every 8 hours; under the every-1-to-4-hour or near-continuous on-chain settlement regimes described earlier, the rate exposure of any single settlement is smaller and this effect is markedly weaker. In the final minutes before a rate settlement, traders who expect to pay a high rate close positions in concert to avoid the payment, causing an instantaneous surge in demand for liquidity. Meanwhile, market makers sharply reduce their quoted depth to avoid bearing unfavorable inventory and rate exposure at the moment of settlement. This instantaneous divergence of supply and demand forms a distinct liquidity-vacuum period around settlement, which can readily be exploited by malicious traders to trigger artificial price spikes and liquidations.

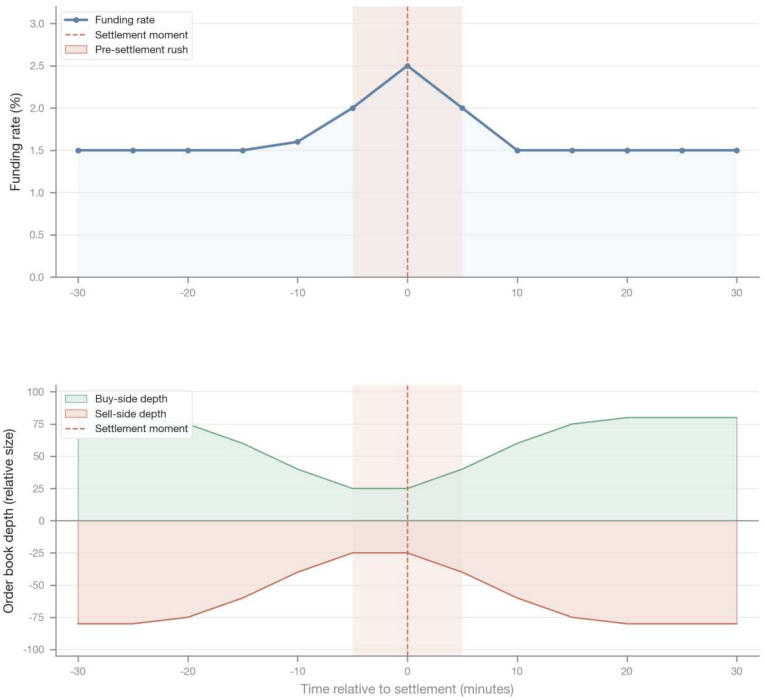


Figure 20-11. The liquidity gap at the moment of funding-rate settlement (seen mainly in low-frequency discrete contracts settling every 8 hours)

Figure 20-11 presents the liquidity gap around settlement: in the minutes before settlement, rate-avoiding closing demand and preemptive market-maker retreat overlap, and depth falls sharply to form a brief vacuum; after settlement completes, market makers quickly replenish, and liquidity returns to normal within minutes. This periodic pulse is a structural by-product of the low-frequency discrete settlement regime of perpetual futures and is insignificant under high-frequency or continuous settlement regimes.

20.4.4 Loop coupling and compound crises

A systemic liquidity collapse in perpetual futures markets is rarely triggered by a single mechanism; it is instead the result of the mutual coupling and cross-amplification of the three positive-feedback loops—the volatility spiral, the liquidation cascade, and

funding-rate distortion. This kind of compound crisis exhibits strong nonlinearity: in the early phase, a market shock may activate only one of the loops, and the system still retains some capacity for self-repair; but once the shock breaches the critical point and all three loops fire at once, the multiplier effect among them causes liquidity to evaporate completely in an instant.

The formation of a compound crisis follows a clear transmission chain. First, an external shock or a large sell order causes the price to plunge, realized volatility rises, and market makers begin to retreat because of the convex rise in inventory costs, setting off the negative spiral between volatility and liquidity. As the price falls further, the system's endogenous leverage climbs passively and hits the liquidation lines of large positions. Compulsory market liquidation orders pour in, not only directly devouring the already-thin order-book depth but also, through the extreme toxicity they generate, driving out the remaining market makers entirely, so that the liquidation cascade erupts in full. In this process, the extreme deviation of the price deeply distorts the funding rate, the arbitrageurs who originally played a stabilizing role are forced to close, and the concentrated exit of basis trades releases into the market a final wave of one-directional liquidity demand that cannot be absorbed. These three mechanisms overlap heavily in time and are causally intertwined, jointly pushing the market into a state of liquidity vacuum.

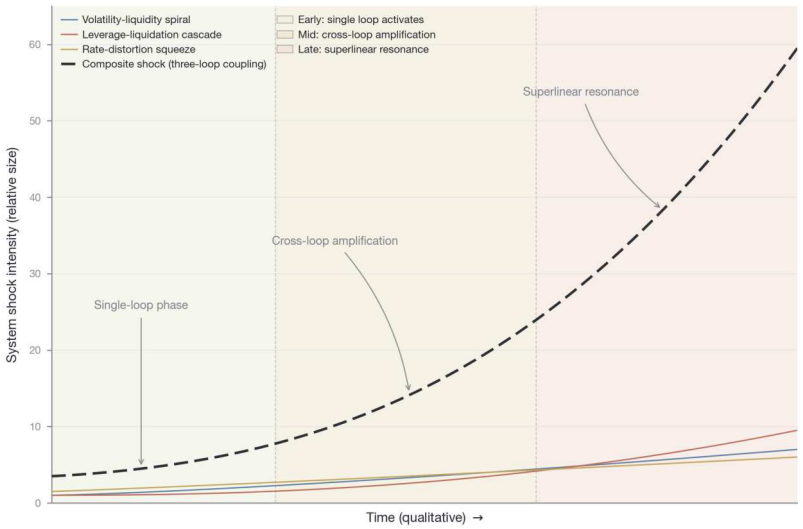


Figure 20-12. The coupling of the three positive-feedback loops and the nonlinear character of a compound crisis

Figure 20-12 reveals the nonlinear character of a compound crisis: in the early phase, the loops develop independently and shock intensity is approximately linear; in the middle phase, the loops cross-amplify (rising volatility simultaneously accelerates liquidation and rate distortion); in the late phase, all three loops resonate fully, the multiplier effect makes shock intensity grow superlinearly, and the market transitions rapidly from a controlled state to complete collapse in a very short time. The "transition" here follows the analogical usage of the liquidation-cascade state transition in Section 20.3.4

and can likewise be tested empirically by "whether shock intensity changes abruptly as the driving variable changes."

The Terra-LUNA collapse of May 2022 is the most extreme empirical instance of this compound-crisis coupling [17]. This event contained two coupled but mechanistically distinct spirals: first, the credit-collapse spiral of an algorithmic stablecoin (UST depeg → LUNA issuance dilution → LUNA selling → further UST depeg), a monetary-level crisis stemming from a design flaw in UST's mint-redeem mechanism; and second, the endogenous liquidity spiral of perpetual futures (LUNA price decline → liquidation cascade → market-maker retreat → liquidity vacuum → further price decline), which is this chapter's core mechanism. Their coupling point is this: the perpetual-futures liquidation cascade drove the LUNA spot price down further, which in turn accelerated the death spiral of the algorithmic stablecoin. The following focuses on the second spiral and its coupling with the first. The trigger was a mild depeg caused by a large sell-off of UST in a Curve liquidity pool; this initial shock raised implied volatility and prompted market makers to widen spreads. As Luna Foundation Guard failed to defend the peg by selling its Bitcoin reserves, market confidence deteriorated sharply and the LUNA price plunged [18], rapidly triggering large-scale liquidation cascades in lending protocols and perpetual futures markets. At the same time, the market's extreme bearishness toward LUNA drove its perpetual-futures funding rate to extreme negative values, which not only failed to attract arbitrageurs to buy but instead forced all cross-market basis traders to take their losses and exit. Surging volatility drove out market makers, relentless liquidation devoured all bids, and rate distortion destroyed the last arbitrage buffer; under the multifactor resonance of all three loops, LUNA's apparent and true liquidity fell to zero simultaneously, its price fell more than 99.9% within just a few days, and the shock reverberated systemically across the entire crypto market [19].

20.5 The split between apparent and true liquidity

Order-book depth and the bid-ask spread are the two most intuitive metrics for assessing market microstructure. In normal markets, these metrics accurately reflect the expected cost that a trader will incur in executing a large order. When the market enters a state of stress, however, these conventional metrics systematically overestimate the true available liquidity [20]. This overestimation is not measurement error but a structural feature of modern high-frequency market-making mechanisms. This section examines the mechanism by which the wedge between apparent and true liquidity forms, revealing why seemingly ample liquidity can suddenly disappear at the very moment the market needs it most.

20.5.1 Ghost liquidity

The order books of modern crypto-asset exchanges are highly dynamic, with order submission and cancellation occurring at frequencies far exceeding the frequency of actual fills. This dynamism is driven largely by high-frequency market makers' algorithms, so that the liquidity depth displayed on the order book contains a large volume of conditional quotes. Degryse, de Winne, Gresse, and Payne (2018) termed this phenomenon

ghost liquidity [21]. Ghost liquidity refers to limit orders that are visible on the order book but are rapidly withdrawn once actual trading demand arrives or when market conditions change slightly. Degryse et al.'s original definition focused on repeated quoting across trading venues followed by immediate cancellation after a fill; building on that cross-venue sense, this chapter extends the concept to millisecond-scale conditional cancellation within a single venue.

The existence of ghost liquidity stems from market makers' defense against adverse-selection risk. In a continuous two-sided quoting mechanism, a market maker that provides liquidity is also exposed to the risk of being "picked off" by informed traders. To balance this risk, high-frequency market makers' algorithms continuously send orders to the market to probe other participants' intentions and the genuineness of market depth. When the market is calm, these probing orders rest on the order book and constitute thick apparent depth. These orders, however, are not unconditional commitments. Once volatility rises or the accumulation of one-directional order flow is detected, the algorithm cancels these quotes within milliseconds. Identifying such "toxic" order flow and dynamically shrinking quotes accordingly is precisely the core of high-frequency market makers' risk control; the volume-synchronized probability of informed trading (VPIN) proposed by Easley, López de Prado, and O'Hara (2012) [22] is a representative tool for measuring order-flow toxicity, and its rapid rise in high-frequency environments often precedes an abrupt contraction in liquidity.

This conditional-quoting strategy causes a severe split between apparent and true liquidity. Figure 20-13 shows the dynamic relationship between order-book depth and actual executable depth before and after a crisis trigger.

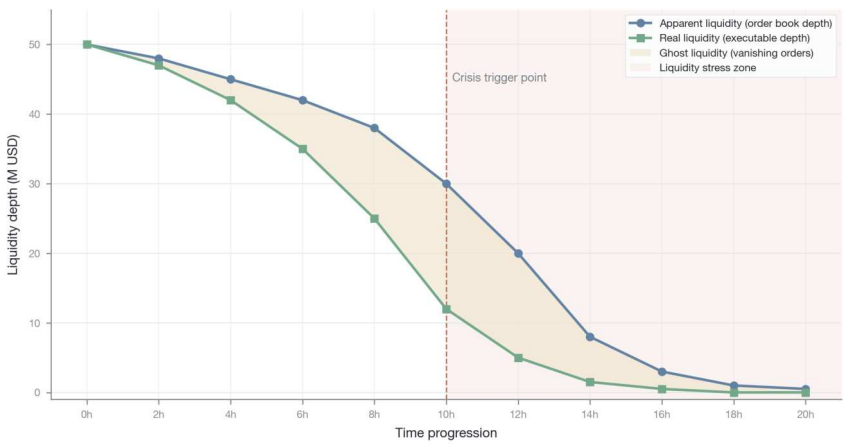


Figure 20-13. The evolution of apparent versus true liquidity before and after a stress event (an illustrative reconstruction by the author, based on public market data, of BTC perpetual depth during the FTX collapse; not data from [15])

Under normal conditions, apparent liquidity and true liquidity overlap closely, and conventional liquidity metrics are reliable. But when the crisis trigger arrives, apparent liquidity begins to decline yet still holds at a certain level, creating the illusion that the market still has absorptive capacity. At the same time, true liquidity plunges abruptly. The vast gap between the two is ghost liquidity. This split means that price-

impact models computed from historical order-book depth will severely underestimate the actual execution cost in extreme conditions, causing risk-management systems to fail.

20.5.2 Hidden liquidity

The opposite of ghost liquidity is a layer of invisible liquidity that also exists in the market: hidden orders and conditional orders. If ghost liquidity is the illusion of being "visible but not executable," then hidden liquidity is a reserve of liquidity that is "invisible but executable." Research by Bessembinder, Panayides, and Venkataraman (2009) shows that the use of hidden orders rises markedly as market volatility, order size, and relative depth increase [23].

The hidden-order mechanism allows a trader to submit a large limit order to the exchange while displaying only a small portion on the public order book (the exposed tip of the "iceberg"), with the remainder hidden inside the matching engine. This mechanism is intended to protect large traders from information leakage and predatory algorithmic trading. For market makers, hidden orders are an important tool for managing large inventories. When a market maker needs to close or build a large position, publicly displaying its full intent would trigger anticipatory reactions in the market and move the price in an unfavorable direction. By using hidden orders, a market maker can release liquidity gradually without setting off market panic.

Under stress scenarios, the existence of hidden liquidity makes measuring true liquidity more complex. On one hand, hidden orders may provide unexpected support at key price levels and slow the pace of a price decline; on the other, when market stress reaches an extreme level, this hidden support may also be rapidly withdrawn. On-chain automated market maker mechanisms have, to some extent, changed the paradigm of hidden liquidity. In a concentrated-liquidity model, although liquidity providers must publicly disclose the range over which their capital is distributed, they can achieve an effect similar to hidden orders by dynamically adjusting the range boundaries. This microstructural difference causes centralized and decentralized exchanges to exhibit different liquidity-decay features under stress testing.

Liquidity type	Visibility	Executability	Behavior in normal markets	Behavior in stressed markets
Apparent liquidity	Fully visible	Partly executable	Stable and ample; metrics reliable	Declines with a lag; false support present
Ghost liquidity	Fully visible	Very low	Constitutes the bulk of order-book depth	Withdrawn within milliseconds, producing a depth vacuum
Hidden liquidity	Invisible	Fully executable	A stable buffer of large orders	High cancellation uncertainty; fragile support

Table 20-1. The microstructural features and state evolution of the three liquidity dimensions (Data

source: compiled based on the research frameworks of Bessembinder et al. [23] and Degryse et al. [21])

Table 20-1 reveals a key asymmetry: in normal markets, ghost liquidity constitutes the bulk of order-book depth, whereas in stressed markets this portion of depth disappears within milliseconds. By contrast, hidden liquidity is invisible under normal conditions, and its behavior under stress is likewise hard to predict. This means that any risk assessment based on the visible depth of the order book faces a dual bias of both overestimation (the false support of ghost liquidity) and underestimation (the potential buffer of hidden liquidity), and in extreme conditions the risk of overestimation far outweighs that of underestimation.

20.5.3 The synchronized retreat of market makers

The disappearance of liquidity in a crisis is not a gradual process of decay but manifests as an instantaneous collective evaporation. The root of this phenomenon lies in the high correlation among different market makers' decisions. The empirical study by Anand and Venkataraman (2016) points out that when market conditions turn unfavorable, market makers tend to shrink their supply of liquidity at the same time, and this behavior produces strong covariation in liquidity both cross-sectionally and over time [6].

The synchronized retreat of market makers can be explained along two dimensions: inventory cost and the coordination game. First, all market makers face similar external environmental constraints. When volatility surges, a liquidation cascade begins, or the funding rate becomes extreme, market makers' inventory-risk costs and adverse-selection costs simultaneously cross their internally set risk thresholds. Because different market makers' risk-management models and input variables (such as price volatility and order-flow imbalance) are highly homogeneous, their algorithms tend to output similar retreat instructions when confronting the same market shock.

Figure 20-14 simulates the evolution of the participation levels of multiple independent market makers during a typical liquidity-shock event.

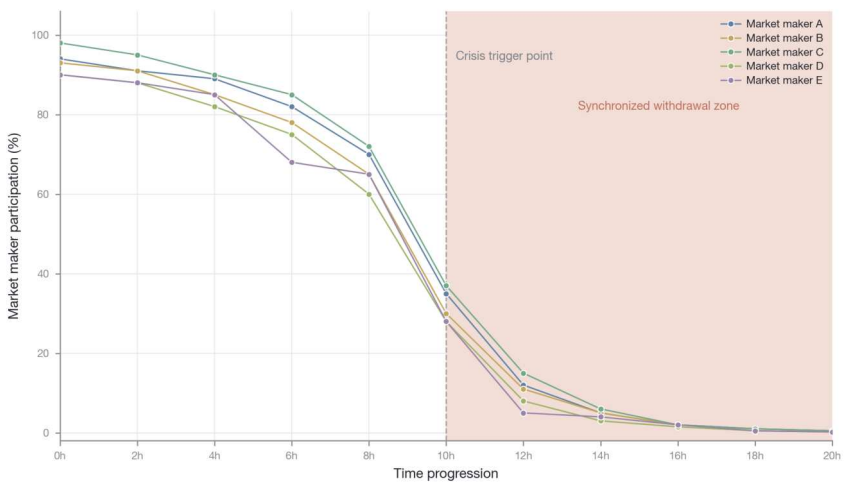


Figure 20-14. The synchronized retreat of market makers' liquidity supply under a stress event (Data source: constructed based on the covariation model of Anand and Venkataraman [6])

Second, there is also an implicit strategic complementarity among market makers, whose game-theoretic roots are the coordination game and informational externality analyzed in Section 20.2.3: when a market maker expects its peers to retreat, whoever stays will absorb all of the toxic order flow alone, so each maker strives to retreat first, and the system thereby transitions instantly from the good equilibrium of "everyone provides liquidity" to the bad equilibrium of "everyone retreats." The game-theoretic derivation is not repeated here.

20.5.4 The failure of liquidity metrics

The split between apparent and true liquidity ultimately causes traditional liquidity-assessment frameworks to fail under stress testing. Conventional liquidity metrics—such as the bid-ask spread, the cumulative order volume within a specified depth, or a price-impact coefficient based on historical trade data—are all built on the assumption that the supply of liquidity is continuous and stable. As discussed above, however, liquidity in perpetual futures markets is highly endogenous, and there is a structural negative correlation between its quantity supplied and market demand. This subsection focuses on a mechanistic diagnosis of this metric failure, leaving the concrete design of quantitative metrics to Section 20.8.

Figure 20-15 shows how this endogeneity causes the stable linear relationship between apparent depth and actual executable depth to become highly nonlinear under stress, with executable depth exhibiting severe marginal diminution, thereby causing risk models based on apparent depth to fail. This relationship can be tested by fitting the slope on paired "required depth–executable depth" samples for stress periods and normal periods; if the slopes in the two regimes do not differ significantly, the marginal-diminution claim does not hold.

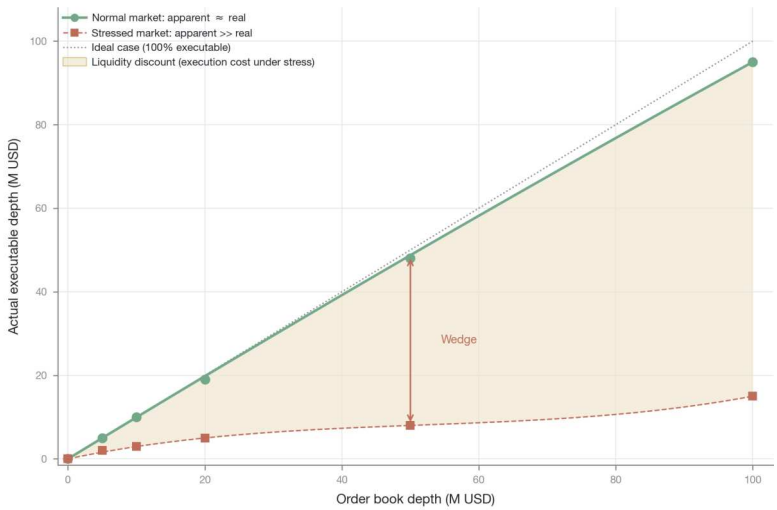


Figure 20-15. The nonlinear divergence between order-book depth and actual executable depth (Data source: derived in combination with Berkowitz's [24] liquidity-risk model; [24] is a working paper and has not been peer-reviewed)

To address this metric failure, risk-management theory has introduced the concept of

liquidity-adjusted value at risk. Bangia, Diebold, Schuermann, and Stroughair (1999) [25] first proposed a methodology for incorporating liquidity risk into the value-at-risk framework, capturing the amplifying effect of liquidity on portfolio losses by overlaying a distribution of execution costs on top of market risk. In the liquidation mechanism of perpetual futures, the underestimation of liquidity risk is especially pronounced. When the system's liquidation engine relies on apparent liquidity to assess an account's safety margin, it may mistakenly conclude that the existing margin is sufficient to cover the cost of closing. But when real liquidation orders pour into the market and trigger the synchronized retreat of market makers, the actual slippage cost far exceeds expectations, exceeds the margin coverage, and sets off chain liquidations.

The traditional liquidity-adjusted value-at-risk framework, however, still assumes that liquidity cost is exogenous (estimating execution slippage under a stress scenario at a fixed discount rate) and cannot capture the nonlinear collapse of endogenous perpetual-futures liquidity under market-maker retreat, so in principle it does not apply to the scenario described in this chapter. Measuring true liquidity requires endogenizing market makers' dynamic response function, simulating the nonlinear rise in inventory costs under a given volatility shock and predicting the critical point of algorithmic order cancellation; future methodological innovation may need to turn toward agent-based modeling or a reflexivity value-at-risk framework [24].

20.6 A five-stage framework of liquidity evaporation

The preceding sections revealed the theoretical foundations of liquidity endogeneity and the three positive-feedback loops that drive its self-dissolution. These mechanisms do not operate in isolation; they interweave along a specific temporal sequence and logical chain, constituting a complete transmission path for a liquidity collapse. The disappearance of liquidity is rarely instantaneous; it instead follows an identifiable five-stage evolutionary model. Understanding this transmission path is decisive for identifying market fragility, designing early-warning indicators, and formulating intervention strategies.

A word on positioning: this section links the preceding mechanisms from an integrative temporal perspective, and the five stages are identifiable phases, on the collapse timeline, of the three loops in Sections 20.2 to 20.4 and the contagion in Section 20.7, rather than a new mechanism independent of the three loops. Adjacent stages can overlap, and not every shock passes through all five stages in sequence (the counterexamples at the end of Section 20.6 and in Section 20.6.4 bear this out). The incremental value of the five-stage framework lies in characterizing the prefatory stage of the "microstructural trigger," as well as the leading indicators between stages (such as the time lag by which spreads and toxicity deteriorate ahead of depth, described in Section 20.6.2), while the internal mechanisms of each stage refer back to Section 20.3 (the liquidation cascade) and Section 20.7 (contagion). Viewed this way, the sudden drying-up of liquidity through synchronized market-maker retreat under stress is consistent with the empirical characterization by Kirilenko, Kyle, Samadi, and Tuzun (2017) [26] of the "Flash Crash" of May 2010: the concentrated exit of high-frequency market makers once their inventory limits were exceeded was the key link by which a single large sell

order evolved into a market-wide evaporation of liquidity. This empirical work provides external corroboration, from outside crypto markets, for this section's microstructural-trigger and synchronized-retreat stages.

Figure 20-16 shows the core transmission path and feedback mechanisms of the five-stage model of liquidity evaporation.

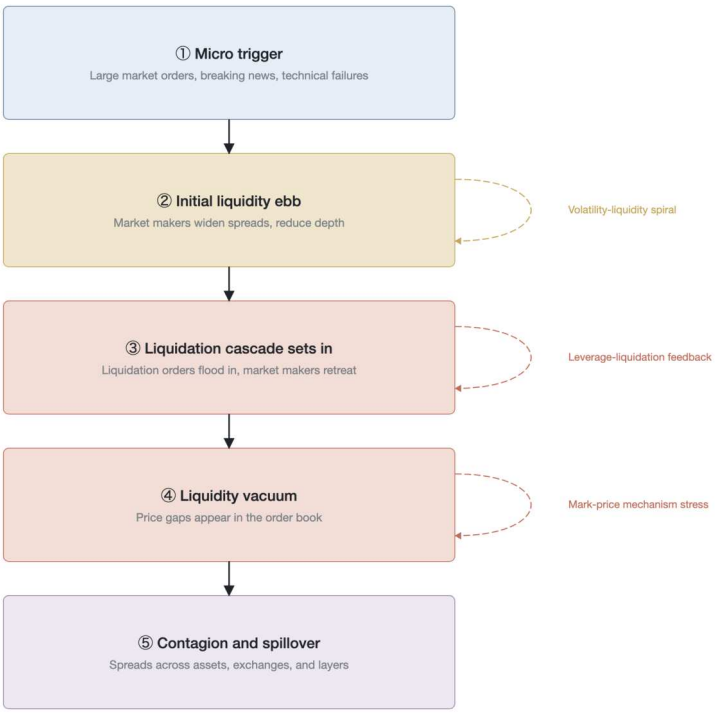


Figure 20-16. A schematic of the five-stage model of liquidity evaporation and its positive-feedback loops

As shown in Figure 20-16, the five stages are not a simple linear progression; a positive-feedback loop intervenes at each transition node: the microstructural trigger to the initial ebb is driven by the volatility-liquidity spiral, the initial ebb to the liquidation cascade is amplified by leverage endogeneity, and the liquidity vacuum to contagion spillover relies on market makers' cross-asset risk-control linkages.

Not every liquidity shock evolves to the fifth stage. During the "Black Thursday" event of March 12, 2020, BitMEX was forced to suspend trading for about 25 minutes because of system overload at the extreme moment when the BTC price plunged to about $3,800. This involuntary trading halt and the subsequent slowing of the liquidation cascade occurred in temporal succession, and can serve as an illustrative example of "an external intervention interrupting a loop at a particular stage." A caveat: in a single event it is hard to separate whether the halt itself or the market's natural bottoming was the principal cause, so this chapter makes no strong causal claim that the halt blocked the cascade. When BitMEX resumed trading, the most violent panic had already

passed, and the market, near stage three, did not continue to evolve toward stages four and five. This case also illustrates that the five-stage framework describes the complete transmission path in the absence of intervention, and that a timely external intervention—whether an intentional circuit breaker or an incidental system outage—can interrupt the positive-feedback loop at a particular stage.

20.6.1 Stage one: the microstructural trigger

The starting point of liquidity evaporation is typically a trigger event at the microstructural level. The event itself need not be systemically destructive, but it is enough to break the prevailing liquidity equilibrium. Such trigger events fall into three categories: the shock of a single very large market order, the injection of sudden negative news, and a technical failure of an exchange's infrastructure. In normal markets, these shocks are quickly absorbed by ample apparent liquidity. When the market is already in a fragile state, however—when the system's effective leverage is elevated, market makers' inventory is severely skewed, or the market is in a naturally thin liquidity trough—a small trigger event can set off a catastrophic chain reaction.

The core feature of a trigger event lies in its instantaneous perturbation of the initial price and the bid-ask spread. For example, a market sell order that exceeds the top-layer depth of the order book directly penetrates several price levels, causing the last traded price to jump anomalously. This price jump not only reflects an instantaneous imbalance of supply and demand but, more important, sends a strong signal of rising volatility to all market participants. Market makers' algorithmic models immediately capture this signal and reassess the current adverse-selection risk and inventory-carrying cost. At this stage, the observable microstructural indicators show a widening of the initial price-jump magnitude and an initial widening of the bid-ask spread, even though overall order-book depth has not yet contracted catastrophically.

20.6.2 Stage two: the initial ebb of liquidity

As the microstructural trigger event pushes realized volatility higher, market makers begin to make a rational first-round response, and the market enters the stage of the initial ebb of liquidity. Based on the inventory-cost convexity analyzed in Chapter 19, rising volatility causes market makers' expected inventory costs to grow nonlinearly. To control their risk exposure, market makers' first defensive measure is to widen the bid-ask spread, thereby increasing compensation for potential adverse selection. If the volatility indicators continue to deteriorate, market makers further reduce the size of the limit orders they post at each price level of the order book—that is, they shrink their market-making depth.

This stage marks the formal onset of the positive-feedback spiral between volatility and liquidity. Market makers' widening of spreads and reduction of depth directly cause subsequent orders of the same size to produce larger price impact. This liquidity-decay-induced price impact further raises the market's effective volatility, forcing more market makers to tighten their quotes [27]. Market makers have not yet withdrawn from the market entirely at this point; although apparent liquidity is impaired, it still maintains basic trading functionality. Observable indicators show the bid-ask spread possibly

widening to three to ten times its normal level, order-book depth falling by more than half, and indicators reflecting order-flow toxicity beginning to climb sharply.

Figure 20-17 shows the evolution of key market-microstructure indicators across the stages of liquidity evaporation.

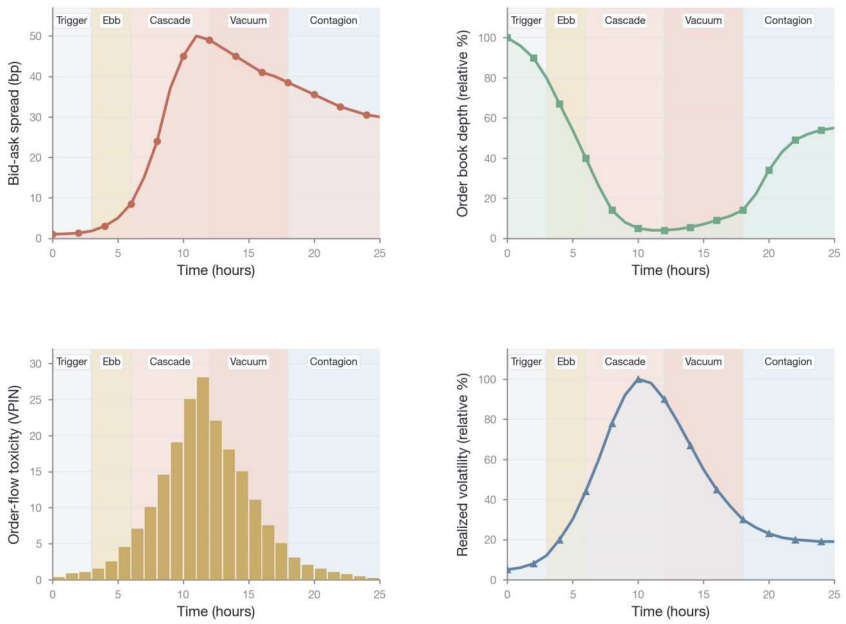


Figure 20-17. The evolution of key indicators in the five-stage model of liquidity evaporation (an illustrative fit by the author based on the FTX collapse case; not data from [15])

In the sense of a model illustration, Figure 20-17 reveals a key early-warning logic: during the transition from stage one to stage two, the bid-ask spread and order-flow toxicity indicators have already begun to deteriorate markedly, but the apparent depth of the order book has not yet fallen abruptly. This time lag among the indicators means that the anomalous widening of spreads and the sharp climb in toxicity can serve as leading signals of a liquidity collapse. Once these signals appear, the transition from stage two to stage three may be completed in a very short time, leaving market participants a very limited window to react.

20.6.3 Stage three: the intervention of the liquidation cascade

When the magnitude of the initial price decline reaches the maintenance-margin requirements of large numbers of high-leverage positions, the market enters the most destructive third stage: the intervention of the liquidation cascade. Its internal mechanisms were detailed in Section 20.3; here we only characterize the identifiable features of this stage. As noted earlier, liquidation orders are a perfectly inelastic, mandatory demand for liquidity. These liquidation instructions, pouring in as market orders, directly devour the order-book depth that has already become thin during the initial-ebb stage. The price is pushed down without resistance to the next liquidation trigger line, setting off a new round of forced closures, and the positive-feedback loop between

leverage and liquidation (Sections 20.3.1–20.3.3) is fully activated.

At this stage, market makers face a core discrimination problem. Confronting the surging one-directional order flow, a market maker must judge whether it is a temporary liquidity shock or the harbinger of a systemic collapse. Because taking on the wrong orders could push their own positions to the brink of liquidation, whereas the cost of retreating in error is merely the loss of some fee income, the vast majority of market makers choose the latter. This collective, rational retreat turns former liquidity providers into bystanders; in extreme cases, market makers may even become demanders of liquidity themselves in order to hedge the one-directional inventory they have accumulated. The observable features of this stage are: liquidation volume climbs sharply, open interest falls sharply, and the exchange's insurance fund begins to be heavily consumed.

20.6.4 Stage four: the liquidity vacuum

The compounded effect of the liquidation cascade and the synchronized retreat of market makers ultimately pushes the market into the liquidity-vacuum stage. At this stage, the continuity of the order book is completely destroyed, and a visible "price gap" appears—an extremely wide price range with no limit-order quotes whatsoever. Any market order attempting to cross this vacuum zone causes an extreme jump-like gap in the price. This price discontinuity not only destroys the market's price-discovery function but also poses a serious challenge to the underlying infrastructure of perpetual futures.

The liquidity vacuum places enormous pressure on the mark-price mechanism. A mark price based on an exponential moving average lags significantly when confronting a price gap, which can lead to misjudgment of a position's health and thereby trigger unjustified liquidations or impede necessary risk control. If the bankruptcy losses caused by liquidation exceed the absorptive capacity of the insurance fund, the exchange is forced to trigger the auto-deleveraging mechanism. The forced closure of positions on the profitable side means that even a trader whose directional judgment was correct is forced to bear the cost of the systemic depletion of liquidity. During this period, the bid-ask spread may widen to hundreds of basis points, and conventional liquidity metrics fail completely.

The FTX data come with one important qualification: some of the liquidity on the FTX platform was built on the basis of an affiliated institution misappropriating client funds for market making, and its collapse was to a large extent a fraud-exposure event (Conlon, Corbet, and Hu, 2026 [28]). The fraud factor, however, does not negate the analytical value of the endogenous liquidity spiral, because similar order-book depth evaporation and spread surges also occurred in the May 2021 crash, which involved no fraud. The following data are used to illustrate the speed and magnitude of liquidity deterioration, not to generalize the special circumstances of FTX.

Market state	Stage	Bid-ask spread	Order-book depth (bid/ask)	Price-impact coefficient
Pre-crisis normal	November 7	0.5 bp	\$50M / \$50M	Very low
Initial ebb	Morning of November 8	2.0 bp	\$45M / \$42M	Rising slightly
Liquidation cascade	Evening of November 8	10.0 bp	\$25M / \$20M	Rising markedly
Liquidity vacuum	Midday of November 9	45.0 bp	\$3M / \$2M	Extremely elevated

Table 20-2. An illustrative evolution of BTC perpetual-futures liquidity metrics during the FTX collapse (Data source: an illustrative reconstruction by the author based on public market data; not data from [15])

The data in Table 20-2 are an illustrative time series meant to reveal the astonishing speed of liquidity deterioration. From the normal state of November 7 to the liquidity vacuum of November 9, the bid-ask spread widened 90-fold in under 48 hours (from 0.5 bp to 45 bp), while order-book depth shrank by about 94%–96% (from \$50M to \$2–3M). The depth decay on the two sides was not symmetric: in the liquidity-vacuum stage, bid-side depth (\$3M) was slightly higher than ask-side depth (\$2M), reflecting the structural feature that sell-side liquidity is consumed first in a plunging-price environment. The depth data in Table 20-2 are an illustrative characterization of the order-book condition of a single BTC perpetual-futures trading pair on the platform directly hit by the FTX crisis, whereas the \$580 million depth cited at the opening of this chapter is Kaiko's aggregate 1%-band depth across the whole market, multiple exchanges, and multiple currencies. The two datasets have different statistical definitions, but both point to the same conclusion: whether at the single-platform or market-wide level, the speed and magnitude of liquidity deterioration far exceed what market participants would expect based on normal experience.

20.6.5 Stage five: contagion spillover

A liquidity crisis is rarely confined to a single asset or a single trading platform; its final stage manifests as cross-dimensional contagion and spillover. Market makers' global risk-control directives spread the liquidity stress of core assets to peripheral assets; the rupture of cross-exchange arbitrage paths causes price-spread anomalies to persist and transmits pressure in reverse; and the price feeds of oracles transmit the extreme volatility of centralized markets to decentralized protocols, triggering on-chain liquidations. The full mechanism of this contagion spillover is developed systematically in Section 20.7.

20.7 The cross-dimensional contagion of liquidity

Once a liquidity crisis erupts, it rarely stays within the single asset or single market in which it originated. The liquidity network of perpetual futures markets forms a highly interconnected complex system through market makers' cross-asset capital allocation, arbitrageurs' correlated strategies, and oracles' price-transmission mechanisms. When

a vacuum in liquidity appears at one node, this stress spreads outward through specific transmission chains, amplifying a local microstructural shock into a systemic liquidity crisis. This contagion process does not occur at random; it follows structural paths from core to periphery, from spot to derivatives, and from centralized to decentralized markets.

Figure 20-18 shows the four main dimensions along which a liquidity crisis spreads outward and the mechanisms by which they interact.

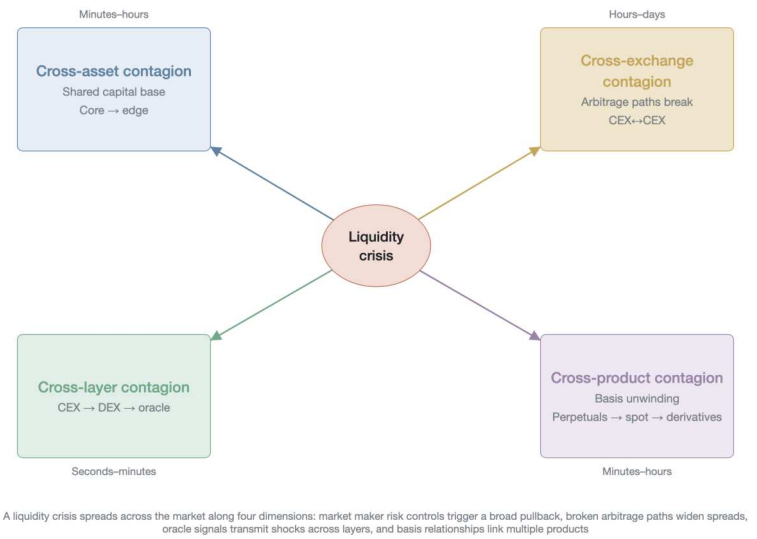


Figure 20-18. A four-dimensional model of liquidity contagion (mechanism reference: Akhtaruzzaman et al. [29]; the four-dimensional framework is the author's extension of their systemic-risk concept, not the original classification in [29])

As shown in Figure 20-18, the four contagion dimensions are not mutually independent but interact closely: cross-asset contagion operates through market makers' unified capital base, cross-exchange contagion through the rupture of arbitrage paths, cross-layer contagion through the oracle system, and cross-product contagion through the concentrated closing of basis trades, each spreading local liquidity stress outward. The following subsections analyze each in turn.

20.7.1 Cross-asset contagion

In crypto-asset markets, the cross-asset contagion of liquidity exhibits a marked asymmetry, typically taking the form of a strong one-directional transmission from core assets to peripheral assets. The root of this phenomenon lies in market makers' capital-management structure and risk-control mechanisms. A market maker typically provides liquidity across multiple assets simultaneously, but its capital base is unified. When a core asset experiences violent price swings and triggers a liquidation cascade, the market maker's market-making positions in that asset face extreme adverse-selection risk and inventory pressure.

To protect its overall capital, a market maker's risk-control system triggers a global directive to shrink. This risk-control mechanism does not distinguish among asset classes but instead requires widening bid-ask spreads and reducing quoted depth across all trading pairs. As a result, even peripheral assets that have themselves experienced no fundamental change or direct selling pressure see their liquidity dry up instantly because of the market maker's defensive retreat.

The speed and scope of cross-asset contagion depend to a large extent on the topology of the market-maker network. Market depth is highly concentrated among a small number of top market makers: according to industry reports from institutions such as Kaiko and Wintermute, the top five market makers provide a substantial share (on the order of 50% to 70%) of order-book depth at major exchanges. These figures come from the proprietary definitions of data vendors and market-making firms, whose methodologies are not public, and some sources (such as the top market makers themselves) have an inherent conflict of interest; this chapter treats them only as an order-of-magnitude reference rather than a precise measurement, cross-referencing them with the "top 3–5 firms reaching depth on the order of 70%" definition in Section 19.5.1 of Chapter 19. This highly concentrated core-periphery structure has an efficiency advantage in normal times: the depth coverage of core market makers reduces the market's overall transaction costs. Research by Acemoglu, Ozdaglar, and Tahbaz-Salehi (2015) [30], however, shows that such a network structure has absorptive capacity under small shocks but experiences cascading, full-scale contagion under large shocks that exceed a critical threshold. When one core market maker fully retreats because of a risk-control directive, its exit simultaneously affects dozens of trading pairs and multiple exchanges, instantly creating a liquidity vacuum across the entire network.

The cross-margin regime further reinforces this cross-asset contagion effect. Under a unified margin account, if the long positions in a core asset held by a trader incur unrealized losses because of a sharp price decline, those losses directly erode the account's total margin balance. When the total margin level falls below the maintenance requirement, the system forcibly closes the account's other asset positions, including peripheral assets whose prices were originally stable. This mechanism converts the price shock of a core asset into forced selling of peripheral assets, and against the backdrop of peripheral-asset liquidity already drained by market makers, this forced selling triggers an even more violent price collapse in the peripheral assets.

Figure 20-19, based on market data from the November 2022 FTX collapse, records the difference in the liquidity-evaporation timeline between core and peripheral assets during this transmission process.

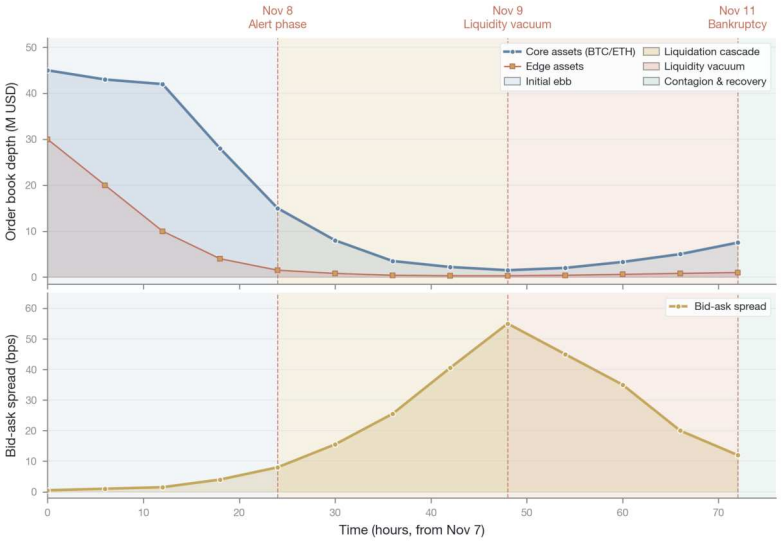


Figure 20-19. The timeline of liquidity evaporation under extreme market conditions (mechanism reference: Galati et al. [31])

In this transmission process, the degree of liquidity deterioration in peripheral assets often far exceeds that in core assets. Because of their large trading volume and diversified participant structure, core assets can still attract some bargain-hunting liquidity providers after an initial liquidity shock. Peripheral assets, however, ordinarily rely on just a few market makers to maintain liquidity, so once those market makers retreat, the market falls into a complete liquidity vacuum, and the bid-ask spread may widen dozens or even hundreds of times, producing an extreme liquidity premium.

An apparent tension arises here: this subsection emphasizes that core assets "can still attract some bargain-hunting liquidity" (a self-repair tendency), whereas Section 20.5.3 emphasizes that market makers' "synchronized retreat causes collapse" (a fragility tendency). The two are not contradictory but correspond to different parameter regimes: when the market is sufficiently deep, participants are sufficiently heterogeneous, and there is still a capital buffer (typically a mild pullback in a core asset), replenishment dominates and the market exhibits self-repair; when the market is thin, participants are highly homogeneous, and the capital buffer is exhausted (typically a peripheral asset or the extreme tail of a core asset), retreat dominates and the market exhibits an abrupt collapse. The dividing variables are mainly the number of heterogeneous participants and the size of the deployable capital buffer. This "retreat-dominated versus replenishment-dominated" demarcation is currently a theoretical inference, and empirical calibration of its threshold is left to future research.

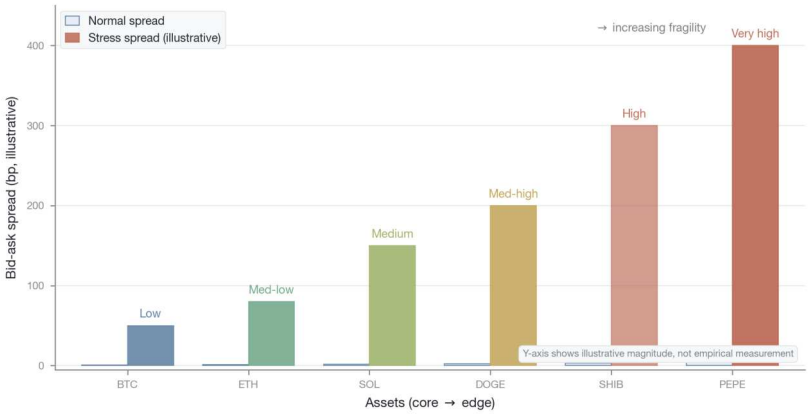


Figure 20-20. A comparison of the features of cross-asset liquidity-transmission chains (mechanism reference: Hautsch et al. [32])

Figure 20-20, taking the liquidity-transmission-chain features of multiple asset classes as its objects of comparison, shows the differentiated impact of cross-asset contagion across core assets, mid-tier assets, and peripheral assets, corroborating this subsection's argument that core assets are relatively resilient while peripheral assets are extremely fragile.

20.7.2 Cross-exchange contagion

Liquidity contagion among centralized exchanges is realized mainly through the failure of the cross-exchange arbitrage mechanism. In normal markets, arbitrageurs play the role of a bridge connecting the liquidity pools of different exchanges. When a particular exchange experiences a brief liquidity imbalance and price deviation in a specific asset, arbitrageurs trade counter-directionally at that exchange while hedging at other exchanges with ample liquidity, thereby dispersing the local liquidity stress across the whole market.

When an extreme liquidity crisis erupts, however, this dispersion mechanism suddenly ruptures. If a major exchange suffers a liquidity run that severely distorts its internal prices, cross-exchange arbitrageurs will attempt to capture this enormous price-spread profit. But the peculiarity of a crisis is that the collapsing exchange is often accompanied by withdrawal restrictions, overloaded trading interfaces, or extreme slippage costs. These frictions prevent arbitrageurs from smoothly completing fund transfers or position hedges, causing the arbitrage path to rupture substantively.

Once the arbitrage path ruptures, arbitrage positions that were originally meant to dampen price spreads turn into one-sided exposures. To control risk, arbitrageurs are forced to urgently close their hedge positions at the other exchanges that are still functioning normally. This large-scale, unexpected closing behavior directly transfers the liquidity stress of the collapsing exchange onto the order books of the other exchanges. At the same time, having observed the collapse of one exchange, the market makers at the other exchanges anticipate the incoming flow of closing orders and the potential systemic risk, and proactively withdraw their quotes. This expectation-based

synchronized retreat causes the liquidity crisis to achieve instantaneous cross-exchange contagion among physically isolated exchanges.

20.7.3 Cross-layer contagion

Liquidity contagion between centralized and decentralized markets characterizes the third dimension of systemic risk. The core hub of this cross-layer contagion is the oracle system. The operation of decentralized finance protocols depends heavily on the external price data that oracles provide, and the vast majority of that data originates from the spot or derivatives markets of centralized exchanges.

When a centralized exchange suffers a liquidity crisis that causes its price to plunge in an instant, this extreme and possibly distorted price is faithfully captured by the oracle and broadcast to the various decentralized lending protocols and synthetic-asset platforms. The smart contracts of decentralized protocols lack the ability to discern anomalous price movements and mechanically trigger on-chain liquidation procedures according to the oracle's price feed. This on-chain liquidation triggered by an external liquidity shock forces liquidators to sell the collateral assets into the liquidity pools of decentralized exchanges, thereby causing on-chain liquidity depletion and price collapse.

As the hub of cross-layer contagion, the oracle has failure modes spanning at least three distinguishable dimensions. The first is latency contagion: the price updates of mainstream oracles are constrained by heartbeat intervals and deviation thresholds, and in extreme volatility the price feed may lag significantly behind the real-time price of centralized exchanges, causing the timing of on-chain liquidation triggers to deviate from the true market state and producing over-liquidation or under-liquidation. The second is manipulation contagion: an attacker can influence an oracle's price feed by manipulating the price of the underlying asset on a centralized exchange, thereby triggering anomalous liquidations in a decentralized protocol. The Mango Markets incident of October 2022 is a representative case of this contagion channel, but its specific mechanism must be characterized accurately: the attacker pumped the price of the MNGO perpetual futures across multiple exchanges, which raised the valuation of their own collateral within the Mango protocol (a long MNGO position), then over-borrowed against it and drained the protocol's funds, involving more than $100 million. The essence of this case is "manipulating the valuation of one's own collateral and then over-withdrawing," not "manipulating a third party's spot price to trigger cascading liquidations of others' positions"; but it likewise reveals the cross-layer fragility whereby "the oracle faithfully transmits a manipulated price." The third is congestion contagion: during extreme conditions, a surge in the underlying blockchain's transaction fees may prevent oracle nodes from updating their price feeds in time, causing on-chain protocols to make liquidation decisions using stale price data.

Reverse contagion also exists, and although it is less frequent, its destructive power should not be dismissed. When a decentralized market suffers a flash-loan attack, oracle manipulation, or a vulnerability in the underlying protocol, on-chain asset prices become extremely distorted. High-frequency arbitrageurs quickly capture this enormous on-chain/off-chain price spread and convert this distorted price signal into real trading orders on centralized exchanges. If the market makers of the centralized mar-

ket fail to identify this false-information shock in time, they may suffer severe losses after taking on the orders, which in turn triggers their risk-control mechanisms and leads to a full retreat. This two-way contagion mechanism shows that centralized and decentralized markets are already deeply bound at the liquidity level, and that fragility at either end becomes a hidden danger for the entire system.

20.7.4 Cross-product contagion

The liquidity crisis of perpetual futures markets ultimately spreads, through inter-product arbitrage relationships, to the spot market and other traditional financial derivatives markets. The most direct link between the perpetual futures and spot markets is the funding-rate arbitrage mechanism. Under normal conditions, large amounts of arbitrage capital earn a steady funding-rate return by holding long spot and short perpetual positions.

When the perpetual futures market falls violently and triggers a long-liquidation cascade, the funding rate rapidly turns from positive to negative, and the negative value may reach an extreme level. At this point, the arbitrage positions that were originally earning a return turn into persistently loss-making positions. To stop losses, arbitrageurs must close their combined positions—buying to close their shorts in the perpetual futures market while selling to close their longs in the spot market. This large-scale spot selling directly transmits the liquidity stress of the perpetual futures market to the spot order book, causing the spot price to plunge in tandem.

This inter-product contagion is not confined to crypto-native markets; it also extends outward along the complexity chain of financial instruments. Violent swings in the spot price are quickly reflected in the regulated futures contracts listed on traditional exchanges, forcing the market makers of traditional financial institutions to adjust their hedge positions. Further, within traditional trading hours, this volatility is transmitted to the related exchange-traded fund (ETF) markets, where, through the lag effect of the creation and redemption mechanism, it continues to affect market liquidity the next day or over a longer period. This tiered transmission from crypto-native derivatives to regulated futures and ETFs is a mechanistic inference; its specific strength and time lag await empirical testing. This cross-product contagion, unfolding in tiers along a temporal sequence, demonstrates the systemic influence of perpetual futures as a source of liquidity risk.

20.8 Institutional roots and quantification

The foregoing mechanistic analysis has revealed the complete transmission chain by which liquidity evaporates under stress. This section integrates these mechanisms into an institutional-level summary, proposes quantitative methods for measuring liquidity risk, and examines the game-theoretic dilemma in market makers' collective behavior, thereby providing a theoretical basis for the design of governance frameworks in subsequent chapters.

The liquidity fragility of perpetual futures markets is not a transitional phenomenon of the market's early development but a structural feature arising from its core in-

stitutional design. Compared with traditional equity or futures markets, perpetual futures amplify the intensity and transmission speed of liquidity shocks along multiple dimensions (Table 20-3).

Dimension	Traditional equity and futures markets	Perpetual futures markets
Leverage multiple	Constrained by regulation (equity margin about 2–3x; index/commodity futures about 10–20x by contract)	Commonly offer high leverage of 50x to 125x
Liquidation mechanism	Margin-call mechanism with a buffer period	Forced market closure executed the moment it is triggered
Trading hours	A fixed closing period exists as a natural shock absorber	Round-the-clock, non-stop trading
Circuit breakers	Price limits or trading halts are common	The vast majority of platforms lack systemic circuit breakers
Market-maker obligations	Some markets have quoting obligations for designated market makers	Purely voluntary participation, with no mandatory quoting requirement
Central clearing	A central counterparty absorbs default risk	No unified clearing; each exchange's risk is independent
Information environment	Standardized disclosure under strict regulation	No unified and transparent disclosure standard

Table 20-3. A comparison of the institutional features of traditional markets and perpetual futures markets (Data source: Joshi [33]; that paper is a working paper / job-market paper, not peer-reviewed, and the relevant institutional comparisons should be cross-checked against public trading rules)

The combination of high leverage and an immediate liquidation mechanism is the core driver of liquidity fragility. As listed in Section 20.1.3 and Table 20-3, perpetual liquidation executes the moment it is triggered, with no margin-call buffer; while this protects the exchange from bankruptcy losses, it shifts enormous liquidity stress onto the market. When the market is in a one-sided sharp decline, the chain liquidation of high-leverage positions produces perfectly inelastic demand for liquidity that rapidly devours order-book depth [34].

In addition, as described in the fourth amplifier of Section 20.1.3, round-the-clock trading and the absence of circuit breakers deprive the market of any respite during extreme volatility. A liquidity shock may erupt during the thinnest-depth periods, such as weekends or the pre-dawn hours, when market makers' response capacity is limited and an initial shock more readily evolves into a systemic liquidity vacuum. This difference in institutional design shows that, without changing the underlying architecture, the fragility of perpetual futures markets cannot be eliminated by trading-volume growth alone.

Moreover, the liquidity fragility of perpetual futures markets is not determined solely by endogenous mechanisms; the macro liquidity environment constitutes an important precondition for the accumulation of fragility. The liquidity deterioration of 2022 did not begin with the FTX event in November but had been accumulating throughout the Federal Reserve's aggressive rate-hiking cycle. Market-wide order-book depth had already declined markedly from early 2022 to just before the FTX event, and the FTX

event merely delivered a final blow to an already fragile base. The dollar-tightening cycle systematically compressed the liquidity base of perpetual futures markets through multiple channels: the contraction of institutional capital's risk appetite caused market-making capital to withdraw, the rising yield on stablecoin reserves raised the opportunity cost of market making, and the failures of crypto-friendly banks directly severed the funding channels between fiat and crypto markets. Ignoring this macro context would overestimate the independent explanatory power of the endogenous mechanisms described in this chapter.

The difference in institutional features is further reflected at the level of measuring liquidity risk. Traditional liquidity metrics can effectively reflect market depth and transaction costs under normal conditions, but they often fail under extreme market conditions: as analyzed in Section 20.5, apparent liquidity systematically overestimates true liquidity under stress, causing severe bias in risk measurement.

To capture the risk of a liquidity collapse more accurately, this section proposes four liquidity-risk quantification metrics oriented toward perpetual futures markets. This section is a design proposal for quantification metrics, intended to indicate operable directions for measurement; their full parameterization and historical backtesting are left to subsequent empirical research, and except for the "market-maker concentration index," for which a minimal formalization is given below, the other metrics remain at the level of conceptual definition.

First, conditional depth measures the expected available liquidity at a given level of volatility or order toxicity. Unlike a static order-book snapshot, conditional depth incorporates market makers' cancellation probability and reflects the dynamic response of liquidity supply to market state variables. When volatility breaches a certain threshold, conditional depth exhibits a nonlinear, abrupt decline, and this metric can provide an early warning of the fragility behind apparent liquidity (see Cont et al. [35] on the framework for liquidity stress testing).

Second, liquidity value at risk borrows a concept from traditional financial risk management to quantify liquidity losses under extreme scenarios. It measures, at a given confidence level, the maximum proportion of liquidity that the market may evaporate within a specified time window [35]. For example, liquidity value at risk at a 95% confidence level indicates that in the worst 5% of market scenarios, order-book depth will shrink to an extremely low proportion of its normal level. This metric transforms liquidity risk from a single price-impact measure into a system-level probability of supply depletion.

Third and fourth, the liquidation-trigger distance and the market-maker concentration index provide risk early warning at the microstructural level. The liquidation-trigger distance measures the gap between the current market price and the next dense liquidation band. The nearer the distance, the higher the probability that microscopic price movements trigger a cascade liquidation, and the greater the risk that the market enters the positive-feedback loop of liquidity consumption. The market-maker concentration index assesses the degree to which market depth depends on a small number of core liquidity providers. As a minimal formal proposal, the market-maker concentration index can be defined as:

$$\mathrm{MMCI}_N = \sum_{i=1}^{N} s_i$$

where s_i is the proportion of total market-making depth provided by the i-th market maker on the target trading pair (ranked in descending order), and N is the number of top market makers included in the statistic. Taking $N = 5$, and combining this with the illustrative reference of "the top five firms accounting for depth on the order of 50%–70%" given in Section 20.7.1, yields the illustrative value $\mathrm{MMCI}_5 \approx 0.5$–$0.7$ (this value inherits the gray-literature nature of the source in Section 20.7.1 and is illustrative only). The closer MMCI_N is to 1, the more market depth depends on a small number of top market makers, and the higher the risk that the market falls into a liquidity vacuum once these market makers retreat in concert because of their risk-control strategies.

The quantitative methods reveal the structural features of liquidity risk, while market makers' collective behavior gives rise to a deeper institutional dilemma. Liquidity in perpetual futures markets exhibits the features of a quasi-public good. All market participants—whether directional traders or arbitrageurs—benefit from tight bid-ask spreads and ample order depth. Yet this resource, with its broad externalities, is supplied entirely and voluntarily by profit-maximizing private market makers. Under systemic stress, this supply mechanism inevitably falls into the classic tragedy of the commons.

From the standpoint of individual rationality, each market maker's decision is driven by its real-time profit equation. When market volatility surges and the toxicity brought by liquidation orders rises sharply, continuing to maintain quotes means bearing enormous inventory risk and adverse-selection cost. At this point, pulling orders and exiting the market is the optimal strategy for protecting one's own capital. When all market makers make the retreat decision based on the same rational logic, however, the result of collective irrationality follows: liquidity across the entire market collapses completely.

In a liquidity-vacuum environment, the price-discovery mechanism fails, and a small sell order can trigger a violent price gap. This systemic collapse harms not only ordinary traders but ultimately the market makers themselves. The inventory positions market makers already hold cannot be closed at a reasonable price in a market that lacks liquidity, causing them to face paper losses far exceeding expectations. This is a classic prisoner's dilemma: cooperation (jointly maintaining liquidity) can prevent a market collapse and is most beneficial for the collective; but for any single market maker, defection (retreating first) is always the dominant strategy. The final Nash equilibrium is necessarily that all market makers retreat in concert and the market falls into a suboptimal state of liquidity depletion [36].

Traditional financial markets mitigate this dilemma by introducing external mechanisms. For example, a central bank acts as lender of last resort in times of crisis, or an exchange mandates quoting obligations for licensed market makers under extreme conditions. In the crypto-asset derivatives market, which is highly decentralized and lacks unified regulation, there is neither a central institution capable of providing unlimited liquidity nor any mandatory market-making constraint. The competitive relationships and information asymmetry among market makers make spontaneous coordinated action impossible. This is the core challenge of liquidity governance. The

global regulatory environment is responding to the institutional deficiencies described above: as of 2025, the European Union's Markets in Crypto-Assets Regulation (MiCA), the Hong Kong Securities and Futures Commission's framework for virtual-asset trading platforms, and Japan's Financial Services Agency, among others, are respectively imposing obligations on licensed exchanges regarding market integrity, orderly trading, and even designated-market-maker-like liquidity guarantees. These emerging regulatory directions may gradually alter the "no-obligation" baseline state described in this chapter, and their specific impact is discussed further in the next chapter.

Integrating the analysis above, this chapter has systematically diagnosed the roots of liquidity fragility in perpetual futures markets. Liquidity is not an exogenously given static environment but an endogenous variable coupled deeply with price, volatility, and leverage. The positive-feedback loops formed by these three elements, under the effect of the institutional amplifiers unique to perpetual futures, can convert a small initial shock into a market-wide liquidity catastrophe. The vast wedge between apparent and true liquidity, together with the tragedy of the commons in market makers' collective action, jointly reveals the deep-seated defects of the existing liquidity-supply system.

The collective retreat of market makers in extreme conditions stems not from irrational behavior but from a structural tendency of the profit equation. Since the sum of individual rationality produces systemic fragility, the solution must go beyond strategy optimization at the microscopic level and turn toward institutional and incentive design at the macroscopic level. The next question is: how can market rules be restructured to change the return-risk asymmetry market makers face in a crisis? How can innovative governance frameworks be designed so that privately supplied liquidity possesses the resilience of public infrastructure? These explorations of liquidity governance and reconstruction are the central topic of the next chapter.

20.9 Chapter summary

This chapter has revealed a fundamental contradiction in the liquidity system of perpetual futures: liquidity disappears on a large scale at precisely the moment it is most needed. This countercyclicality does not stem from participants' irrational panic but is a structural consequence of market microstructure and the market maker's profit equation. By dissecting the mechanisms of liquidity endogeneity, this chapter has deconstructed the positive narrative of liquidity supply established in the previous chapter, showing how a seemingly stable supply system can fail systematically under extreme stress.

Liquidity in perpetual futures markets is highly endogenous. It is no longer the exogenous background condition assumed by traditional theory but a dynamic variable coupled deeply with price, volatility, and leverage. The reflexivity triangle formed by these three elements produces mutually reinforcing positive-feedback loops: a price decline triggers rising volatility, market makers' inventory costs grow convexly, forcing them to widen spreads or withdraw depth, and the retreat of liquidity in turn amplifies the price impact of subsequent orders and drives volatility higher still; at the same time, the price decline erodes long margin, effective leverage climbs passively, and

forced liquidation is triggered, with liquidation orders pouring in as perfectly inelastic demand for liquidity that both devours the remaining order-book depth and drives out the remaining market makers through high toxicity; and the extremization of the funding rate in a one-sided market distorts the long-short position structure, triggering the concentrated exit of arbitrage capital and adding another layer of pressure to the liquidity collapse.

The institutional features of perpetual futures amplify this endogenous fragility. Leverage as high as a hundredfold, market liquidation executed the moment it is triggered, the absence of the natural stabilizer of expiry convergence, and a round-the-clock trading environment lacking circuit breakers together constitute a highly sensitive risk-amplification system; these designs, which in normal times confer capital efficiency and trading convenience, turn into catalysts of liquidity evaporation in a crisis. Under their effect, a vast wedge opens between apparent and true liquidity: the ample depth displayed on the order book often contains large amounts of ghost liquidity that disappears in an instant under stress, causing conventional risk metrics to systematically underestimate true liquidity risk on the eve of a crisis.

The collapse of liquidity is not an instantaneous random event but follows a five-stage path from microstructural trigger to systemic contagion. An order that exceeds the market's absorptive capacity or an unexpected external shock acts as the initial trigger, causing market makers to tighten their quotes initially; the subsequent price decline activates the liquidation cascade, and forced sell orders resonate with market-maker retreat; once depth is exhausted, a price gap appears on the order book and the market falls into a liquidity vacuum; finally, the local depletion spreads rapidly, through market makers' cross-asset risk control, the cross-margin mechanism, and arbitrageurs' cross-market positions, to other assets, other exchanges, and even the spot and traditional derivatives markets.

The essence of this catastrophe is a classic tragedy of the commons. Liquidity, as a quasi-public good that benefits the whole market, depends entirely on the voluntary supply of private market makers. In extreme conditions, each market maker's retreat decision based on the profit equation is an individually rational choice, but the synchronized retreat of all market makers produces the result of collective irrationality—a full collapse of market-wide liquidity. Lacking an external lender-of-last-resort role and mandatory market-making obligations, the market cannot break this prisoner's dilemma on its own.

The analysis in this chapter shows that the liquidity fragility of perpetual futures markets is a systemic problem. That market makers supply liquidity amply in calm times and shrink it sharply in extreme conditions is a foreseeable outcome of their rational, optimal response to return-risk asymmetry; but foreseeable does not mean unchangeable. Since the collapse is driven by a specific profit equation and specific institutional features, restructuring these underlying logics may inject new resilience into the market. How can more effective incentive mechanisms be designed to change market makers' behavior in a crisis? How can innovative governance frameworks be introduced to overcome the tragedy of the commons in liquidity supply? These explorations of liquidity governance and institutional reconstruction constitute the core content of the next

chapter.

References

[1] Kaiko Research. (2023). *Looking back on FTX's impact*. https://research.kaiko.com/insights/looking-back-on-ftxs-impact

[2] Kaiko Research. (2022, November 15). *The Alameda Gap*. Kaiko Research Weekly, 2022-11.

[3] Akyildirim, E., Conlon, T., Corbet, S., & Goodell, J. W. (2023). Understanding the FTX exchange collapse: A dynamic connectedness approach. *Finance Research Letters*, *53*, 103643. https://doi.org/10.1016/j.frl.2023.103643

[4] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201–2238.

[5] Soros, G. (1987). *The Alchemy of Finance: Reading the Mind of the Market*. John Wiley & Sons.

[6] Anand, A., & Venkataraman, K. (2016). Market conditions, fragility, and the economics of market making. *Journal of Financial Economics*, *121*(2), 327–349. https://doi.org/10.1016/j.jfineco.2016.03.006

[7] Ho, T., & Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, *9*(1), 47–73. https://doi.org/10.1016/0304-405x(81)90020-9

[8] Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217–224. https://doi.org/10.1080/14697680701381228

[9] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100. https://doi.org/10.1016/0304-405x(85)90044-3

[10] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

[11] Diamond, D. W., & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, *91*(3), 401–419. https://doi.org/10.1086/261155

[12] Morris, S., & Shin, H. S. (1998). Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review*, *88*(3), 587–597.

[13] Geanakoplos, J. (2010). The leverage cycle. *NBER Macroeconomics Annual*, *24*(1), 1–65. https://doi.org/10.1086/648285

[14] Shleifer, A., & Vishny, R. W. (1992). Liquidation values and debt capacity: A market equilibrium approach. *The Journal of Finance*, *47*(4), 1343–1366. https://doi.org/10.1111/j.1540-6261.1992.tb04661.x

[15] Deribit Insights. (2021). *The (Flash) Crash of May 19, 2021*. https://insights.deribit.com/exchange-updates/the-flash-crash-of-may-19-2021/

[16] Perpetual Protocol Team. (2021). 2021/2/21 BTC Flash Crash. *Medium*. https:

//blog.perp.fi/2021-2-21-btc-flash-crash-149eef35f7f8

[17] Briola, A., Vidal-Tomás, D., Wang, Y., & Aste, T. (2023). Anatomy of a Stablecoin's failure: The Terra-Luna case. *Finance Research Letters*, *51*, 103358. https://doi.org/10.1016/j.frl.2022.103358

[18] Uhlig, H. (2022). A luna-tic stablecoin crash (No. w30256). *National Bureau of Economic Research*.

[19] Diop, P. O. (2024). An econometric and time series analysis of the USTC depeg's impact on the LUNA classic price crash during spring 2022's crypto market turmoil. *Commodities*, *3*(4), 24.

[20] Allen, F., & Carletti, E. (2008). The role of liquidity in financial crises. In *Proceedings – Economic Policy Symposium – Jackson Hole* (pp. 379–412). Federal Reserve Bank of Kansas City. https://doi.org/10.2139/ssrn.1268367

[21] Degryse, H., de Winne, R., Gresse, C., & Payne, R. (2018). High frequency trading and ghost liquidity. *SSRN Electronic Journal*.

[22] Easley, D., López de Prado, M. M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *The Review of Financial Studies*, *25*(5), 1457–1493. https://doi.org/10.1093/rfs/hhs053

[23] Bessembinder, H., Panayides, M. A., & Venkataraman, K. (2009). Hidden liquidity: An analysis of order exposure strategies in electronic stock markets. *Journal of Financial Economics*, *94*(3), 361–383. https://doi.org/10.1016/j.jfineco.2009.02.001

[24] Berkowitz, J. (2000). *Incorporating liquidity risk into value-at-risk models* (Working Paper). University of California, Irvine.

[25] Bangia, A., Diebold, F. X., Schuermann, T., & Stroughair, J. D. (1999). Modeling liquidity risk with implications for traditional market risk measurement and management. *Risk*, *12*(12), 68–73.

[26] Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, *72*(3), 967–998. https://doi.org/10.1111/jofi.12498

[27] Bouchaud, J. P., Farmer, J. D., & Lillo, F. (2009). How markets slowly digest changes in supply and demand. *Handbook of Financial Markets: Dynamics and Evolution*, 57–160. https://doi.org/10.1016/b978-012374258-2.50006-3

[28] Conlon, T., Corbet, S., & Hu, Y. (2026). The collapse of the FTX exchange: The end of cryptocurrency's age of innocence. *The British Accounting Review*, *58*(3), 101277.

[29] Akhtaruzzaman, M., Boubaker, S., Nguyen, D. K., & Rahman, M. R. (2022). Systemic risk-sharing framework of cryptocurrencies in the COVID–19 crisis. *Finance Research Letters*, *47*(B), 102787. https://doi.org/10.1016/j.frl.2022.102787

[30] Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review*, *105*(2), 564–608. https://doi.org/10.1257/aer.20130456

[31] Galati, L., Webb, A., & Webb, R. I. (2024). Financial contagion in cryptocurrency

exchanges: Evidence from the FTT collapse. *Finance Research Letters*, *67*(A), 105747. https://doi.org/10.1016/j.frl.2024.105747

[32] Hautsch, N., Scheuch, C., & Voigt, S. (2024). Building trust takes time: Limits to arbitrage for blockchain-based assets. *Review of Finance*, *28*(4), 1345–1380. https://doi.org/10.1093/rof/rfae004

[33] Joshi, R. (2025). *Exogenous Stress and Runs on Perpetual Futures: A Global Games Approach*. https://assets.zyrosite.com/dWxb3MBxOpUo84q9/rj_final_jmp-yITW2IpzBgnp7Swl.pdf

[34] Chen, E., Ma, M., & Nie, Z. (2024). Perpetual future contracts in centralized and decentralized exchanges: Mechanism and traders' behavior. *Electronic Markets*, *34*(1), 1–25.

[35] Cont, R., Kotlicki, A., & Valderrama, L. (2020). Liquidity at risk: Joint stress testing of solvency and liquidity. *Journal of Banking & Finance*, *118*, 105871.

[36] Ostrom, E. (2008). The challenge of common-pool resources. *Environment: Science and Policy for Sustainable Development*, *50*(4), 8–21.

The Public-Goods Dilemma and Incentive Design of Liquidity Governance

In July 2025, Hyperliquid, a decentralized perpetual futures exchange, recorded monthly trading volume of approximately $320 billion (corresponding to roughly $10 billion per day); the total value locked (TVL) in its HLP vault crossed $500 million by the end of 2025 (approximately $300 million to $400 million in July 2025; TVL figures are drawn from the DefiLlama hyperliquid-hlp dashboard, accessed June 2026) [1]. These figures signal that the on-chain central limit order book (CLOB) architecture now rivals centralized exchanges (CEXs) in execution efficiency. More importantly, they validate an entirely new paradigm of liquidity governance: through protocol rules and an automated vault mechanism, retail capital is converted directly into the deep liquidity that the market requires.

This democratization of liquidity provision has not come without cost, however. During the JELLY token attack of March 26, 2025, the rigidity of the HLP vault's algorithmic strategy forced it to passively absorb a short position with a notional value of approximately $15.3 million, and its maximum unrealized loss briefly reached approximately $13.5 million (the former is the notional size of the short position absorbed, the latter the peak unrealized loss; the two use different measurement bases) [1]. This extreme event reveals an inherent tension in liquidity-provision mechanisms between everyday efficiency and crisis resilience, which in turn frames the chapter's central question: how do different market architectures acquire liquidity through institutional design, and what vulnerabilities do these designs exhibit when confronting market volatility?

In the traditional architecture of centralized exchanges, liquidity is acquired through private agreements signed between the platform and a small number of professional market makers—in essence a classic principal-agent problem [2]. Under normal conditions, these contracts deliver very high liquidity quality; yet in extreme market conditions they dry up instantly, as market makers exercise their exemption rights and retreat en masse, falling into the public-goods dilemma discussed in Chapter 20 (the full mechanics of its information asymmetry, adverse selection, and moral hazard are detailed in Section 21.2).

By contrast, decentralized exchanges (DEXs)—exemplified by Hyperliquid—treat liquidity as a product that can be manufactured. Through an engineered architecture com-

prising a capital layer, a strategy layer, and a distribution layer, they produce liquidity directly on-chain and democratize its provision, allowing ordinary investors to participate passively in market making (the fundamental differences between this activity's risk-return profile and that of a traditional index fund are detailed in Section 21.4.1). While this approach eliminates centralized privilege and enhances transparency, it also introduces new dimensions of risk through the predictability of its algorithms and the rigidity that comes from a lack of human intervention during crises (the three-layer architecture is detailed in Section 21.3, and the reconstruction of returns and transfer of risk in Section 21.4).

Adopting the perspective of institutional economics, this chapter offers a systematic analysis of two sharply contrasting approaches to liquidity governance. The analysis begins with the principal-agent dilemma inherent in the centralized-exchange procurement model and its structural deficiencies. Building on this, it deconstructs the three-layer architecture of decentralized-exchange liquidity engineering and, using the HLP vault as an in-depth case study, examines the reconstruction of returns and transfer of risk that accompany the democratization of liquidity provision. It then evaluates the effectiveness of existing incentive mechanisms by quantitatively simulating the survival boundary of an independent market maker operating in an on-chain order book environment. Finally, a four-dimensional framework—effectiveness, incentive compatibility, transparency, and resilience—provides an analytical tool for comparing the relative merits of the two paradigms and their coexistence. After reading this chapter, readers will understand how market architectures differ in the way they allocate risk-bearing, distribute returns, and assign crisis responsibility, and will recognize how, within decentralized finance, code-driven liquidity infrastructure resolves the traditional crisis of trust while giving rise to new systemic vulnerabilities.

21.1 The core problem of liquidity governance

The analysis in Chapter 20 shows that the countercyclical nature of liquidity provision—liquidity contracts precisely when the market needs it most—produces a systemic tragedy of the commons. Building on this, this section distills the central tension of liquidity governance and refines Chapter 20's characterization of liquidity as a *quasi-public good* into that of a common-pool resource: under normal conditions it is approximately non-excludable, but during windows of stress it becomes rivalrous as order book depth is rapidly consumed. The section then surveys the two institutional responses that have emerged in crypto-asset markets and constructs a four-dimensional evaluation framework for systematic comparison.

21.1.1 The tragedy of the commons

As discussed in Chapter 20, the microstructure of crypto-asset markets contains a fundamental contradiction: the supply of and demand for liquidity are negatively correlated. Under normal conditions, market makers face relatively low adverse selection risk and are willing to supply ample liquidity. When an extreme event triggers sharp price swings and a surge in trading demand, however, market makers' risk exposure expands abruptly, and their optimal rational strategy is to withdraw quotes to protect

their own capital. This phenomenon—in which liquidity vanishes precisely when it is needed most—is not merely the rational choice of individual agents but the root cause of systemic crises.

In decentralized exchanges, this phenomenon manifests as a classic tragedy of the commons. In economic terms, liquidity resembles a common-pool resource more than a pure public good: all traders benefit from a deep order book or an ample liquidity pool (non-excludability), yet once a limit order is consumed it no longer exists (instantaneous rivalry). Although market makers can repost quotes on a millisecond timescale, in extreme conditions quotes regenerate far more slowly than they are consumed, which makes the tragedy-of-the-commons framework fully applicable at high-frequency timescales and during periods of market stress. The costs and risks of providing and maintaining this resource are borne by individual liquidity providers. Capponi and Jia (2025) [3] show that on blockchain infrastructure, liquidity providers face a distinctive collective-action dilemma: the value that arbitrageurs extract from a liquidity pool is mutualized as a loss across all liquidity providers, whereas the transaction cost and time delay of withdrawing liquidity must be borne by each provider alone. Facing adverse market conditions, an individual liquidity provider has a strong incentive to retreat ahead of the others, and this individually rational run ultimately causes aggregate liquidity to collapse. To mitigate this adverse selection risk, liquidity providers typically adopt a more convex, defensive pricing curve, but doing so inevitably reduces capital efficiency under normal conditions.

The market cannot resolve this tragedy of the commons on its own. In the absence of a binding coordination mechanism, market makers fall into a prisoner's dilemma: if all of them maintain their quotes, the market may weather the crisis smoothly, but the moment any one begins to retreat, the adverse selection risk borne by those who remain rises exponentially. The empirical study by Anand and Venkataraman (2016) [4] confirms that under adverse market conditions, systematic withdrawal by market makers significantly degrades market quality and amplifies price volatility. How to overcome the limits of individual rationality—through external intervention or internal mechanism design—so as to guarantee a liquidity floor during crises therefore constitutes the central problem of liquidity governance. This problem bears not only on the survival of any single trading platform but also directly determines the level of systemic risk across the entire decentralized finance system.

21.1.2 Two institutional approaches

In response to the systemic failure of liquidity provision, crypto-asset markets have evolved two sharply contrasting institutional responses. Centralized exchanges primarily adopt a procurement model, acquiring liquidity services through private agreements signed with a small number of professional market makers. Under this model, the exchange plays the role of principal, using economic incentives—rebates, fee discounts, and privileged access to advanced APIs—to require market makers to fulfill quoting obligations under specified market conditions. The essence of this arrangement is to treat liquidity as a professional service that can be purchased from external suppliers.

In execution, however, the procurement model faces severe information asymmetry and

moral hazard. An exchange can neither monitor a market maker's internal risk models and true intentions in real time nor fully verify whether the maker exerted its best efforts during a crisis. Alexander (2025) [5] notes that the design of market-maker incentives directly affects the liquidity and price stability of crypto assets, but that an ill-suited incentive structure can lead market makers to invoke exemption clauses and retreat legally under extreme conditions. This governance structure, grounded in bilateral contracts, can deliver very high market depth under normal conditions, yet it repeatedly proves fragile during tail-risk events. When a systemic shock strikes, the market maker's rational choice is to trigger the abnormal-market-conditions clause and suspend its quoting obligations—precisely at the moment the market needs liquidity most.

By contrast, decentralized exchanges exemplified by Hyperliquid have charted an institutional path built on an engineering model. This model abandons reliance on external suppliers and instead manufactures liquidity directly on-chain through protocol rules and smart contracts. Its core lies in decomposing liquidity provision into three programmable modules: capital aggregation, strategy execution, and revenue distribution. By establishing a public vault, the protocol allows any user to deposit assets and become a liquidity provider, thereby democratizing liquidity provision. A built-in algorithmic engine replaces the human market maker, automatically posting two-sided quotes on the order book according to preset logic. This paradigm shift transforms liquidity governance from a contractual game between people into code-driven systems engineering.

The engineering model eliminates the agent's discretion through hard-coded rules. Under normal market conditions, code execution is immune to emotion and will not withdraw quotes on the basis of sentiment or profit motive, thereby providing the market with a degree of floor liquidity (under extreme boundary conditions this execution logic may still be overridden by governance intervention; see Section 21.4.6). This rigidity, however, brings new challenges: an algorithm cannot process complex cross-market information as flexibly as a human market maker, and vault capital may suffer large directional losses in a one-sided market. In essence, then, the engineering model trades hard capital losses for the rule-verifiability of liquidity provision.

21.1.3 A four-dimensional evaluation framework

A systematic comparison of the two liquidity-governance paradigms requires a multidimensional evaluation framework—one that measures not only market quality under normal conditions but also system performance under extreme conditions. Specifically, an institutional design can be assessed along four dimensions: effectiveness, incentive compatibility, transparency, and resilience.

Effectiveness measures an institutional design's ability to supply ample, low-cost liquidity in normal markets. By bringing in top quantitative teams with advanced technology and deep experience, the procurement model typically achieves extremely narrow bid-ask spreads and enormous order book depth. By contrast, the engineering model—constrained by on-chain computing resources and algorithmic rigidity—is at present often slightly inferior in capital efficiency and pricing precision.

Incentive compatibility concerns whether the interests of individual participants remain

aligned with the system's overall objectives. In the context of the tragedy of the commons, the central question is whether the mechanism can internalize the externality of liquidity withdrawal. The procurement model binds market makers through contractual obligations, but their profit-maximizing instinct keeps them searching for loopholes to evade those obligations. The engineering model, by tightly binding retail capital to protocol revenue, creates a shared-fate configuration of interests, yet incentive misalignment persists at the point where arbitrageurs extract value. In a CLOB environment, the mechanism of value extraction differs from that of an automated market maker (AMM) but is no less significant. Drawing on empirical evidence from traditional limit order book markets, Aquilina, Budish, and O'Neill (2022) [6] find that latency arbitrage accounts for roughly 33% of the effective bid-ask spread and imposes an implicit cost of about 0.5 basis points on each trade. Foucault, Kozhan, and Tham (2017) [7] define such fleeting arbitrage opportunities as *toxic arbitrage*, demonstrating that the speed advantage of high-frequency arbitrageurs, while improving price efficiency, systematically intensifies market makers' adverse selection risk. In crypto-asset spot markets, a spread decomposition based on Bitfinex data by Tiniç et al. (2023) [8] shows that adverse selection costs account for roughly 10% of the effective spread, with BTC exhibiting the highest inventory-cost share among all crypto assets. By selling a speed advantage to particular market makers through API privileges and colocation services, exchanges also entrench this value-extraction structure at the institutional level. Although Capponi and Jia (2025) [3] report a more extreme value-transfer ratio in the AMM setting (the median arbitrageur transfers 96% of profit to the infrastructure layer), the value-extraction mechanisms of AMMs and CLOBs differ fundamentally in nature: a CLOB market maker can actively cancel orders, and the value erosion it faces stems chiefly from latency arbitrage and information asymmetry, rather than from the passively borne losses and rebalancing costs of liquidity providers in an AMM.

Transparency reflects market participants' ability to observe the system's operating state and the enforcement of its rules. The procurement model rests on confidential agreements and undisclosed privileges, leaving ordinary traders unable to know a market maker's true obligations or the subsidies it receives. This information black box not only raises doubts about market fairness but also conceals latent systemic risk. The engineering model, by contrast, deploys all rules, parameters, and fund flows on the blockchain, achieving full open-sourcing and auditability. Any user can monitor the vault's asset position and the algorithm's execution logic in real time, and this informational symmetry is a precondition for building trustless liquidity infrastructure.

Resilience directly determines a system's ability to survive a severe shock. Resilience requires that a liquidity-provision mechanism function well not only when markets operate stably but also provide a solid floor at moments of crisis. The fragility of the procurement model lies in its dependence on commercial entities that possess free will—entities that, when faced with the threat of insolvency, will inevitably choose self-preservation. The engineering model, through deterministic algorithmic execution and the continuous locking of public capital, builds a liquidity base that is not driven by emotion. Although this rigidity can cause a vault to endure a substantial drawdown in a one-sided market, it objectively preserves a counterparty of last resort for the market.

Table 21-1 integrates these four dimensions into a unified evaluation matrix, listing

for each dimension its core concern and the specific performance it requires in normal markets and in crisis markets. This framework serves throughout the remainder of the chapter as a comparative tool for systematically assessing the relative merits of the CEX procurement model and the DEX engineering model along each dimension.

Evaluation dimension	Core concern	Required performance in normal markets	Required performance in crisis markets
Effectiveness	Transaction cost and depth	Extremely narrow spreads; low slippage on large orders	Maintain basic two-sided quoting
Incentive compatibility	Alignment of interests	Market-making returns exceed the opportunity cost of capital	Expected return from holding quotes exceeds the cost of default
Transparency	Degree of informational symmetry	Rules public; fund flows traceable	Risk exposures transparent; the liquidation process auditable
Resilience	Shock resistance	Adapt to everyday volatility; parameters adjust automatically	Avoid sentiment- or profit-driven discretionary withdrawal; keep any crisis intervention transparent and auditable

Table 21-1. A four-dimensional framework for evaluating liquidity governance (Data source: constructed by the author; analytical framework adapted from Madhavan (2000) [9])

These four dimensions are not independent but mutually constraining: pursuing maximal effectiveness often requires sacrificing some transparency to protect a market maker's strategy secrets, while enhancing resilience may require accumulating redundant capital under normal conditions and lowering short-term capital efficiency. Whether procurement or engineering, then, each model is at bottom a trade-off among these four dimensions. The central direction for future innovation in liquidity governance is to preserve the transparency and resilience of the engineering model while gradually improving its effectiveness and incentive compatibility through more advanced algorithms and mechanism design.

To prevent confusion among the several numbered frameworks that follow, a brief map is useful. The four-dimensional framework—effectiveness, incentive compatibility, transparency, and resilience—is the evaluative backbone that runs through the entire chapter. The three-layer architecture—capital layer, strategy layer, and distribution layer—is the descriptive skeleton of the DEX engineering model. The five risk categories and three revenue engines that appear later in Section 21.4 are a further decomposition of the internal workings of the HLP vault. Three superficially similar binary contrasts must be kept distinct, because they compare different objects: the *two paradigms* refers to the two institution-level approaches of CEX procurement and DEX engineering; the *two routes* refers to the two submodels within a DEX, namely "protocol as market maker" and "protocol recruits market makers"; and the *active-passive dual track* refers to the two participant identities of the independent market maker and the vault par-

ticipant. Each numbered framework can be mapped back onto the four-dimensional backbone.

21.2 The principal-agent model of CEX liquidity procurement

Within the microstructure of cryptocurrency markets, the relationship between a centralized exchange and its market makers forms the central axis of liquidity creation. This relationship is not a simple sale-and-purchase contract, however, but is deeply rooted in a principal-agent problem set within an environment of information asymmetry. As principal, the exchange's fundamental objective is to provide platform users with narrow bid-ask spreads, a deep order book, and stable liquidity under extreme market conditions. As agent, the market maker is motivated by profit maximization and risk minimization, which inevitably leads it, under certain market conditions, to take actions contrary to the exchange's overall interests. This section dissects the liquidity-procurement model of centralized exchanges and analyzes how market-maker agreements design incentives under the constraints of adverse selection and moral hazard.

21.2.1 Information asymmetry

Principal-agent theory provides an analytical framework for understanding the liquidity-procurement mechanism of centralized exchanges. Within this framework, the exchange delegates the task of liquidity provision to professional market-making firms but confronts a fundamental obstacle: it cannot fully observe, in real time, the true quality and intent of a market maker's conduct. This information asymmetry exists not only ex ante (before signing) but persists ex post (after signing) throughout the entire trading process.

Figure 21-1 presents the internal structure of this principal-agent relationship: information asymmetry manifests ex ante as adverse selection and ex post as moral hazard, and when extreme market conditions trigger the exemption clause, the two erupt simultaneously, causing liquidity to evaporate rapidly at the moment it is needed most.

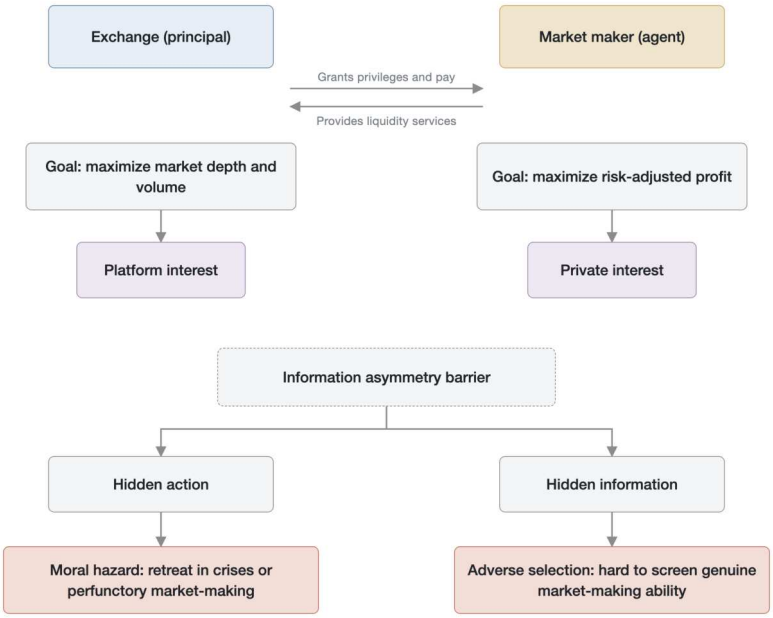


Figure 21-1. The structure of the CEX principal-agent relationship (A conceptual structural diagram constructed by the author from principal-agent and microstructure theory; nodes and arrows denote mechanism relationships, not measured data; the adverse selection branch follows Glosten & Milgrom 1985 [2])

In classical microstructure theory, the principal risk a market maker faces is defined as adverse selection risk. Under the classical model proposed by Glosten and Milgrom in 1985, a market maker supplying liquidity cannot distinguish whether its counterparty is an informed trader holding private information or a noise trader who trades solely out of liquidity needs [2]. To compensate for the expected loss that necessarily arises when trading against informed traders, the market maker must set a bid-ask spread, the width of which is directly related to the proportion of informed traders in the market. Kyle developed this framework further in 1985, constructing a three-party dynamic model comprising a single informed insider, random noise traders, and competitive market makers, and showing that the informed trader hides its informational advantage by spreading trades over time, while market makers can absorb the market's information only gradually, through continuous price adjustment [10].

In the principal-agent relationship of a centralized exchange, however, the dimension of information asymmetry undergoes a profound inversion. Although the exchange possesses global order-flow data within its own platform, modern cryptocurrency market makers typically conduct high-frequency cross-market arbitrage across dozens of centralized and decentralized platforms at once, which makes the exchange the relatively *uninformed* party. Exploiting its cross-platform informational advantage and order-execution capability, a market maker can behave, on the surface, in ways that comply with the agreement on a single exchange while in substance pursuing strategies that may harm that exchange's long-term liquidity. This informational disadvantage is espe-

cially pronounced in cryptocurrency markets, because extreme fragmentation prevents any single platform from obtaining complete global price-discovery information, and the cross-platform data-aggregation capability of top market makers is precisely what constitutes their core competitive barrier.

This bidirectional information asymmetry gives rise to two classic economic problems. The adverse selection problem appears at the contracting stage: the exchange struggles to identify accurately which market makers possess the genuine ability to maintain quotes amid extreme volatility and which are merely speculators relying on simple statistical-arbitrage strategies to extract rebates in calm markets. The moral hazard problem appears after the agreement takes effect: a market maker may exploit its informational advantage and technical means to appear to fulfill its quoting obligations while using devices such as ghost liquidity to evade any substantive risk-bearing. In practice, ghost liquidity takes at least three distinct forms [11]. The first is quote flickering, in which extremely short-lived limit orders are canceled before they can be filled; this may be legitimate risk management or a deliberate display of phantom liquidity. The second is signal manipulation, in which only the best price level in a multi-tier order stack carries genuine trading intent while the deeper orders exist to fabricate order book depth. The third is cross-platform double-counting, in which the same source of capital posts orders on multiple platforms simultaneously, causing aggregate market liquidity to be systematically overestimated. The three carry entirely different implications for principal-agent analysis: the first is detectable through a cancellation-rate metric, the second requires dynamic analysis of multiple depth tiers, and the third lies beyond the monitoring reach of any single exchange. To meet these challenges, exchanges have had to design complex market-maker agreements that align the two parties' interests through finely calibrated incentive and constraint mechanisms.

21.2.2 The structure of the agreement

The market-maker agreement is the core contractual instrument through which a centralized exchange addresses the principal-agent problem. These agreements are generally not disclosed to the public, but by analyzing compliance filings made public by regulators such as the Commodity Futures Trading Commission (CFTC), together with prevailing industry practice, one can clearly outline their internal structure of interest exchange [12]. Every clause of a market-maker agreement is, in essence, an act of pricing and allocating risk for liquidity, an intangible asset.

As shown in Figure 21-2, the market-maker agreement distilled from public CFTC compliance filings rests on three pillars—obligations, compensation, and exemptions—with monitoring and enforcement mechanisms at the base forming a safeguard layer; the boundary of obligation drawn by the exemption clause is precisely the source of systemic fragility in extreme market conditions.

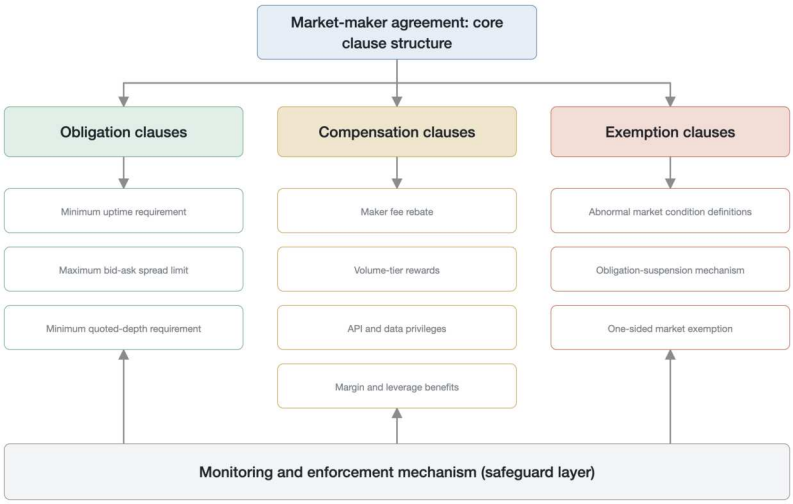


Figure 21-2. The clause structure of a market-maker agreement (A schematic of the three pillars—obligations, compensation, and exemptions—with the monitoring-and-enforcement safeguard layer at the base; not measured data; clauses drawn from the CFTC market-maker agreement filing [12] and Section 21.2.2 of the text)

The structure of the agreement generally rests on three pillars. The first is the obligations clause, the minimum behavioral baseline set by the exchange as principal. Typical obligations include a maximum-spread constraint, a minimum-depth requirement, and a strict uptime guarantee. An agreement might, for example, require the market maker to maintain continuous two-sided quotes for the life of the contract (the specific uptime threshold varies by exchange and reflects prevailing industry practice), to keep the bid-ask spread within a specified number of basis points, and to provide a minimum quantity of liquidity at the best bid and offer [12]. These quantitative metrics aim to ensure that retail investors can execute trades at a reasonable slippage cost for the vast majority of the time. In addition, to prevent market makers from manufacturing false prosperity through high-frequency cancellation strategies, some agreements impose strict caps on order lifetime or on the order-cancellation rate. On latency, agreements typically stipulate that a market maker's quote-update response time must fall below a specified threshold, so that stale quotes do not become an arbitrage tool for other traders during rapid market swings.

In return for bearing these obligations, the compensation clause forms the agreement's second pillar. A centralized exchange pays its liquidity-procurement costs mainly through the fee structure and technical privileges. On fees, the maker rebate is the central economic incentive. When a market maker's limit order is executed, it not only pays no trading fee but receives back a portion of the fee the exchange charges the taker. This negative-fee structure directly subsidizes the market maker's adverse selection costs and inventory risk [5]. On technology, exchanges commonly grant core market makers API privileges, including higher request rate limits, lower system latency, and in some cases a priority matching channel or a dedicated order-flow data feed. These privileges effectively transfer a system-level microstructure advantage to the market

maker as an implicit subsidy. Some exchanges also extend inventory-financing facilities to top market makers, lending them digital assets held on the platform at very low interest rates, thereby lowering the makers' cost of capital and enabling them to maintain quotes across more trading pairs at once.

The agreement's third pillar—the exemption clause—reveals the fragility of this contractual relationship. Analysis of public market-maker agreements shows that a market maker's quoting obligations are automatically suspended under a range of circumstances, including a halt in the calculation of underlying market data, the absence of any direct underlying market from which to derive a price, a determination by the exchange that unforeseeable international financial or political-economic conditions have made accurate pricing impossible, and force majeure events such as war, natural disaster, or technical failure [12]. This clause design is reasonable in legal and operational terms, but it introduces structural risk: the moment an *abnormal condition* is triggered is often precisely the moment the market is undergoing violent price swings and demand for liquidity peaks.

Table 21-2 summarizes the core clauses of these three pillars and, from the perspective of principal-agent theory, annotates the specific function each clause serves in constraining market-maker behavior. The progression from the obligations clause to the exemption clause clearly displays the institutional logic by which risk-bearing is transferred, step by step, from the market maker to the exchange and the end user.

Clause type	Core content	Principal-agent function
Maximum-spread constraint	Upper limit on the two-sided bid-ask spread (in basis points)	Limits the market maker's ability to pass adverse selection costs on by widening the spread
Minimum-depth requirement	Minimum quantity at each quote level	Prevents the market maker from offering token "paper liquidity"
Uptime requirement	Maintain continuous quotes for the life of the contract (specific threshold varies by exchange)	Ensures the temporal continuity of liquidity
Maker rebate	Negative-fee mechanism; fees rebated in proportion to volume	Compensates the market maker's adverse selection costs and inventory risk
API privileges	Priority channel; higher rate limits	Lowers the market maker's technical operating costs
Abnormal-market exemption	Suspends all obligations under extreme conditions	Transfers tail risk back to the exchange and users

Table 21-2. Principal-agent function analysis of the core clauses of a market-maker agreement (Data source: CFTC market-maker agreement filing [12])

21.2.3 Adverse selection

Before a market-maker agreement is signed, the exchange faces a classic adverse selection dilemma. Because a market maker's true strategy, risk appetite, and capital depth are closely guarded commercial secrets, the exchange struggles to distinguish ex ante between the *high-quality* maker willing to bear risk amid volatility and the

low-quality maker that merely harvests rebates in calm periods. If the incentives the exchange offers are too generous, they will inevitably attract a wave of high-frequency speculators bent on exploiting system loopholes; if the terms are too onerous, genuinely capable institutional market makers may exit for competing platforms with richer incentives. This dilemma is especially acute in cryptocurrency markets because, unlike the designated-market-maker system of traditional financial markets, a crypto exchange's market-maker program is usually non-exclusive: makers can operate on several platforms at once, so the exchange's screening mechanism perpetually faces the threat of cross-platform arbitrage.

To mitigate this adverse selection problem, leading centralized exchanges have developed a complex screening and dynamic-tiering mechanism. The first element is an entry barrier based on historical performance. Some top exchanges' market-maker programs, for example, require applicants to have met an extremely high trading-volume threshold over the preceding 30 days, taking this as an initial signal of capital strength and execution capability [13]. Such a high threshold effectively performs the role described in Spence's signaling model, because for a speculator lacking genuine market-making ability, the cost of fabricating such large and sustained volume is prohibitively high. At the same time, some exchanges require applicants to submit detailed strategy descriptions and risk-management frameworks and assess their system stability and emergency-response capability through in-depth interviews with the exchange's technical team.

Second, exchanges make wide use of a performance-based dynamic tiering system. Market makers are no longer treated as homogeneous liquidity providers but are sorted into tiers—typically ranging from ordinary to platinum or diamond—according to their actual performance over the preceding evaluation period. The assessment metrics include not only share of trading volume but also finer microstructure dimensions of liquidity quality, such as the proportion of time with orders posted within a specified depth, quote continuity during sharp volatility, the order-cancellation rate, and the stability of spread maintenance. A market maker's tier directly determines the rebate rate and level of API privileges it enjoys in the next period. This dynamic-game mechanism forces market makers to weigh long-term interests against short-term arbitrage. If a market maker responds to a bout of market turbulence by canceling orders en masse to avoid inventory risk, its rating in the system is swiftly downgraded, costing it future high-rebate privileges—a loss that often far exceeds what it avoided by canceling during that single crisis.

Although these screening mechanisms raise the overall quality of liquidity providers to some degree, they cannot eliminate adverse selection entirely. Because cryptocurrency markets are extremely fragmented, a market maker can readily engage in regulatory arbitrage across exchanges—dumping high-risk inventory onto platforms with weaker safeguards while maintaining only risk-free statistical-arbitrage strategies on the platform with the richest incentives. This cross-platform concealment of strategy keeps any single exchange's screening mechanism perpetually at an informational disadvantage. The more fundamental problem is that a market maker's *true capability* is itself a dynamic concept that shifts with the state of the market: a maker that excelled in the 2021 bull market may, amid the systemic deleveraging wave of 2022, be unable to main-

tain quotes because of capital depletion. The effectiveness of historical performance as a screening signal is often sharply diminished in the face of a structural market shift.

21.2.4 Moral hazard

If adverse selection is ex ante information asymmetry, moral hazard is the market maker's use of its informational advantage, after the agreement takes effect, to harm the principal's interests. In calm day-to-day markets, this moral hazard typically manifests as a proliferation of *ghost liquidity*. By placing large numbers of readily cancelable limit orders in the order book, a market maker superficially satisfies the agreement's minimum-depth and maximum-spread requirements and thereby legally captures the exchange's rebates and privileges. These orders, however, tend to carry extremely high cancellation rates: at the slightest anomaly in market price, or when a genuine taker order attempts to fill, this seemingly deep liquidity vanishes within milliseconds [11]. In substance, the market maker fulfills its paper obligations while refusing to bear any substantive adverse selection risk.

A moral hazard of greater consequence erupts during extreme market crises. When the market suffers a violent shock and volatility spikes, the exchange most needs its market makers to supply stable liquidity to cushion the price collapse. Yet it is precisely at this critical moment that the exemption clause in the market-maker agreement so often becomes the maker's basis for a legal retreat. Analysis of public market-maker agreements shows that a maker's quoting obligations are suspended once the exchange determines that *abnormal market conditions* exist or that an event has rendered accurate pricing impossible [12]. This power of determination is typically exercised unilaterally by the exchange, and the definition of *abnormal market conditions* is itself so vague that it leaves enormous room for interpretation on both sides.

This institutional design produces a structural contradiction: the market maker receives generous subsidies when the market least needs liquidity yet is legally released from its obligations when the market needs liquidity most. Academic research shows that even a tightly regulated designated market maker does not necessarily behave in ways beneficial to the market when confronting an extreme event akin to a *flash crash*. When high-frequency traders withdraw in unison, designated market makers can provide some cushion at first, but in the face of enormous inventory-devaluation risk and extreme information asymmetry they too will ultimately choose to widen spreads sharply or even cease quoting altogether [14]. In cryptocurrency markets, which lack the central bank's lender-of-last-resort support found in traditional financial markets, such collective liquidity withdrawal often triggers catastrophic cascading liquidations, amplifying what began as a localized price move into a systemic evaporation of liquidity.

Another dimension of moral hazard lies in the market maker's active management of inventory risk. Under normal market conditions, a maker covers its adverse selection costs through high-frequency bid-ask spread income and controls its net inventory exposure through cross-platform hedging. When the market enters a one-sided trend, however, the maker's hedging costs rise sharply, and its quoting within the scope of its contractual obligations effectively becomes a compulsory absorption of losses. In this situation, the maker has a strong incentive to evade such passive risk-bearing by

various means, including deliberately triggering the technical conditions of the exemption clause, applying to the exchange for a temporary waiver, or effectively exiting the market by widening spreads to an extreme. The root of this behavior is that the market-maker agreement never truly resolves the central paradox of liquidity provision: the cost of providing liquidity is highest precisely when the market needs it most.

21.2.5 Proprietary market making

When the procurement model fails in a crisis because of the exemption clause, many centralized exchanges do not stop at contractual adjustment but move to the opposite extreme: internalizing the market-making function. This section shows that this path likewise fails to resolve the principal-agent dilemma and instead introduces a more serious conflict of interest. Confronted with the unreliability of external market makers in a crisis, these exchanges adopt a more radical solution—building a proprietary market-making team or forming a deep equity-binding relationship with a particular market maker. On the surface, this model appears to eliminate the principal-agent problem entirely, since the principal (the exchange) and the agent (the market maker) become a single economic interest. Far from solving the problem, however, this internalization gives rise to a more serious structural conflict of interest.

When the exchange simultaneously plays the multiple roles of trade matcher, rule maker, and market participant, its proprietary market-making desk inevitably gains an absolute informational advantage over other participants. First, the proprietary desk can obtain global order-flow data with zero latency, including every client's hidden orders, stop-loss levels, and leverage-liquidation thresholds. Second, it enjoys physical-level minimum latency in the system architecture and may even hold priority-execution rights inside the matching engine itself. This extreme asymmetry of advantage creates the structural conditions for market manipulation in various forms. The regulatory frameworks of traditional financial markets—such as the U.S. Volcker Rule, that is, Section 619 of the Dodd-Frank Act—rest precisely on a keen awareness of this conflict of interest: the rule bars regulated banking institutions from short-term proprietary trading but grants an explicit exemption for genuine market-making activity. Its institutional essence is to require banks, at the same trading desk, to draw a strict line between client-driven market making and proprietary trading undertaken for the bank's own account, and to prevent proprietary trading from evading the ban under the guise of market making through conditions such as inventory caps and compensation design.

Historical experience shows that proprietary trading at cryptocurrency exchanges is often accompanied by serious compliance risk. Some platforms use proprietary accounts for wash trading, artificially inflating volume through matched trades between their own accounts to manufacture the illusion of a thriving market and attract retail investors [15]. In certain extreme conditions, an exchange's proprietary market-making desk may exploit its precise knowledge of clients' liquidation levels to deliberately launch a targeted price shock, force client positions into liquidation, and reap outsized profits amid the resulting volatility. Such conduct betrays the exchange's fundamental character as neutral market infrastructure and transforms the principal-agent problem from "the market maker deceiving the exchange" into "the exchange and market maker together

deceiving the end user."

A deep equity-binding relationship between an exchange and a particular market maker produces a similar conflict of interest even short of full internalization. When a market maker holds equity in the exchange, or the exchange grants a particular maker exclusive information privileges unavailable to other participants, the relationship has in substance breached the boundary of a fair market, turning the entire platform into a rent-seeking instrument for a particular interest group.

21.2.6 Structural deficiencies

Taken together, the existing liquidity-procurement model of centralized exchanges suffers from structural deficiencies that are difficult to overcome. The root difficulty is that liquidity—a highly dynamic, state-dependent public good—cannot be effectively constrained and priced through a static, rigid bilateral contract.

Within the principal-agent framework, the exchange disciplines its market makers through complex incentive and penalty mechanisms yet can never break through the barrier of information asymmetry. Drawing on their technical advantage and cross-market vantage point, market makers can always find strategies that satisfy the letter of the agreement while minimizing their own risk. This space for compliance arbitrage is endogenous to the agreement's design, because any incentive mechanism based on observable metrics must confront Goodhart's law: once a metric becomes a control target, it ceases to be a good metric. When the exchange makes *uptime* and *spread width* its core metrics for evaluating market makers, the makers focus on excelling along those dimensions while relegating what truly matters—liquidity quality, that is, resilience under stress—to secondary importance.

The very existence of the exemption clause is, moreover, a fundamental admission of this contract's impotence before an extreme crisis. Each exemption clause is an implicit *insurance contract* that transfers tail risk from the market maker back to the exchange and the end user. And when the exchange internalizes this contradiction through proprietary market making, it falls into a still more serious conflict of interest, harming market fairness and transparency.

These structural deficiencies show that reliance on the contractual design and microstructure adjustments of centralized institutions alone cannot fundamentally resolve the problem of liquidity provision in cryptocurrency markets. So long as the market maker's profit motive diverges fundamentally from the exchange's character as a public platform, and so long as room for information asymmetry and regulatory arbitrage remains, this principal-agent-based procurement mechanism will remain in a fragile dynamic equilibrium. This predicament also provides the deep real-world backdrop for the rise of alternative liquidity-provision mechanisms in decentralized finance, such as the automated market maker: in an on-chain protocol, rules are encoded as immutable smart contracts, and liquidity providers' behavior is constrained by mathematics rather than by contract, which fundamentally alters the informational structure of the principal-agent relationship.

21.3 The three-layer architecture of DEX liquidity engineering

Unlike the CEX procurement model, decentralized exchanges treat liquidity as a product that can be manufactured through protocol mechanisms. Taking Hyperliquid as its prototype, this section deconstructs the capital layer, strategy layer, and distribution layer of this engineering approach and compares the two DEX liquidity routes.

21.3.1 The paradigm shift

The fundamental divergence between CEXs and DEXs in liquidity governance stems from differing conceptions of the nature of liquidity. As discussed in Section 21.2, a CEX treats liquidity as a *service* to be *procured* from external professional institutions through contractual constraints and fee rebates; its effectiveness depends heavily on a principal-agent relationship under information asymmetry [16], and it faces systemic risk when, under an extreme shock, market makers invoke the exemption clause to withdraw liquidity [17].

By contrast, a new generation of DEXs exemplified by Hyperliquid proposes a wholly different governance paradigm: treating liquidity as a product that can be *manufactured* through protocol mechanisms. From this engineering vantage point, liquidity is no longer the preserve of a few professional institutions but the systematic output of a decentralized protocol that converts raw material (retail capital) into a finished good (order book quotes) [3]. The core of this paradigm shift is to reconstruct market making—once dependent on human trust and private gamesmanship—as an automated process driven by smart contracts and algorithms. By establishing a transparent system of rules, a DEX reduces the moral hazard of the traditional model and democratizes liquidity provision [18]. The engineering approach does not eliminate the principal-agent problem, however; it changes its form and its distribution of risk. Under the vault model, depositors delegate their capital to protocol code rather than to a human market maker, and the focus of information asymmetry shifts from *the market maker's true intent is unobservable* to *the effectiveness and safety of the algorithmic strategy cannot be fully verified*. This transformation reduces the agent's discretion but introduces new dimensions of risk, including algorithmic rigidity, parameter governance, and smart-contract vulnerabilities.

In this engineering system, participants are redefined. On one side, ordinary investors can become indirect liquidity providers by injecting funds into the protocol's vault; on the other, technically capable independent market makers can still connect through the API to compete directly in the market. The two are not mutually exclusive but together form the foundational base and the elastic margin of the DEX liquidity ecosystem. By decoupling capital aggregation, strategy execution, and revenue distribution, this system design offers a structural alternative for resolving the public-goods dilemma of liquidity.

21.3.2 The capital layer

In the three-layer architecture of liquidity engineering, the capital layer is the power source of the entire system. Its core mechanism is the liquidity vault, which aggregates dispersed, small-scale retail funds into market-making capital with scale effects. In the traditional CEX model, a market maker must bring vast proprietary funds or raise institutional financing to meet the steep entry threshold, which turns liquidity provision into a highly concentrated oligopolistic game [19]. The vault mechanism breaks this barrier through smart contracts and democratizes the supply of capital.

Investors need only deposit stablecoins into the designated vault smart contract to receive token certificates representing their share of the pool. These funds no longer sit idle but are marshaled by the protocol as the underlying assets that support subsequent market-making activity. From the investor's perspective, the process resembles buying an index fund focused on a market-making strategy (the limits of this analogy are clarified in Section 21.4.1); from the protocol's perspective, the vault is a bridge between retail capital and professional market-making demand [20].

Vault design is not fixed; protocol developers must make trade-offs along several dimensions. A single-vault model, for example, maximizes the efficiency of capital utilization but readily transmits risk across assets when making markets in multiple assets; a multi-vault model lets investors target their exposure to the risk characteristics of different assets but may fragment liquidity. In addition, to maintain the stability of market-making capital, a vault typically introduces redemption limits or a capacity cap. These design details directly determine the capital layer's resilience under extreme market conditions, because allowing investors to withdraw funds without limit during a crisis would turn the vault itself into a fuse that accelerates a market collapse.

21.3.3 The strategy layer

The capital layer completes the aggregation of funds; the strategy layer is responsible for converting this static capital into dynamic two-sided quotes on the order book. In Hyperliquid's architecture, this conversion is performed automatically by a built-in market-making strategy engine. The engine is, in essence, an automated algorithm running at the protocol layer: it reads market data in real time, computes the optimal bid-ask spread, and posts the corresponding orders on the CLOB [21].

The strategy engine's core task is to balance the depth of liquidity supplied against the risk the market maker faces. Its key parameters include quote width, quote depth, inventory-management thresholds, and the ratio of capital allocated across assets. When the market is in a normal state, the engine keeps spreads relatively narrow to capture trading fees; when volatility spikes, it automatically widens spreads or reduces posted size to lower adverse selection risk [2]. Unlike the passive pricing curve of a traditional AMM, the strategy engine can actively manage inventory: by adjusting the asymmetry of its quotes, it steers order flow to work off one-sided risk exposure.

This "protocol as market maker" design makes the strategy engine functionally equivalent to a professional quantitative market-making team, yet its operating logic is driven entirely by open, transparent code. On-chain auditability eliminates the black-box op-

erations of the traditional model and ensures fairness in execution. This transparency brings an endogenous paradox, however: publicly visible strategy parameters are easily reverse-engineered and gamed by high-frequency traders or maximal extractable value (MEV) searchers [22]. Moreover, because parameter changes typically must pass through a governance process, the strategy engine's response to a sudden black-swan event is often slower than that of a human-operated centralized market maker.

This architecture also involves a trust assumption that is easily overlooked. Hyperliquid currently uses a centralized sequencer to process trade instructions, and the sequencer controls the ordering and execution priority of orders. As a protocol-native component, the HLP strategy's order submission and execution path may enjoy priority, which is functionally similar to the API privileges in a CEX market-maker agreement. This fact means that the claim in Section 21.3.1 of delegation "to code rather than to a person" requires further qualification: although the strategy code is publicly auditable, it runs on sequencer infrastructure controlled by the core team. Trust is not eliminated entirely but shifted—from trust in an individual market maker's conduct to trust that the sequencer operator will not abuse its execution priority. As Hyperliquid advances the progressive decentralization of its sequencer, the strength of this trust assumption will gradually diminish; but at the current stage, it is inaccurate to describe HLP as a fully trustless liquidity-provision mechanism.

Beyond the capital, strategy, and distribution layers, the system as it actually operates also contains a risk-control layer that cuts across all three. The strategy engine's execution is not unconstrained; the protocol imposes multiple risk-control parameter limits on it, including a maximum position size for any single asset, a cap on the vault's overall net exposure, a maximum notional value per trade, and a leverage ceiling for specific asset classes. These constraints are not part of the market-making algorithm's internal logic but a safety boundary that sits independently above the strategy. The design quality of the risk-control layer directly determines the vault's ability to survive at boundary conditions: the core lesson of the JELLY and POPCAT incidents is not that the market-making strategy itself was flawed but that the risk-control parameters failed to effectively constrain extreme position concentration in long-tail assets at the position-building stage (the full mechanism by which this imbalance amplifies the vault's tail risk is detailed in Section 21.4.3).

21.3.4 The distribution layer

The final link in liquidity engineering is the distribution layer, which returns the revenue generated by market-making activity to the capital layer's participants fairly and transparently. In Hyperliquid's economic model, the fee-distribution module is the incentive hub that sustains the entire system over the long run. The vault's revenue comes mainly from two sources: first, the bid-ask spread and the share of trading fees captured when the strategy engine facilitates trades on the order book; and second, the liquidation penalties the vault earns by taking over defaulted positions when large-scale forced liquidations occur in the market [10].

These two revenue sources display marked complementarity across different phases of the market cycle. In a low-volatility normal market, market-making spreads and fees

are the stable base of returns; in high-volatility or extreme conditions, although adverse selection risk increases market-making losses, a surge in liquidation penalties forms a powerful countercyclical hedge. The distribution module converts these net proceeds, in proportion to each investor's vault shares, into growth in the net asset value of those shares.

This mechanism builds a self-reinforcing incentive loop. Active trading generates fee revenue, which draws more retail capital into the vault; the vault's growing scale deepens the order book's liquidity, which in turn attracts more traders. Within this loop, the inflow of liquidation penalties enables the vault to replenish capital quickly after a crisis, strengthening its ability to take over the next round of liquidated positions [23]. By tightly binding liquidation management to market-making revenue, this design gives the protocol a stronger capacity for self-repair in the face of a systemic shock than a traditional CEX.

21.3.5 Overall operation

The preceding three subsections analyzed the internal logic of the capital, strategy, and distribution layers in turn; the core value of the three-layer architecture lies in their coordinated operation. Figure 21-3 places the system components of the three-layer architecture and their interactions in a single view, arranged top-down along the flow of funds, so that one can trace how capital moves and changes state across the layers.

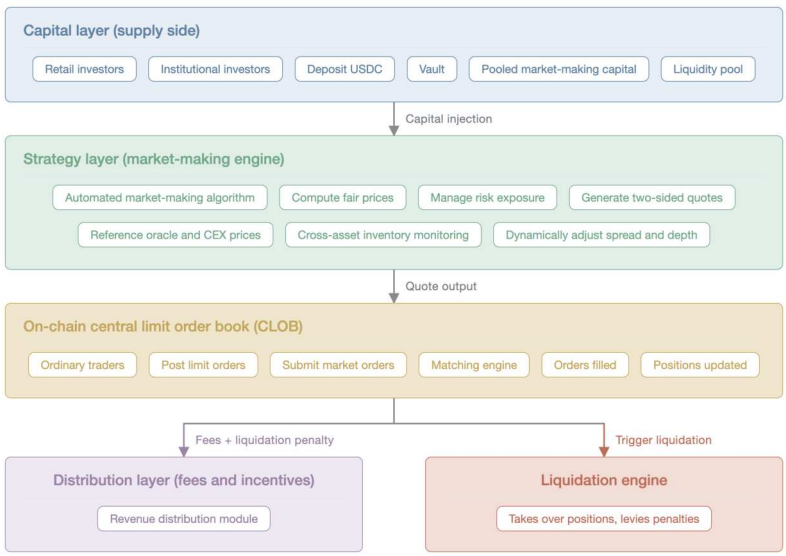


Figure 21-3. The three-layer architecture of Hyperliquid's liquidity engineering (Data source: Hyperliquid protocol documentation [24])

The process begins when user A deposits USDC into the protocol's HLP vault. Once this capital enters the capital layer, it merges with the funds of many other users to form a large liquidity pool, and user A receives HLP share tokens representing their proportional contribution. The strategy layer then takes over this aggregated capital.

According to preset algorithmic parameters, the market-making strategy engine assesses the current market's volatility and the balance between long and short forces and computes the optimal bid and ask prices and posted quantities. The engine converts these parameters into concrete instructions and posts buy and sell orders simultaneously on the on-chain CLOB.

When trader B submits a market buy order on the order book, the system matches it against a sell order posted by the strategy engine. Once the trade completes, trader B has paid a trading fee and the strategy engine has captured the bid-ask spread. At this point the flow of funds enters the distribution layer. The fee-distribution module aggregates the fee revenue and spread profit generated by this trade. If a liquidation of trader C occurs during this period, the liquidation engine intervenes and collects a liquidation penalty, and these funds likewise flow into the distribution module. Finally, following its rules, the distribution module adds the total proceeds proportionally to the vault's total assets; the net asset value of the HLP shares held by user A rises accordingly, completing the full loop from capital contribution to value capture.

Figure 21-4 presents the full end-to-end flow of this trade lifecycle, in which every link is executed automatically by smart contracts without human intervention.

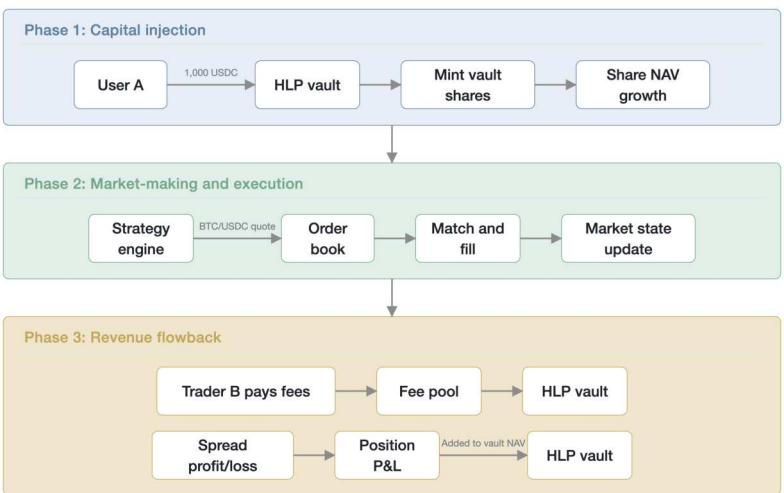


Figure 21-4. The end-to-end flow of funds: user deposit → order book quoting → revenue distribution (Data source: Hyperliquid protocol documentation [24])

21.3.6 Two routes

As the decentralized CLOB has developed, two sharply contrasting models of liquidity governance have emerged so far. The model exemplified by Hyperliquid's HLP vault may be called "protocol as market maker." Under this model, the protocol not only provides the trading infrastructure but also directly operates a built-in market-making engine. Its core advantages are control over liquidity and higher resilience during crises. Because the strategy engine is hard-coded, its execution logic is unaffected by market sentiment: under normal conditions it quotes continuously according to its rules and

does not withdraw orders on the basis of sentiment or profit motive, thereby providing the market with a rule-verifiable liquidity floor (under extreme boundary conditions this logic may still be overridden by governance intervention; see Section 21.4.6) [25].

The contrasting model, exemplified by dYdX v4, is "protocol recruits market makers." In this architecture, the protocol concentrates on building a high-performance on-chain matching engine and leaves liquidity provision entirely to external, independent market makers. To attract these professional firms, the protocol typically designs complex token-incentive programs and a tiered fee-rebate structure [26]. This model can deliver very high liquidity quality under normal conditions, because external market makers possess more advanced hardware and greater flexibility in adjusting their strategies.

The fundamental difference between the two lies in their distribution of risk and their governance structure. The "protocol as market maker" model socializes strategy risk (such as an algorithmic bug or a parameter error), spreading it across all vault participants, while lowering the entry threshold to a minimum and achieving full democratization. The "protocol recruits market makers" model, by contrast, isolates risk inside independent professional firms, but its liquidity provision still faces a principal-agent dilemma similar to that of a CEX: when the market collapses, token incentives are often insufficient to offset the market makers' inventory losses, and liquidity evaporates in an instant [27]. These two routes are not strictly opposed; future DEX architectures are more likely to converge—using a built-in vault to provide a base of antifragile liquidity while introducing external market makers through incentives to raise overall market efficiency.

21.4 Case study: the HLP vault

The HLP vault is the archetypal product of the liquidity-engineering paradigm, converting professional market-making activity into a standardized yield instrument that ordinary investors can access. This section examines it in depth along five dimensions: positioning, returns, risk, historical performance, and governance.

21.4.1 Index-style investing

The evolutionary history of financial products is a history of democratization—of turning institutional privilege into mass-market products. In traditional equity markets, index funds and exchange-traded funds (ETFs) aggregated retail capital and turned passive investing from a complex, institution-only operation into a standard product that ordinary investors could access at will [28]. This paradigm shift not only lowered the barrier to investment but also reshaped market microstructure, making passive capital an important cornerstone of market liquidity. In the context of decentralized derivatives exchanges, the HLP vault in the Hyperliquid protocol represents a similar structural innovation. The vault turns participation in perpetual futures market making—a highly specialized, capital-intensive activity—into a yield product that any user can access permissionlessly [29]. More precisely, this mechanism is a structured instrument for capturing a liquidity risk premium: it aligns with the democratizing logic of index funds along the dimension of lowering the barrier to market-making par-

ticipation, yet it differs fundamentally from index investing in its risk-return profile. Its essence lies in using protocol-level engineering to fully automate the fundraising, strategy execution, and revenue distribution of market-making activity.

In terms of mechanism design, the HLP vault is not a simple pool of funds but an autonomous agent capable of system-level economic behavior. Once users deposit dollar-pegged stablecoins into the vault, the protocol's built-in engine takes over this capital and executes a two-sided quoting strategy simultaneously across all perpetual futures pairs on the platform [30]. This feature of diversified market making across many contracts closely parallels the logic of a traditional index fund that buys a weighted basket of constituent stocks. Liquidity providers need neither expertise in quantitative trading nor low-latency infrastructure; in effect, they delegate the right to formulate strategy and manage risk to the protocol code [31]. In this process, a vault share becomes, in essence, a new asset class representing passive exposure to the profit pool of the entire exchange's market making.

Analogizing market-making activity to index investing, however, requires clarifying the fundamental difference in their risk-return profiles. The traditional index investor is a bearer of market risk whose returns derive from the long-run appreciation of asset prices; the vault's liquidity provider, by contrast, is a seller of liquidity insurance in market microstructure. A market maker's core profit comes from capturing the bid-ask spread and from funding rate arbitrage, and these returns compensate it for bearing adverse selection risk and inventory risk [32]. When users deposit funds into the vault, they are not buying a market trend but systematically selling liquidity. This role reversal means that the vault's net asset value does not depend directly on crypto assets' bull-bear cycle but is instead closely tied to the market's volatility characteristics, trading volume, and the degree of participation by competing market makers. Thus, although this structured capture of a liquidity risk premium democratizes the form of participation, it far exceeds traditional passive-investment instruments in the complexity of its risk assessment. From the perspective of behavioral finance, analogizing the vault to an *index fund* carries a framing-effect risk: to a retail investor, *index fund* connotes low risk, long-term holding, and passive diversification, whereas a vault depositor actually bears adverse selection risk, the risk of a nonlinear inventory blowup, and whale-manipulation risk. Its return distribution, moreover, exhibits a negative skew and fat tails quite unlike those of index investing.

21.4.2 Sources of return

The HLP vault's return structure is not a one-dimensional linear sum but is driven jointly by three independent engines with distinct cyclical characteristics. This multi-engine design is the core reason the vault can maintain a relatively stable net asset value across different market states.

The first return engine is market-making spread income, the most basic profit source of all market-making activity. In a normal market, the vault's strategy engine continuously posts buy and sell limit orders on the order book to earn the tiny difference between bid and ask prices [33]. Because the vault's capital is large and can provide liquidity across all trading pairs at once, these high-frequency, thin-margin trades accumulate

over a vast base to form the vault's principal source of returns in low-volatility periods. Spread income is highly sensitive to market conditions, however. When volatility spikes, a one-sided market causes the maker's posted orders to be swept through repeatedly and inventory to accumulate rapidly on the unfavorable side; spread income then not only falls sharply but may even turn negative [34].

The second return engine is the protocol fee share. As a protocol-native vault, the HLP vault enjoys a special right to a share of revenue, capturing a fixed proportion of traders' fees on the platform [35]. The distinguishing feature of this income stream is that it is decoupled from the actual performance of the market-making strategy and driven purely by the platform's total trading volume. In economic terms, it can be viewed as a *participation reward* the protocol pays the vault for providing base liquidity. When the market turns highly volatile, market-making spread income comes under pressure, but the surge in trading volume drives the fee share up markedly, thereby hedging the potential losses of the market-making strategy to some degree.

The third return engine is liquidation penalties and the management of defaulted positions. In a leveraged trading market, when a trader's margin ratio falls below the maintenance-margin requirement, the position is forcibly liquidated and charged an additional liquidation penalty [36]. This income stream requires distinguishing two components of quite different natures. In a normal market, the penalties generated by a large number of small, dispersed liquidations constitute risk-controlled fee income with low correlation to the vault's inventory risk. In extreme conditions, however, the large penalties triggered by a whale's liquidation or by cascading liquidations are entirely different in nature: as the vault collects these penalties, it is forced to take on an enormous directional position, so that liquidation-penalty income and inventory risk are amplified in the same instant. This *poisoned yield* means that the net effect of liquidation penalties as a return engine depends heavily on subsequent market movements: if the price reverts, the vault both collects the penalty and realizes an inventory gain; if the price continues to move away, the penalty income falls far short of offsetting the inventory loss. The JELLY incident is a textbook case of the latter.

Figure 21-5 depicts the relative contribution shares of these three return engines across three states: the normal market, the high-volatility market, and the extreme crisis. Liquidation penalties become the core support for net asset value during the extreme-crisis phase and can partially offset market-making losses under certain conditions; but this compensation is ex post, uncertain, and heavily dependent on the speed of the market's reversion, and it should therefore not be understood as a natural countercyclical hedge.

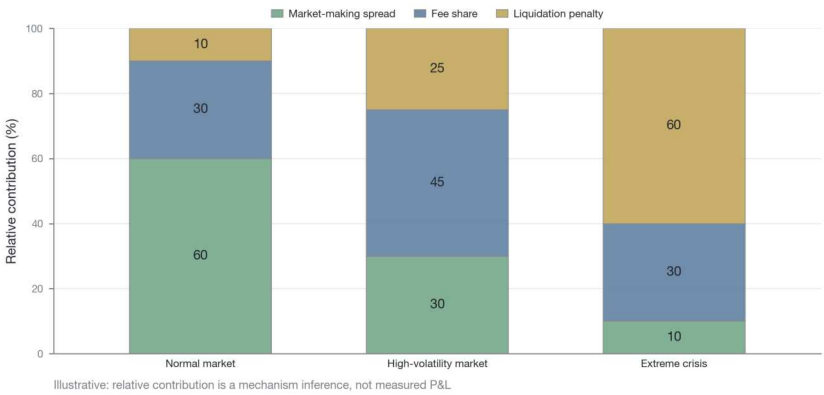


Figure 21-5. The cyclical characteristics of HLP's three return engines (Conceptual illustration: the relative contribution shares of the three engines are derived from mechanism analysis, not measured profit and loss; the mechanism follows Hyperliquid Protocol vaults [29], supplemented by the HLP security glossary entry [35]; on-chain profit-and-loss data for the vault can be queried through DefiLlama and Hyperliquid's official vault dashboard)

The combination of these three engines builds a subtle internal hedging mechanism. In a normal market, the market-making spread provides stable positive cash flow; in a high-volatility market, the rise in the fee share offsets the decline in the spread-capture rate; and in an extreme crisis, the explosive growth of liquidation penalties becomes the key buffer that absorbs inventory losses. This nonlinear feature of the return structure gives the vault's net-asset-value curve a resilience distinct from that of a single market-making strategy.

21.4.3 Risk exposure

Although the multi-engine return structure provides an intrinsic risk buffer, the essential nature of liquidity provision means the vault cannot be immune to systemic risks in market microstructure. Pooling capital and entrusting it to automated strategy execution in fact exposes liquidity providers to five distinct categories of risk.

The first is adverse selection risk. In the order book model, a market maker perpetually faces the predicament of trading as counterparty to traders who hold an informational advantage. When the market price makes a genuine move, the maker tends to sell too low or buy too high, and this structural loss caused by information asymmetry is termed the adverse selection cost [8]. Because the vault's strategy is driven by public code and its position is fully transparent on-chain, its quoting logic is readily predicted and gamed by high-frequency arbitrageurs and MEV searchers, so its adverse selection exposure is amplified (the full analysis of this transparency paradox is detailed in Section 21.4.6).

The second is the nonlinear blowup of inventory risk. The vault makes markets in dozens of perpetual futures pairs at once, and under normal conditions the independent fluctuations of different assets achieve a statistical diversification of inventory risk [37]. In cryptocurrency markets, however, cross-asset correlations tend to converge sharply toward one under extreme conditions. When the entire market suffers a macro shock, the vault may accumulate one-sided inventory in the same direction across all trading pairs at once. This synchronized deterioration of multidimensional inventory causes the

diversification effect to fail in an instant, pushing the vault into a state of severe capital depletion. If the one-sided market persists longer than the vault's strategy-adjustment cycle, the paper loss turns into an irreversible, substantive loss of capital.

The third category is extreme-event and whale-manipulation risk. Unlike a centralized-exchange market maker, which can actively sever its connection to the system at any time through human intervention, the protocol vault's automated nature means it cannot subjectively refuse a trade. This *always-on* property, while providing a liquidity floor, also makes the vault a potential target for malicious manipulation. If a trader with vast capital exploits the mechanical nature of the vault's quote depth to deliberately engineer an extreme price dislocation in a poorly liquid long-tail asset—forcing the vault to buy at the top or dump at the bottom—the vault's capital will be rapidly depleted [38]. Such attacks exploit the rigid weakness of the vault's strategy in responding to irrational trading behavior and constitute the largest source of the vault's tail risk.

The fourth category is mark-price and oracle-deviation risk. HLP's liquidation engine relies on the mark price to trigger forced liquidations, and the mark price is computed from an exponentially smoothed moving-average component of the spot index prices from several external exchanges together with the basis (the funding rate is a function of the basis between the mark price and the index price, not an input to the mark-price calculation; see Chapter 14). In a normal market, the mark price is highly consistent with the tradable price on the platform, but under extreme conditions the two can diverge significantly. The key attack step in the JELLY incident exploited precisely this mechanism: by manipulating the price of the JELLY token on an external spot exchange, the attacker drove the mark price on the Hyperliquid platform far from the actual tradable price, causing the vault to take on a position based on the distorted mark price and to incur an enormous unrealized loss. This risk is especially pronounced in low-liquidity assets, because the capital required to manipulate an external spot market is disproportionate to the asset's open interest on the perpetual futures platform.

The last is strategy and algorithm risk. The vault's operation depends heavily on a preset matrix of parameters and code logic. Parameter settings—such as quote width, depth allocation, and the inventory-skew coefficient—must adapt to an ever-changing market microstructure. If the market state shifts fundamentally while the strategy parameters are not updated in time through the governance process, the vault's quoting will deviate from the optimum and cause a continuous drain of funds [39]. At a deeper level, smart-contract security is the foundational risk of all on-chain financial infrastructure: any vulnerability in the vault contract, the strategy contract, or the oracle price-feed contract could allow funds to be extracted directly or cause the strategy engine to execute erroneous instructions. The largest fund-loss events in the crypto industry's history (such as cross-chain bridge attacks) all stemmed from security flaws at the contract layer rather than the strategy layer, and this risk is further amplified in the cross-chain liquidity layer discussed in Section 21.6.4.

These five risk categories are not evenly distributed across all market states. Figure 21-6 presents, in matrix form, how the severity of each risk category evolves across three market environments—normal, high-volatility, and extreme crisis; strategy and

contract risk runs through all market states and is difficult to hedge with a capital buffer.

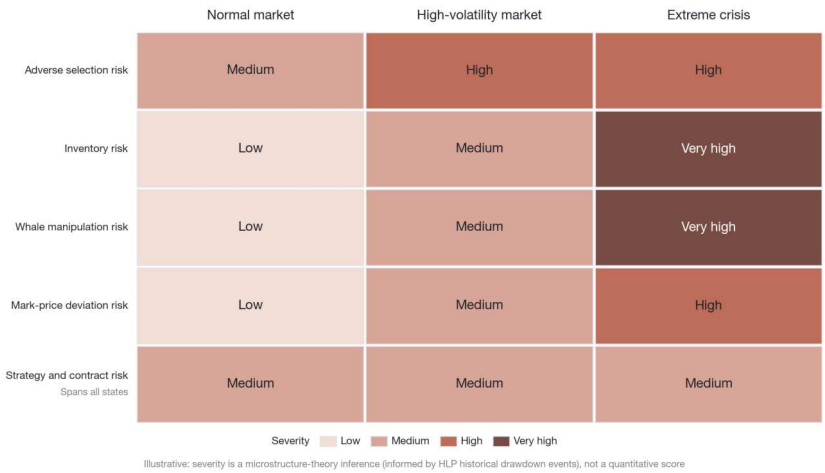


Figure 21-6. The HLP five-category risk-exposure matrix (Conceptual illustration: the severity of the five risk categories across three market states is derived from microstructure theory, not a quantitative score; theoretical basis: Tiniç et al. 2023 [8] and Fournier & Jacobs 2020 [37]; severity references the HLP vault's historical drawdown events, as of June 2026: the March 2025 ETH whale-liquidation event with a loss of approximately $4 million, the March 2025 JELLY token attack with a maximum unrealized loss of approximately $13.5 million [1][57], and the November 2025 POPCAT manipulation event with approximately $4.9 million in bad debt [38][58])

21.4.4 Historical performance

Empirical analysis of the vault's historical performance is important for testing the effectiveness of its mechanism. Since the vault went live, its net-asset-value curve has recorded not only the profitability of the market-making strategy across different market cycles but also the fragility boundary of the protocol architecture in the face of extreme shocks.

As of October 2025, the HLP vault's cumulative profit and loss was approximately $121.8 million, corresponding to a total return of about 50% (over the period) and a full-cycle annualized compound growth rate of about 42% (data from the DefiLlama hyperliquid-hlp dashboard, on an all-time cumulative profit-and-loss basis, accessed June 2026; the total return uses the vault's average assets under management over the period as the denominator, and the 42% annualized figure is significantly inflated by an early low-base effect, as qualified below). On this basis, only 6 of the 64 observation weeks recorded a negative return, with an average weekly return of about 0.84%. This performance fits the classic profile of a high-frequency market-making strategy: accumulating returns through thin-margin trades with a very high win rate. The smoothness of the net-asset-value curve is broken at certain points, however, by sharp downward gaps. On March 12, 2025, a whale deposited $4.3 million in margin to build a 50x-leveraged ETH long position (notional value approximately $200 million to $230 million) and then deliberately withdrew the margin to trigger liquidation, forcing the HLP vault to absorb the closing liquidity of approximately 113,000 ETH at

an extremely unfavorable price, causing a direct loss of approximately $4 million [36]. Just two weeks later, on March 26, in the JELLY token attack, the attacker exploited the thin liquidity of a long-tail asset and manipulated the spot price across platforms, forcing the vault's short position to a maximum unrealized loss of approximately $13.5 million [1][57]. The POPCAT manipulation event of November 2025 further exposed the vault's fragility in the face of market manipulation: through a fake buy wall and flash cancellations, the attacker caused approximately $4.9 million in bad debt [38][58]. Although liquidation penalties and fee income partially offset these losses after the fact, the deep drawdowns in net asset value faithfully reflect the systemic cost the vault bears as a lender of last resort. In favorable liquidation events, the vault can likewise earn significant gains; for example, the forced liquidation of a $700 million BTC long position on February 1, 2026, brought the vault a single-day profit of approximately $15 million (data for this event are from KuCoin [38] and Longbridge [59] reports of February 2026 and the Hyperliquid vault dashboard, accessed June 2026). This profit event deserves symmetric attention within a risk-analysis framework, however: it shows that HLP had taken on an enormous directional position at that moment, and had the BTC price moved by an equal magnitude in the opposite direction, the vault could have suffered a loss of comparable size. This event reflects the level of risk concentration permitted by the vault's risk-control parameters more than any superiority at the strategy level.

Several important qualifications are needed in interpreting the performance data above. The 42% annualized compound growth rate is significantly inflated by a low-base effect: the vault's early assets under management were far smaller than in later periods, so the high early return rates contributed disproportionately to the full-cycle compound growth rate. As assets under management grew from approximately $150 million at the end of 2024 to a peak of approximately $512 million in May 2025 (the size curve is a single-source estimate, with specific figures pending verification against DefiLlama and official dashboard historical data), the expected return per unit of capital was noticeably diluted: the full-year return for 2024 was about 50%, whereas the per-unit-capital annualized return in 2025 fell back markedly (estimates across sources and measurement bases range from roughly 10% to 22%, the differences arising mainly from whether fees are deducted, whether token incentives are counted, and differing statistical windows). Moreover, the observation period covered a strong phase of the crypto market, and market-making strategies generally perform well in an environment of high trading volume and moderate volatility. The presence of liquidation-penalty income makes the vault functionally long market volatility, and the roughly $15 million single-day profit event noted above shows that the vault bears a significant directional risk exposure—which is in tension with the characterization in Section 21.4.1 of the vault as "not depending directly on crypto assets' bull-bear cycle."

From the perspective of attack-surface analysis, although the JELLY and POPCAT incidents both resulted in financial losses for the vault, their attack vectors differ in essence. JELLY was a liquidity-exhaustion attack: the attacker built a large short position and deliberately triggered a self-liquidation, forcing the vault to take on the entire position as counterparty, and then amplified the vault's unrealized loss through cross-platform spot manipulation. After the JELLY incident, the protocol implemented

risk-control measures such as lowering the maximum leverage and adding position-opening limits. The POPCAT incident eight months later, however, used a different attack vector—creating a price illusion with a fake buy wall and then flash-canceling—which was market manipulation rather than liquidity exhaustion, a category the JELLY-era fixes did not cover. The systemic weakness both incidents exposed is that HLP's quote depth in long-tail assets is severely disproportionate to the scale of externally manipulable capital.

Figure 21-7 marks the key market events over the course of the vault's net-asset-value evolution, revealing how external shocks map into profit-and-loss fluctuations in the internal capital pool.

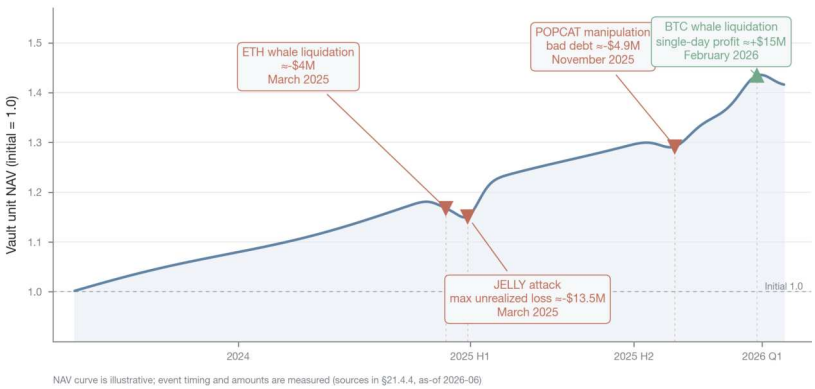


Figure 21-7. The HLP net-asset-value curve (illustrative trajectory) with key events annotated (The net-asset-value curve is an illustrative trajectory with an initial value of 1.0; the shape of the curve references Hyperliquid's official vault dashboard and DefiLlama [29] and is not a day-by-day measured series; the timing, direction, and amount of the four overlaid key events are measured, as of June 2026: the March 2025 ETH whale liquidation with a loss of approximately $4 million [36], the March 2025 JELLY attack with a maximum unrealized loss of approximately $13.5 million [1][57], the November 2025 POPCAT manipulation with approximately $4.9 million in bad debt [38][58], and the February 2026 whale liquidation with a single-day profit of approximately $15 million; event sources are given in Section 21.4.4 of the text [59])

Further return-attribution analysis (based on inference from the protocol mechanism, lacking direct verification from on-chain itemized profit-and-loss data) suggests that the dynamic switching of profit composition across market states may be the key to the vault's sustained long-run profitability. In a normal market of moderate volatility, the mechanism implies that bid-ask spread capture contributes the bulk of the positive returns, reflecting that the vault's pricing algorithm should be reasonably competitive at handling ordinary order flow. When the market enters a high-volatility quadrant, however, the return structure reverses markedly. The contribution of market-making spreads shrinks sharply, while the fee share driven by surging trading volume and the liquidation penalties triggered by large-scale liquidations become the core pillars supporting net asset value.

Figure 21-8 provides a quantitative breakdown of the vault's return attribution under two typical market environments.

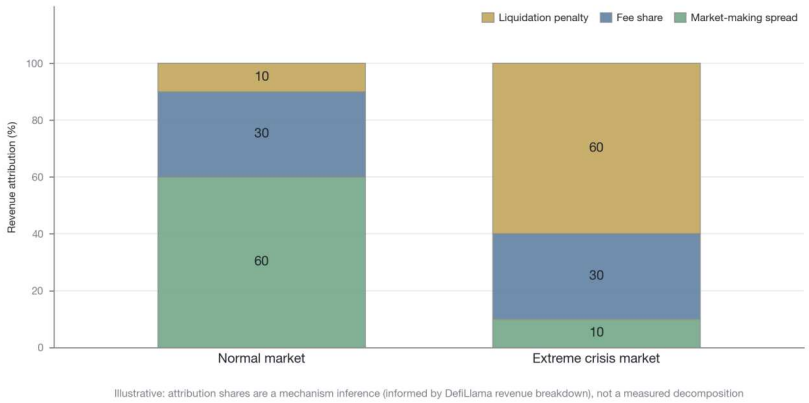


Figure 21-8. HLP return-attribution breakdown (Conceptual illustration: the return-attribution shares for the two scenarios are derived from mechanism analysis, lack direct verification from on-chain itemized profit-and-loss data, and are not a measured decomposition; the primary anchor for the attribution is the Hyperliquid vault dashboard and DefiLlama revenue line items [29])

This dynamic return-compensation mechanism explains why the vault can quickly repair its net asset value after a deep drawdown. It shows that the protocol-level distribution-rule design—internalizing fees and liquidation penalties as vault revenue—is in essence a systematic cross-subsidy for liquidity providers who bear tail risk. Strip away this non-market-making income, and the long-run viability of an automated strategy that relies on spread capture alone in the on-chain environment would be greatly diminished.

21.4.5 Comparison with traditional products

Comparing the liquidity vault of a decentralized protocol with innovative products in traditional financial markets helps locate its coordinates within the broader financial spectrum. In traditional finance, although ETFs greatly popularized passive investing, direct participation in market-making activity was always regarded as an institution-only domain. More recently, some structured products have used derivatives packaging to offer retail investors exposure resembling market-maker returns (such as short-volatility strategies or covered-call funds), and such products may be loosely termed *market-making ETFs* [40].

The HLP vault differs markedly from such traditional products in its underlying logic. Traditional structured products typically simulate market-maker returns by executing complex option strategies in over-the-counter markets or on derivatives exchanges, and their operation is entirely a black box to investors, who see only the final change in net asset value and cannot see through to the underlying position distribution and execution details. The protocol vault, by contrast, achieves genuine on-chain transparency. Every order the vault posts, every fill, every inflow and outflow of funds, and its real-time inventory state are publicly queryable on a block explorer [31]. This extremely transparent information structure significantly reduces the information asymmetry between manager and investor found in traditional finance (though, as Section 21.4.6 notes, transparency itself does not eliminate information asymmetry entirely but shifts

it into a strategy-level disadvantage in predictability).

In entry threshold and intermediation cost, the protocol vault displays the efficiency advantage of a decentralized architecture. Traditional products, constrained by strict regulatory frameworks, often require very high minimum subscription amounts and charge steep management fees and performance fees. The protocol vault, through smart contracts, achieves zero-threshold fundraising and zero management fees [29]. After deducting network friction costs, returns are distributed to liquidity providers strictly in proportion. This disintermediated profit-distribution model raises the efficiency of capturing market-making returns to a level traditional finance can scarcely reach.

Innovation comes with limitations that cannot be ignored, however. Traditional structured products, though costly and opaque, are backed by strict regulatory endorsement and well-developed investor-protection mechanisms; when a product faces extreme risk, a clear path of legal recourse exists. The protocol vault, by contrast, operates in an unregulated, crypto-native environment, where the principle that code is law means liquidity providers must bear all technical risks and economic losses themselves. Moreover, traditional products typically provide standardized risk metrics (such as a value-at-risk (VaR) calculation), whereas the protocol vault currently lacks a unified risk-assessment framework, making it difficult for ordinary investors to gauge their true level of risk exposure accurately.

21.4.6 Governance dilemmas

As a piece of code-driven, community-funded financial infrastructure, the protocol vault faces a series of complex trade-offs at the governance level. These dilemmas arise from a fundamental contradiction between the vault's public-good character and the strategic secrecy it needs to survive in a competitive market.

The foremost conflict is the transparency paradox. From the liquidity provider's standpoint, extreme transparency is the foundation of trust; they need to know exactly how their funds are being used. In a zero-sum trading market, however, strategic transparency often amounts to an exploitable weakness. When the vault's quoting logic, inventory thresholds, and rebalancing frequency are public to the entire network, professional quantitative teams can readily construct targeted counter-strategies—front-running the vault before it adjusts its position, or imposing a liquidity squeeze when the vault is forced to close out [8]. This adverse selection cost created by transparency is ultimately borne by the liquidity providers. The protocol must therefore find a fragile balance between "reassuring investors" and "preventing arbitrage by counterparties."

Another core dilemma is the locus of control over parameters. Although the vault's strategy engine runs automatically, its core parameters (such as the spread multiplier and inventory tolerance) still require human setting. Handing control over parameters entirely to the core team would in effect degrade the vault into a centrally managed black-box fund, betraying the original intent of decentralization. But entrusting parameter adjustments to a decentralized autonomous organization (DAO) through token voting would leave the vault unable to respond in time to a rapidly shifting market microstructure, given the lengthy governance process [39]. Under extreme conditions, a delay of a few minutes in adjusting a parameter can mean losses of millions of dollars.

This tension between "decentralizing control" and "maximizing execution efficiency" is a governance challenge shared by all on-chain automated market-making protocols today.

The dilemma over parameter control is especially acute in extreme events. The emergency handling of the JELLY incident in March 2025 revealed a phenomenon of "crisis governance reversion": when the vault faced an unrealized loss of approximately $13.5 million that threatened the safety of the entire vault's roughly $230 million in assets (the vault's assets under management at the end of March 2025, when the JELLY incident occurred), Hyperliquid's validators voted in an emergency to forcibly settle all JELLY positions at $0.0095 (far below the then-prevailing market price of about $0.50) and delisted the token from the platform. This intervention is functionally equivalent to the manual risk-control intervention of a CEX and stands in inherent tension with the engineering positioning of delegation "to code rather than to a person" in Section 21.3.1. From the perspective of institutional analysis, DEX liquidity engineering today in fact operates in a hybrid mode: governed by code rules under normal conditions, reverting to human discretion in a crisis. This arrangement has its counterpart in traditional finance—for example, a central bank follows policy frameworks such as the Taylor rule under normal conditions but exercises emergency discretion in a crisis. But the analogy also exposes a legal uncertainty: if the protocol claims to be decentralized, a validator vote that unilaterally alters the terms of a trade may constitute breach under contract law; if it concedes that a centralized decision-making entity exists, that entity may face legal charges of market manipulation. The boundary of legitimacy for crisis governance reversion is an unresolved institutional gap in DEX governance architecture today.

The vault mechanism also faces an existential threat distinct from market-microstructure risk: legal-characterization risk. The HLP vault aggregates retail capital, manages it uniformly through an algorithmic strategy, and distributes returns in proportion to shares—a structure that may satisfy all four prongs of the Howey test under U.S. securities law: an investment of money (depositors inject USDC), a common enterprise (funds are pooled into a single pool), an expectation of profit (depositors expect to earn returns through market-making activity), and reliance on the efforts of others (returns depend on the protocol code and the core team's strategy settings). In most jurisdictions, such a structure could be deemed a collective investment scheme or a fund product. If vault shares were classified as securities, the protocol would face the compliance requirement of registering as an investment company or fund manager, which would fundamentally alter the viability of the DEX liquidity-engineering model, because the permissionless participation, zero management fees, and instant global access described in Section 21.3 would all be difficult to sustain under the current securities-regulatory framework.

In addition, an endogenous conflict exists between liquidity providers' redemption rights and the vault's system resilience. To ensure the free entry and exit of funds, a vault typically allows users to redeem their capital at any time after a very short lock-up period [29]. When the market suffers an extreme crisis and the vault faces a severe drawdown, however, investors' flight to safety often triggers a large-scale wave of redemptions. This procyclical withdrawal of capital not only weakens the vault's ability

to take over liquidated positions but also forces the strategy engine to close out and realize positions when liquidity is at its worst, entering a negative feedback loop of "NAV decline → capital redemption → forced liquidation → further NAV decline." Unlimited redemption rights thus in fact constitute the vault's principal source of systemic risk in extreme states. How to design a reasonable exit mechanism (such as a dynamic redemption fee or crisis-period withdrawal limits) that both protects investors' interests and preserves the vault's capital-buffer capacity is a pressing question for liquidity governance.

21.4.7 Division of labor among vaults

As decentralized derivatives exchanges expand in scale, an architecture in which a single vault bears all liquidity responsibilities gradually exposes a problem of internal cross-contamination of risk. A market-making strategy pursues high-frequency, low-volatility, stable returns, whereas liquidation takeover involves low-frequency, high-volatility, tail-risk exposure. Mixing these two lines of business, with their sharply different risk-return profiles, in the same capital pool leaves liquidity providers unable to allocate precisely according to their own risk preferences. A natural conception of architectural evolution to address this is to introduce specialized division of labor, stripping the liquidation function out of the market-making vault to form an independent liquidation liquidity provider (LLP) vault [41]. This division of labor, however, is a matter of mechanism design and evolutionary conception rather than current deployment: as of June 2026, Hyperliquid's official documentation shows that its liquidator vault remains a constituent strategy of HLP rather than an independent, firewall-isolated vault standing alongside HLP [29]; the analysis below should accordingly be understood as a conceptual exploration of specializing the liquidation function.

This division-of-labor architecture achieves an economic isolation and repricing of risk. HLP concentrates on two-sided order book quoting in normal markets, earning returns by capturing the bid-ask spread and a share of trading fees; its risk exposure is concentrated mainly in ordinary adverse selection and short-term inventory fluctuations, suiting risk-neutral investors who seek relatively stable cash flow. The LLP vault, by contrast, serves specifically as the protocol's lender of last resort, taking over defaulted positions when traders are forcibly liquidated and compensating itself for the extreme market risk it bears by collecting liquidation penalties. This design attracts investors with a higher risk tolerance who are willing to endure long periods of idle capital in exchange for capturing the crisis premium.

Figure 21-9 illustrates the division of labor in this conception between the HLP market-making vault and the LLP vault: HLP handles two-sided order book quoting in normal markets, with income drawn mainly from market-making spreads and fees; the LLP vault focuses on taking over defaulted positions at moments of crisis, earning returns through liquidation penalties; and a risk-isolation firewall between the two ensures that the shock of the liquidation flow is not transmitted to the inventory balance of the market-making strategy.

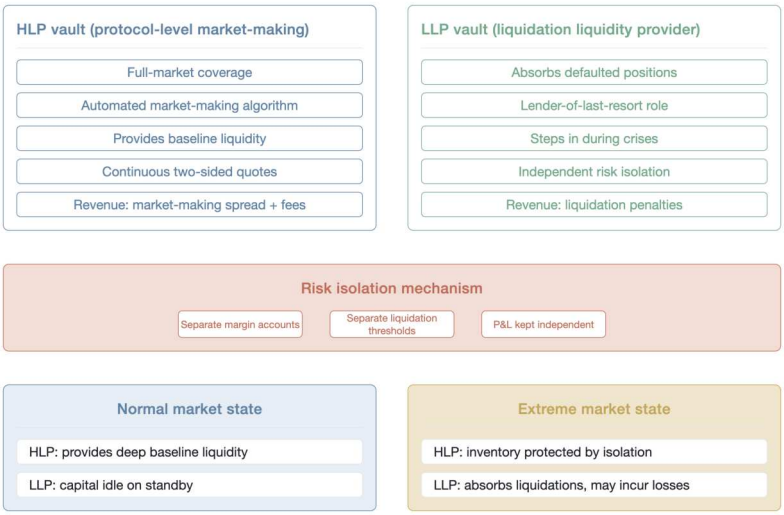


Figure 21-9. The division-of-labor architecture of HLP and LLP (Schematic: the division of labor and risk isolation between HLP and an independent liquidation vault is a matter of mechanism design and evolutionary conception; as of June 2026, Hyperliquid's liquidator vault is a constituent strategy of HLP rather than an independent vault [29]; theoretical basis: Chitra 2025, Autodeleveraging [41])

Specialized division of labor not only meets differentiated demand on the capital side but also produces a marked synergy at the level of strategy execution. Under the single-vault model, when cascading liquidations occur in the market, the enormous flood of defaulted positions instantly breaks the inventory balance of the market-making strategy, forcing the engine to stop quoting or even to take orders in the opposite direction to close out—draining order book depth at precisely the moment the market needs liquidity most. Once an independent liquidation vault is introduced, defaulted positions are isolated and taken over, and the market-making vault's inventory state is spared the direct shock of the liquidation flow. This isolation mechanism lets the market-making strategy keep running normally for longer under extreme conditions, thereby raising the systemic resilience of the entire exchange's liquidity supply. The risk-isolation effect of the division-of-labor architecture may fail in an extreme tail event, however. When the whole market suffers a one-sided move that triggers cascading liquidations, the liquidation vault is forced to take on a large volume of same-direction positions at the same time that the HLP market-making vault is accumulating inventory in that same direction through normal market making; the profit and loss of the two vaults then become highly positively correlated under extreme conditions, and the diversification effect disappears exactly when it is most needed. This resonance risk within the system is directly related to the countercyclical nature of liquidity discussed in Chapter 20, and mitigating it requires introducing directional constraints at the vault-design level (such as a hedging requirement for the liquidation vault) or a time-delay mechanism (such as delaying the takeover of liquidated positions to avoid a short-term price shock).

Through capital aggregation, automated strategy execution, and specialized risk isolation, the protocol vault mechanism accomplishes a deep deconstruction and reshaping

of the traditional market-making model. It demonstrates that in a decentralized network, complex financial-intermediation services can be transformed into transparent protocol rules. Yet engineering liquidity provision does not eliminate risk; it changes the form of its distribution. In the market's future evolution, this new kind of delegation relationship—one that takes code as the agent—must still withstand the stress test of more extreme cycles.

21.5 A survival simulation for the independent market maker

The core goal of liquidity engineering is to convert dispersed capital into continuous two-sided quotes through protocol rules. Within this architecture, the vault represents a passive route to participation, allowing ordinary users to delegate their funds to the protocol's strategy engine. The on-chain central limit order book, however, also allows independent market makers to connect directly through the API and run their own strategies. These two modes of liquidity provision are complementary in microstructure terms. To assess the effectiveness of the protocol's incentives in attracting active market makers, this section simulates the operation of a mid-sized independent market maker on the Hyperliquid platform, quantifies its risk-return structure, and explores its boundary of profitability under different market scenarios.

21.5.1 Simulation setup

Assessing a market maker's ability to survive requires a parameter system consistent with the features of the on-chain environment. The parameter settings below are calibrated against empirical ranges in the public literature: the volatility assumption references the empirical range of roughly 60%–75% for the annualized realized volatility of BTC perpetual futures [42]; the spread and adverse selection ratios reference the BTC perpetual futures order book microstructure data reported by He et al. (2022) [43] (an average bid-ask spread below 0.05 basis points and a price impact below 0.3 basis points for a $500,000 order; the basis-point convention here follows the original paper's definition of relative spread, and readers should note the conversion between basis points and percent when citing) as well as the adverse selection decomposition of the effective spread in Tiniç et al. (2023) [8]; and the funding rate distribution references the BIS working paper by Schmeling et al. (2023) [44]. The simulation is intended to provide an illustrative analysis rather than a precise quantitative forecast, and the specific figures vary significantly across market cycles and platforms. This simulation constructs the profile of a typical mid-sized independent market maker with an initial capital of $750,000, focused mainly on perpetual futures pairs of mid-cap crypto assets (such as ETH or SOL). The market maker employs a classic inventory-management strategy, and its infrastructure has an order-execution latency of between 200 and 500 milliseconds. This latency is upper-middle for the on-chain environment; although it cannot rival the microsecond-level colocated execution of a traditional centralized exchange, it is sufficient to maintain competitive quotes on most decentralized limit order books.

The strategy parameters directly determine the market maker's behavior on the order book. The base two-sided spread is set at ± 3 basis points, a width that provides effective liquidity while preserving a basic profit margin for the maker. On order depth, the maker maintains \$50,000 to \$100,000 of posted size on each side of the book. Inventory management is the core of the market-making strategy: when the absolute value of one-sided inventory exceeds \$200,000, the strategy engine triggers a quote-skewing mechanism that adjusts the asymmetry of the bid-ask spread to attract offsetting orders and push inventory back toward neutral. In addition, to guard against extreme tail risk, when short-term market volatility exceeds three standard deviations, the strategy automatically suspends two-sided quoting and enters a risk-averse state.

The market-environment assumptions abstract real historical data into three typical scenarios. The normal market occupies about 65% of trading time and is characterized by steady volatility and randomly distributed order flow. The high-volatility market accounts for about 25% and usually accompanies macroeconomic data releases or major industry events; it is characterized by a pronounced one-sided price trend and a markedly narrowed effective spread. The extreme-crisis scenario, though only 10% of the time, typically manifests as liquidity exhaustion and a liquidation cascade and is the critical period that tests a market maker's ability to survive. These three scenarios form the base environment for the revenue and cost calculations that follow. The 10% probability assigned to the extreme scenario is based on the empirical frequency of the 2024–2025 crypto perpetual futures market, in which roughly 10% of trading weeks saw an extreme event with intraday volatility above 5%. Sensitivity analysis shows that as the frequency of tail events rises, the market maker's expected annualized return converges rapidly toward breakeven (the specific critical value is illustrative only and depends on the parameter settings), which reflects the decisive influence of tail-risk pricing on any judgment of profitability.

21.5.2 The revenue side

An independent market maker's revenue has three main components, which differ in stability and predictability. The first source is the liquidity-provider rebate offered by the protocol. To attract market makers, many decentralized exchanges adopt a maker rebate—awarding a fee bonus for orders that provide liquidity and are ultimately filled [5]. This income is relatively stable, and its size depends directly on the maker's average daily volume and the rebate rate the protocol sets. In the simulation, this income is treated as the maker's unconditional base return and forms the first line of defense in covering fixed operating costs.

The second source is bid-ask spread capture, the most central way a traditional market maker profits. By quoting both a bid and an ask, the maker earns the difference on two-sided fills. The theoretical spread does not equal the return actually captured, however. The capture rates and loss rates below are all calibrated values for this simulation, not empirical observations. In a normal market, the relatively high randomness of order flow typically keeps the maker's effective spread-capture rate between 40% and 60%. This rate measures the share of orders filled on both sides (the gross capture rate); after deducting adverse selection costs and operating friction, the net profit margin

is about 0–30%. When the market turns highly volatile, the strengthening one-sided trend often leaves the maker able to fill only passively on the unfavorable side, and the effective capture rate falls sharply to 20%–30%. In an extreme crisis, with the market in a one-directional wave of large-scale selling or buying, the maker can barely achieve two-sided matching, and the spread-capture rate approaches zero.

The third profit-and-loss component is net funding rate income. The funding rate is the core mechanism used to anchor the spot price in perpetual futures markets, with longs and shorts periodically exchanging payments according to the price deviation [43]. For a market maker, this income is highly random. Strictly speaking, the funding rate should be classified as a conditional profit-and-loss item tied to inventory direction rather than a pure revenue item: a neutral market maker's net funding rate income is close to zero when its long and short positions are roughly balanced, and it becomes a stable source of income only when the maker holds a systematic directional tilt—at which point the maker also bears directional risk [44]. Its sign depends on the product of the maker's current inventory direction and the direction of the funding rate. If the maker holds inventory opposite to the market's dominant trend, it usually receives a positive funding rate payment; otherwise it must pay. Research shows that in certain market environments, funding rate arbitrage can itself constitute an independent profit strategy [45]. In a pure market-making strategy, however, funding rate income tends to be one of the more volatile components of the profit-and-loss statement and cannot serve as a stable profit expectation.

21.5.3 The cost side

While pursuing the revenue above, the market maker must bear costs and risks along several dimensions. The adverse selection loss is the maker's principal implicit cost. When a trader with an informational advantage enters the market, the maker often fills against it at an unfavorable price, after which the market price moves adversely and the maker incurs a paper loss [46]. In the on-chain environment, because all trade data and smart-contract states are publicly visible, the maker's quoting intent is more easily exploited by high-frequency arbitrageurs or MEV searchers. The adverse selection loss rate for on-chain market making is therefore typically higher than at a traditional centralized exchange. In a normal market, this loss accounts for about 30%–40% of the theoretical spread; in a high-volatility scenario, about 60%–80%; and in an extreme scenario, it may exceed 100%, meaning the maker loses substantively on every fill.

Inventory-fluctuation loss is the second core cost. In the course of providing two-sided quotes, a market maker inevitably accumulates one-sided positions. Until inventory reverts to neutral, these positions are fully exposed to market price fluctuations [47]. The simulation shows that inventory-fluctuation loss is significantly positively correlated with the volatility of the underlying asset and with the maker's average inventory-holding time. In a market with a strong one-sided trend, the maker tends to accumulate a large volume of disadvantaged positions, and the effectiveness of the inventory-management mechanism is severely tested. If the price deviates faster than the quote-skewing strategy can attract offsetting orders, the maker faces enormous unrealized losses.

In addition, the market maker must bear the friction costs and institutional costs of on-chain execution. Although some appchains or layer-2 networks significantly lower the per-trade fee, the cumulative execution cost remains a non-negligible expense for a maker that must update its quotes at high frequency. Institutional costs include the infrastructure investment needed to sustain around-the-clock operation and the special risk premium that may arise in a decentralized protocol, such as an unexpected loss from an anomalous triggering of the liquidation engine. These fixed and semi-fixed costs further raise the maker's profitability threshold.

21.5.4 Scenario analysis

Substituting the revenue and cost elements above into the three specified market scenarios yields the results in Figure 21-10: in the normal market the net-asset-value curve rises smoothly; in the high-volatility scenario, profit and loss alternate around the zero axis with markedly greater variance; and in the extreme-crisis scenario the net asset value falls sharply—vividly displaying the asymmetry of tail risk.

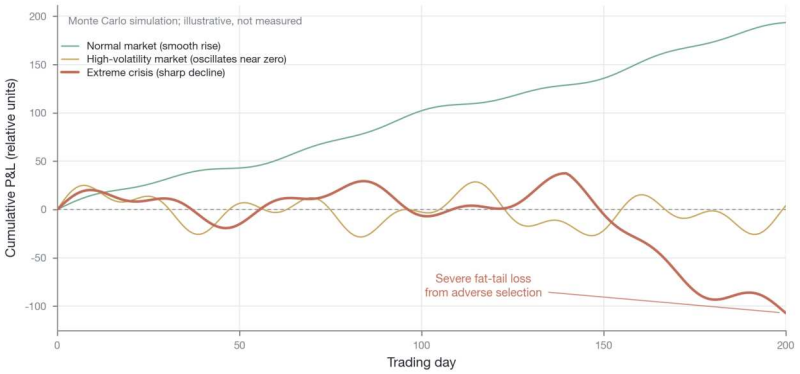


Figure 21-10. Profit-and-loss simulation for an independent market maker: three scenarios (Data source: Monte Carlo simulation; the volatility parameters are calibrated against the empirical range of 60%–75% for the annualized realized volatility of BTC perpetual futures [42], the adverse selection ratio against the effective-spread decomposition reported by Tiniç et al. 2023 [8], and the funding rate distribution against the BIS working paper [44])

In a normal market, the maker can serve effectively as a liquidity intermediary. Here the randomness of order flow keeps two-sided fills relatively balanced, and spread capture is sufficient to cover adverse selection losses and execution costs. The simulation shows that in this scenario the maker's combined average daily return reaches 0.05%–0.15% of its capital, and the net-asset-value curve trends steadily upward. This indicates that in a calm market environment, the protocol's incentives and the maker's profit motive achieve good incentive compatibility.

When the market enters the high-volatility scenario, the maker's profit structure deteriorates markedly. The formation of a one-sided trend quickly unbalances inventory, and the adverse selection loss rate climbs to 60%–80% of the theoretical spread. Although rising trading volume may bring more rebate income, this increment is often insufficient to offset the paper losses caused by inventory fluctuations. At this stage,

the maker's average daily return oscillates violently around the breakeven line, and the variance of the net-asset-value curve increases significantly. The maker's core task shifts from pursuing profit to controlling risk, and the strategy centers on how to clear disadvantaged inventory quickly through aggressive quote-skewing.

In the extreme-crisis scenario, market microstructure fractures. Large-scale one-directional orders flood the order book, the maker's effective spread-capture rate drops to nil, and the adverse selection loss far exceeds 100%. More seriously, in an environment of liquidity exhaustion the inventory-management mechanism fails completely, and the maker cannot find a counterparty willing to take the other side. The simulation shows that in this scenario the maker faces a significant single-day loss and the net-asset-value curve drops sharply. This asymmetric tail risk is precisely the fundamental reason many independent market makers choose to cancel their quotes and exit the market in the early stage of a crisis.

21.5.5 The profitability boundary

A market maker's ability to survive is highly sensitive to several microstructure parameters. As shown in Figure 21-11, a sensitivity analysis of four parameters in turn—capital size, base spread, market volatility, and competitive intensity—can delineate the profitability boundary of an independent market maker in the on-chain environment, which also indirectly reflects the lower bound of the protocol's incentive effectiveness.

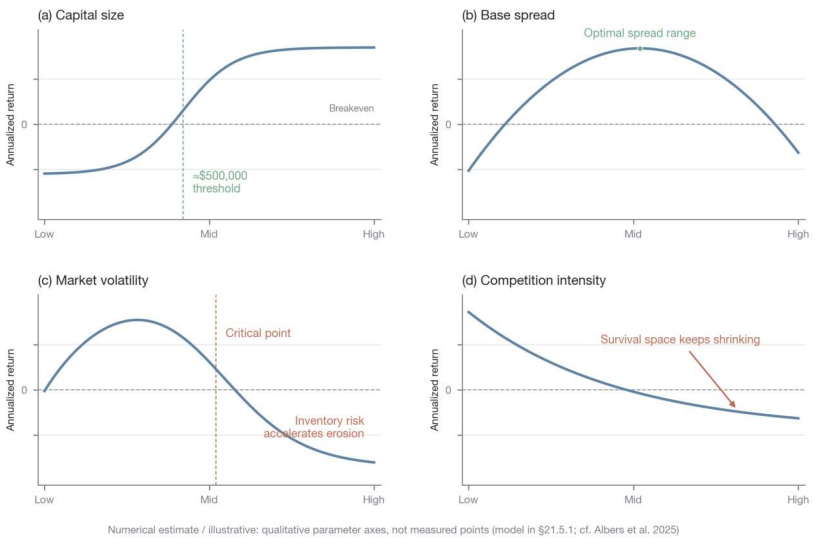


Figure 21-11. Sensitivity analysis of the profitability boundary (Data source: numerical calculation; model parameters are given in Section 21.5.1; the trade-off between fill probability and adverse selection references the empirical study of Binance BTC perpetual futures by Albers et al. 2025 [48])

Capital size is the foundation of a market maker's risk-bearing capacity. The analysis shows a nonlinear positive relationship between capital size and the annualized return. A smaller capital base means the maker more easily hits its risk limits when inventory becomes unbalanced and is thus forced to close out at unfavorable prices. Only when capital size exceeds a certain threshold (about $500,000 in the simulation) (this refers

to total capital, including margin and operating reserves; at an initial margin rate of 5%–10%, it can support a notional market-making position of roughly $5 million to $10 million, and the simulation covers only a single mid-cap trading pair) can the maker cover its fixed costs while retaining enough of a cushion to absorb the short-term drawdowns of a high-volatility market.

The base spread directly determines the maker's gross margin. Given an adverse selection probability, too narrow a spread makes the expected return on each trade negative and traps the maker in structural losses. Yet the spread is also tightly constrained by market competition. Rising competitive intensity forces makers to narrow their quotes to compete for fills, compressing profit margins across the whole industry [49]. As the number of active makers grows, each maker's share of order flow shrinks and the profitability boundary shifts upward, requiring superior execution efficiency or lower operating costs for a maker to survive.

Market volatility is the most active exogenous variable affecting the profitability boundary. Moderate volatility helps increase trading volume and opportunities for spread capture, but once volatility exceeds the strategy model's tolerance ceiling, inventory risk is amplified exponentially. The sensitivity analysis clearly displays the trajectory of returns decaying rapidly as volatility rises. This means the protocol's incentives must be able to provide extra risk compensation during high-volatility periods; otherwise a rational independent market maker will inevitably choose to exit, causing liquidity provision to fracture exactly when it is most needed.

21.5.6 Active versus passive

Having assessed the independent market maker's survival, the analysis now turns to the route by which an ordinary user deposits funds directly into the protocol vault (passive participation). These two modes of liquidity provision differ fundamentally in source of return, risk control, and barrier to participation. As shown in Table 21-3, the independent market maker has full autonomy over its source of return and risk control, able to adjust its strategy dynamically according to the market state or even to retreat; the vault participant, by contrast, delegates these decisions entirely to the protocol code, gaining a much lower technical barrier and a relatively smooth return curve at the cost of losing the flexibility to exit actively during a crisis.

Dimension	Independent market maker (active)	Vault participant (passive)
Source of return	Fully self-controlled; depends on the maker's own strategy for spread capture and rebates	Controlled uniformly by the protocol engine; shares in aggregate fees and liquidation penalties
Risk control	Dynamically adjusted; can suspend quoting or retreat at any time	Passively borne; constrained by protocol parameters and a withdrawal cooldown
Technical barrier	Very high; requires low-latency infrastructure and quantitative R&D capability	Very low; participation requires only an on-chain transfer
Expected return	Highly dispersed; may earn excess profit or face severe loss	Relatively smooth; reflects the market-average market-making return
Behavior in a crisis	Tends to cancel quotes to avoid unknown tail risk	Strategy rigidity keeps capital continuously exposed, providing floor liquidity

Table 21-3. Risk-return comparison of the independent market maker and the vault liquidity provider (Data source: protocol documentation and inference from microstructure theory [50][51]). Note: this table is organized along the "active-passive participant identity" axis; its "source of return" and "risk control" correspond to the effectiveness and incentive-compatibility dimensions of the four-dimensional framework in Section 21.1.3, and "behavior in a crisis" corresponds to the resilience dimension

From an economic perspective, these two models are not simple substitutes but form complementary layers of the on-chain liquidity ecosystem. The independent market maker represents "elite-tier liquidity": through refined algorithms and very low latency, it provides the tightest bid-ask spreads in normal markets and dominates the price-discovery process. Because it has full autonomy, however, its liquidity provision tends to prove fragile in an extreme crisis.

By contrast, vault participants pool long-tail capital to form "base-tier liquidity." Although the protocol engine's strategy may be less sharp than that of a top independent market maker, its core feature is the rule-verifiability of code execution. Under normal market conditions, vault capital still maintains quotes according to established rules, providing the market with an indispensable liquidity floor. This "active-passive" dual-track architecture lets a decentralized exchange pursue everyday efficiency while gaining a degree of systemic resilience.

21.5.7 Judging effectiveness

On the basis of the simulation above (a single-agent profit-and-loss simulation on the independent-market-maker side) and the vault analysis in Section 21.4, a comprehensive judgment can be made of the protocol-level liquidity incentives. Assessing the effectiveness of an incentive mechanism requires distinguishing two different goals: *attracting* and *retaining*.

On attraction, economic incentives such as rebates and fee shares prove markedly effective in normal markets. As long as the protocol can offer a stable expectation of trading volume, the clear financial return is enough to overcome the initial technical

barrier and friction costs of an independent market maker entering the on-chain environment. The simulation data confirm that, under reasonable parameter settings, an active market-making strategy can achieve a capital return exceeding the risk-free rate, which explains why an emerging decentralized exchange can amass considerable order book depth in a short time.

On retention, however—especially in the face of an extreme market crisis—financial incentives alone are often insufficient to offset the potential losses. When the expected value of adverse selection losses and inventory risk exceeds the marginal utility of the rebate, the rational maker's sole optimal solution remains retreat. This means traditional incentives cannot fundamentally resolve the public-goods dilemma of liquidity provision. Against this backdrop, the protocol vault mechanism reveals its distinctive structural value. The vault exists not to beat independent market makers on efficiency but to provide the market with a compulsory crisis buffer by fixing a market-making obligation onto a portion of capital. The conclusion this section's single-agent simulation can directly support is only that independent market makers tend to retreat in an extreme crisis. The claim that "the vault's floor liquidity absorbs concentrated selling pressure, reduces independent market makers' one-sided inventory risk, and thereby induces them to stay" is a theoretical inference not yet proven by this simulation; testing it would require a multi-agent simulation or empirical data that incorporate the interaction between the vault and independent market makers. Under this mechanism hypothesis, the presence of floor liquidity could improve the survival environment for independent market makers and induce some that had planned to retreat to stay instead, thereby enhancing the stability of the entire market microstructure at the macro level.

21.6 Comparison and coexistence of the two paradigms

The preceding sections analyzed the internal logic of CEX procurement and DEX engineering in turn. This section places the two paradigms within a unified analytical framework for systematic comparison, discusses the substantive meaning of liquidity democratization, and looks ahead to the two paradigms' coexistence and future trends.

21.6.1 Systematic comparison

The central problem of liquidity governance is how to sustain market makers' willingness to participate across different market states. Centralized and decentralized exchanges take sharply different institutional paths to this problem. A centralized exchange *procures* liquidity from professional institutions through bilateral private agreements, whereas a decentralized exchange exemplified by Hyperliquid *engineers* the supply of liquidity through smart contracts and a vault mechanism. The two paradigms display systematic differences along dimensions such as normal-market liquidity quality, crisis resilience, transparency, and barrier to entry.

The quality of liquidity in a normal market is the foremost indicator of market-microstructure efficiency. Through strict capital thresholds and market-making obligation clauses, a centralized exchange can screen for top market makers with

high-frequency trading capability and maintain extremely narrow bid-ask spreads and deep order books on major pairs; a decentralized exchange, constrained by on-chain settlement latency and algorithmic rigidity, can usually attain only a medium-to-high level of normal-market liquidity quality [16]. This difference often reverses under extreme conditions, however: a centralized exchange's agreements generally contain an abnormal-market-conditions exemption that allows market makers to withdraw quotes amid violent volatility, intensifying market fragility just when liquidity is needed most; a decentralized exchange's protocol-level market-making vault, by contrast, quotes continuously according to preset rules and does not retreat on the basis of subjective judgment, providing the market with a liquidity floor—though this may still be overridden by governance intervention under extreme boundary conditions.

Table 21-4 consolidates the analysis above into a systematic comparison of the two paradigms across six core dimensions. CEXs and DEXs exhibit a complementary distribution of strengths across several dimensions: the CEX is superior in normal-market liquidity quality and crisis-response speed, while the DEX holds structural advantages in transparency and barrier to entry; and the difference between them in crisis liquidity resilience—that a CEX's market makers may legally retreat whereas a DEX's vault will not withdraw its orders—constitutes the core finding of this chapter's analysis.

Dimension	CEX procurement	DEX engineering (Hyperliquid type)
Normal-market liquidity quality	Very high	Medium-to-high
Crisis liquidity resilience	Falls sharply (retreat + exemption)	Has a floor, but can be overridden by governance intervention
Transparency	Very low	Very high
Barrier to entry	Very high	Low
Conflict of interest	Severe	Lower but present
Crisis-response speed	Fast (human intervention)	Slow (rule rigidity)

Table 21-4. Systematic comparison of CEX procurement versus DEX engineering (Data source: Aquilina et al. (2024) [16] and protocol-mechanism analysis). Note: this table is a projection of the four-dimensional framework of Section 21.1.3 along the "CEX vs. DEX" axis—"normal-market liquidity quality" corresponds to effectiveness, "crisis liquidity resilience" and "crisis-response speed" correspond to resilience, "transparency" matches the dimension of the same name, and "conflict of interest" corresponds to incentive compatibility; "barrier to entry" is an incremental dimension of this table relative to the four-dimensional framework

On transparency and conflict of interest, a centralized exchange's principal-agent model contains inherent information asymmetry. As rule maker and referee—sometimes even competing directly in the market through a proprietary market maker—the exchange's dual role gives rise to a serious conflict of interest. A decentralized exchange, by contrast, encodes all rules in smart contracts, with strategy parameters and fund flows fully visible on-chain, greatly reducing the cost of trust [3]. This high transparency also brings the risk of the algorithm being gamed, however, and in the face of an unknown crisis, rule rigidity makes its response far slower than a centralized exchange's human intervention.

21.6.2 The democratization of liquidity

The engineering paradigm of decentralized exchanges has not only changed how liquidity is produced but also structurally advanced the *democratization of liquidity*. This concept is not mere marketing rhetoric but comprises three interwoven dimensions: democratization on the supply side, democratization of information, and democratization of risk-bearing [52].

Democratization on the supply side breaks professional market makers' monopoly on market-making returns. In the traditional architecture, market making requires a very high capital threshold and complex infrastructure, and ordinary investors can only be consumers of liquidity (takers) who pay the spread. Through the liquidity-vault design, a decentralized exchange lets any user holding assets inject funds proportionally, aggregating retail capital into institutional-grade market-making capital. This mechanism allows ordinary participants to capture returns at the market-microstructure level, achieving a shift in identity from consumer to producer.

Democratization of information seeks to weaken the informational privilege among market participants. A centralized exchange often provides particular market makers with better order-flow data or lower latency through premium interfaces, and this opaque allocation of privilege intensifies the market's adverse selection risk. In the on-chain order book model, all order-submission, matching, and settlement data are publicly queryable, and the strategy engine's parameters are disclosed through the governance process. Although this high degree of transparency erodes the excess profit that comes from an informational advantage, it significantly reduces the systemic crisis of trust caused by information asymmetry (as noted in Section 21.3.3, privileges such as the on-chain sequencer's execution priority have not been eliminated entirely) [53].

Democratization of risk-bearing is an inevitable result of the liquidity-vault mechanism. A professional market maker usually must bear extreme inventory risk and price-fluctuation risk alone, whereas the vault mechanism spreads these risks, by share, across all liquidity providers. This risk-diversification effect raises the system's capital-carrying capacity under normal conditions but conceals a deep cost. Retail liquidity providers often lack the ability to price complex market-making risks and are easily drawn by high yields into underestimating the tail risk of extreme conditions. At the same time, because setting the strategy parameters requires a high degree of expertise, governance power may ultimately concentrate in a few core developers, diverging at the decision-making level from the original intent of full democratization.

21.6.3 The pattern of coexistence

The two paradigms of liquidity governance are not absolute substitutes but display marked complementarity across different asset classes and market states. The evolution of market structure suggests that centralized procurement and decentralized engineering are forming a pattern of coexistence based on comparative advantage.

In trading major assets (such as Bitcoin and Ethereum), the centralized exchange—by virtue of its very high efficiency and low-latency advantage in normal markets—still dominates price discovery. Professional market makers can maximize profit on these

highly liquid assets and therefore have a strong incentive to participate in a centralized exchange's liquidity procurement. When the market suffers an extreme shock that causes professional market makers to retreat en masse, however, the decentralized exchange's protocol-level vault takes on the function of supplying floor liquidity. Its non-retreating character may cause paper losses in the short run, but it provides a last-resort floor for the price continuity of the entire crypto market. This combination of everyday quality and a crisis floor constitutes the complete resilience of the market structure for major assets.

For long-tail assets, the divergence between the two paradigms is more pronounced. Because long-tail assets have high volatility and unstable trading demand, professional market makers face very high inventory risk and adverse selection costs and are often unwilling to provide liquidity services for them on a centralized exchange. Here the engineering advantage of the decentralized exchange comes fully into play. Through permissionless vault creation and community-driven capital aggregation, a long-tail asset can quickly cross the cold-start threshold of bootstrapping liquidity [3]. In this setting, the decentralized mechanism fills the void that centralized capital is unwilling to enter and becomes the only effective path to price discovery for an innovative asset.

The analysis above rests on an implicit assumption: that the liquidity systems of the CEX and the DEX operate independently, so that their complementarity still holds under extreme conditions. In real markets, however, several channels of risk transmission exist between the two systems and may cause the complementary structure to fail synchronously at the moment it is most needed. The same market maker often operates on both Binance and Hyperliquid, sharing a capital pool and risk limits. When the CEX side suffers a large loss, the maker may be forced to withdraw liquidity from the DEX side at the same time to satisfy its overall risk-control requirements, forming a cross-platform liquidity contagion. More critically, a CEX market maker's triggering of the exemption clause to exit and a redemption run by DEX vault depositors may be set off synchronously by the same macro event (such as systemic deleveraging or a regulatory shock). During the FTX collapse of November 2022, DEX trading volume surged briefly while order book depth likewise contracted sharply, providing an empirical precedent for this kind of resonant failure. The coexistence of the two paradigms is therefore more accurately described as a *conditional complementarity*: complementarity holds under moderate market stress but may fail on both sides during a systemic liquidity shock.

Moreover, as regulatory frameworks gradually clarify, compliance considerations will further entrench this divergence. Centralized exchanges will increasingly tend to serve institutional clients and regulated assets, and their liquidity-procurement agreements will become more standardized and legalized; decentralized exchanges will continue to serve as liquidity infrastructure for borderless, censorship-resistant assets, relying on cryptographic engineering to sustain their distinctive niche.

21.6.4 Future trends

The evolution of the microeconomics of liquidity has not stalled. As the technology matures and market depth increases, three emerging trends are reshaping the frontier

of decentralized liquidity engineering.

AI market-making agents are sharply lowering the technical barrier for independent market makers. Traditional high-frequency market-making strategies require a large quantitative team and expensive computing resources, whereas a new generation of automated algorithmic models can adaptively adjust quoting parameters and inventory-management strategies through machine learning [54][55]. The spread of such intelligent tools will let more small and mid-sized quantitative teams, and even individual developers, connect to decentralized order books as independent market makers, supplementing the more flexible "elite-tier liquidity" beyond the protocol vault. In the four-dimensional framework, this trend mainly improves the effectiveness dimension, converging the DEX's normal-market liquidity quality toward the CEX standard.

Building a cross-chain liquidity layer acts directly on two dimensions: effectiveness and incentive compatibility. At present, liquidity is fragmented across numerous isolated blockchain networks and specific protocols, resulting in low capital utilization. A new generation of cross-chain aggregation architectures schedules liquidity across networks through standardized messaging and asset-locking mechanisms [56]. Once this underlying interoperability is fully realized, a liquidity vault will be able to allocate capital dynamically based on the real-time yields of each chain, evolving from a market maker in a single market into a liquidity provider for the global market.

The rise of the liquidity-as-a-service model is breaking the traditional binding of signed contracts. In this model, liquidity is no longer an appendage tied to a specific exchange but is abstracted into an independent resource that can be called through a standardized interface. A project team or an emerging trading platform no longer needs to conduct cumbersome bilateral negotiations with market makers but instead *rents* the market-making depth it needs directly from a decentralized liquidity network. This modular architecture for liquidity provision not only greatly lowers the startup cost of a new market but also marks liquidity's transformation from a passively procured service into programmable, composable financial infrastructure. In the four-dimensional framework, the liquidity-as-a-service model improves both transparency (a standardized interface replaces private agreements) and resilience (a modular architecture allows rapid switching among liquidity sources), but its incentive compatibility depends on whether the pricing mechanism can accurately reflect the true cost of liquidity across different market states.

21.7 Chapter summary

Chapters 19 through 21 together build a complete analytical framework for understanding the problem of liquidity provision. This chapter has examined two sharply contrasting governance responses: the CEX liquidity-procurement model and the DEX liquidity-engineering model. The former forms a principal-agent relationship by signing private agreements with market makers; it is efficient and flexible under normal conditions, but the adverse selection and moral hazard bred by information asymmetry erupt in concentrated form during periods of extreme volatility, when market makers invoke the exemption clause to retreat legally—shedding their supply obligation pre-

cisely when the system needs liquidity most. The latter, exemplified by Hyperliquid's HLP vault, uses a three-layer architecture of capital, strategy, and distribution layers to turn liquidity provision into a protocol-native automated process, lowering the barrier to entry; it does not eliminate the principal-agent problem, however, but transforms it into a new kind of trust relationship between depositors and protocol code, introducing new dimensions of risk such as algorithmic rigidity, parameter governance, and smart-contract vulnerabilities.

The core finding of this chapter is that the distinctive governance value of the DEX engineering model is not superior efficiency but the verifiability and ex post auditability of its execution rules. Under normal market conditions, the strategy engine executes deterministically according to preset rules and does not withdraw orders on the basis of sentiment or profit motive; but the JELLY and POPCAT incidents show that under extreme boundary conditions, the protocol may still override code execution through governance mechanisms such as a validator vote. This hybrid mode of "code governance in normal times, human discretion in crises" resembles a CEX's human intervention in its outcome, but its process is transparent and auditable on-chain—and this difference, rather than "the determinism of execution," constitutes the true distinction between the two models. Within the four-dimensional framework of effectiveness, incentive compatibility, transparency, and resilience, CEX procurement and DEX engineering form a complementary market structure: the former provides efficient elite liquidity under normal conditions, and the latter provides resilient floor liquidity in a crisis.

As the paradigms of liquidity production and governance evolve, the market's central tension gradually shifts from how to maintain order book depth to how to cope with sharp swings in asset prices. A market maker's inventory risk, the liquidation engine's throughput capacity, and the preservation of vault capital all ultimately depend on the system's ability to price and absorb volatility. When liquidity provision becomes a test of endurance against extreme price movements, a microstructure analysis of volatility itself becomes an unavoidable subject—and this lays the groundwork for the discussion of volatility and derivatives pricing in Part Eight.

References

[1] Binance Square (community submission; author Louis Wang / Biteye Core Contributor; not an institutional Binance Research report). (2025). *In-depth research on Hyperliquid: The DeFi legend from zero to trillion.* Accessed June 2026. https://www.binance.com/en/square/post/28571815431049

[2] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100. https://doi.org/10.1016/0304-405x(85)90044-3

[3] Capponi, A., & Jia, R. (2025). Liquidity provision on blockchain-based decentralized exchanges. *The Review of Financial Studies*, *38*(10), 3040–3085. https://doi.org/10.1093/rfs/hhaf046

[4] Anand, A., & Venkataraman, K. (2016). Market conditions, fragility, and the economics of market making. *Journal of Financial Economics*, *121*(2), 327–349. https:

//doi.org/10.1016/j.jfineco.2016.03.006

[5] Alexander, A. (2025). The impact of market-making incentives on crypto liquidity & price stability. *SSRN Electronic Journal*. https://ssrn.com/abstract=5143194 https://doi.org/10.2139/ssrn.5143194

[6] Aquilina, M., Budish, E., & O'Neill, P. (2022). Quantifying the high-frequency trading "arms race." *The Quarterly Journal of Economics*, *137*(1), 493–564. https://doi.org/10.1093/qje/qjab032

[7] Foucault, T., Kozhan, R., & Tham, W. W. (2017). Toxic arbitrage. *The Review of Financial Studies*, *30*(4), 1053–1094. https://doi.org/10.1093/rfs/hhw103

[8] Tiniç, M., Sensoy, A., Akyildirim, E., & Corbet, S. (2023). Adverse selection in cryptocurrency markets. *Journal of Financial Research*, *46*(2), 497–546.

[9] Madhavan, A. (2000). Market microstructure: A survey. *Journal of Financial Markets*, *3*(3), 205–258. https://doi.org/10.1016/s1386-4181(00)00007-0

[10] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1336.

[11] Delphi Digital. (n.d.). *What is ghost liquidity?* https://members.delphidigital.io/learn/ghost-liquidity

[12] Commodity Futures Trading Commission. (2025). *Market maker agreement: North American Derivatives Exchange (Nadex) d/b/a Crypto.com Derivatives North America* (CFTC org-rules filing, 2025-01; agreement text is the September 2022 version). https://www.cftc.gov/sites/default/files/filings/orgrules/25/01/rules01022513031.pdf

[13] Binance. (2019). *Introducing the Binance Market Maker Program*. https://www.binance.com/en/support/announcement/detail/360034573691

[14] Bellia, M., Christensen, K., Kolokolov, A., Pelizzon, L., & Renò, R. (2022). *Do designated market makers provide liquidity during a flash crash?* (SAFE Working Paper No. 270). Leibniz Institute for Financial Research SAFE.

[15] Chainalysis. (2025). *Market manipulation: Suspected wash trading*. https://www.chainalysis.com/blog/crypto-market-manipulation-wash-trading-pump-and-dump-2025/

[16] Aquilina, M., Foley, S., Gambacorta, L., & Krekel, W. (2024). Decentralised dealers? Examining liquidity provision in decentralized exchanges. *Bank for International Settlements Working Papers*, No. 1227. https://bis.org/publ/work1227.pdf

[17] Biais, B., Foucault, T., & Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial Economics*, *116*(2), 292–313. https://doi.org/10.1016/j.jfineco.2015.03.004

[18] Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press.

[19] O'Hara, M. (2015). *Market microstructure theory*. John Wiley & Sons.

[20] Malinova, K., & Park, A. (2024). Learning from DeFi: Would automated market makers improve equity trading? *SSRN Electronic Journal*. https://ssrn.com/abstract=4531670

[21] Sim, J. Y., Chun, H., & Park, J. (2025). Toward unified liquidity in cryptocurrency markets: A comparative survey and architectural perspective. *International Journal of Advanced Smart Convergence*, *14*(2), 33–45.

[22] Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., … & Juels, A. (2020). Flash boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. *IEEE Symposium on Security and Privacy (SP)*, 910–927. https://doi.org/10.1109/sp40000.2020.00040

[23] Klages-Mundt, A., & Minca, A. (2021). (In)Stability for the blockchain: Deleveraging spirals and stablecoin attacks. *Cryptoeconomic Systems*, *1*(2). https://cryptoeconomicsystems.pubpub.org/pub/klages-mundt-blockchain-instability https://doi.org/10.21428/58320208.e46b7b81

[24] Hyperliquid Foundation. (2025). *Hyperliquid documentation*. https://hyperliquid.gitbook.io/hyperliquid-docs/

[25] Foucault, T., Kadan, O., & Kandel, E. (2013). Liquidity cycles and make/take fees in electronic markets. *The Journal of Finance*, *68*(1), 299–341.

[26] dYdX Foundation. (2023). *v4 deep dive: Rewards and parameters*. https://www.dydx.xyz/blog/v4-rewards-and-parameters

[27] Colliard, J. E., & Foucault, T. (2012). Trading fees and efficiency in limit order markets. *The Review of Financial Studies*, *25*(11), 3389–3421. https://doi.org/10.1093/rfs/hhs089

[28] Braun, B. (2016). From performativity to political economy: Index investing, ETFs and asset manager capitalism. *New Political Economy*, *21*(3), 257–273.

[29] Hyperliquid. (2023). *Protocol vaults*. https://hyperliquid.gitbook.io/hyperliquid-docs/hypercore/vaults/protocol-vaults

[30] Hyperliquid. (2023). *Hyperliquidity Provider (HLP): Democratizing market making*. https://medium.com/@hyperliquid/hyperliquidity-provider-hlp-democratizing-market-making-bb114b1dff0f

[31] Chitra, T., Diamandis, T., Sheng, N., Sterle, L., & Yusubov, K. (2025). Perpetual demand lending pools. *arXiv preprint arXiv:2502.06028*. https://arxiv.org/abs/2502.06028 https://doi.org/10.48550/arXiv.2502.06028

[32] Gornall, W., Rinaldi, M., & Xiao, Y. (2025). Perpetual futures and basis risk: Evidence from cryptocurrency. *SSRN Electronic Journal*. https://ssrn.com/abstract=5036933

[33] Stoikov, S., & Sağlam, M. (2009). Option market making under inventory risk. *Review of Derivatives Research*, *12*(1), 55–79. https://doi.org/10.1007/s11147-009-9036-3

[34] Hendershott, T., & Seasholes, M. S. (2007). Market maker inventories and stock prices. *American Economic Review*, *97*(2), 210–214.

[35] Zealynx. (2026). *Hyperliquidity Provider (HLP)*. Blockchain Security Glossary. https://www.zealynx.io/glossary/hyperliquidity-provider

[36] WisdomTree Prime. (2025). *The great whale slap: How a whale offloaded $4M in*

losses to Hyperliquid's HLP vault. https://www.wisdomtreeprime.com/blog/the-great-whale-slap-how-a-whale-offloaded-4m-in-losses-to-hyperliquids-hlp-vault/

[37] Fournier, M., & Jacobs, K. (2020). A tractable framework for option pricing with dynamic market maker inventory and wealth. *Journal of Financial and Quantitative Analysis*, *55*(5), 1667–1698. https://doi.org/10.1017/s0022109019000462

[38] KuCoin. (2026). *Maximizing the liquidation alpha: How Hyperliquid's HLP vault converts whale losses into liquidity provider yield*. https://www.kucoin.com/news/articles/maximizing-the-liquidation-alpha-how-hyperliquid-s-hlp-vault-converts-whale-losses-into-liquidity-provider-yield

[39] Fan, S., Min, T., Wu, X., & Cai, W. (2023). Towards understanding governance tokens in liquidity mining: A case study of decentralized exchanges. *World Wide Web*, *26*(3), 1181–1200.

[40] Chinco, A., & Sammon, M. (2024). The passive-ownership share is double what you think it is. *Journal of Financial Economics*, *157*, 103860. https://doi.org/10.2139/ssrn.4188052

[41] Chitra, T. (2025). Autodeleveraging: Impossibilities and optimization. *arXiv preprint arXiv:2512.01112*. https://doi.org/10.48550/arXiv.2512.01112

[42] Catania, L., & Grassi, S. (2022). Forecasting cryptocurrency volatility. *International Journal of Forecasting*, *38*(3), 878–894. https://doi.org/10.1016/j.ijforecast.2021.06.005

[43] He, S., Manela, A., Ross, O., & von Wachter, V. (2022). Fundamentals of perpetual futures. *arXiv preprint arXiv:2212.06888*. https://doi.org/10.48550/arXiv.2212.06888

[44] Schmeling, M., Schrimpf, A., & Todorov, K. (2023). Crypto carry. *BIS Working Papers No. 1087*. https://www.bis.org/publ/work1087.pdf

[45] Werapun, W., Karode, T., Suaboot, J., Arporntthip, T., & Sangiamkul, E. (2025). Exploring risk and return profiles of funding rate arbitrage on CEX and DEX. *Blockchain: Research and Applications*. Advance online publication.

[46] Ho, T., & Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, *9*(1), 47–73. https://doi.org/10.1016/0304-405x(81)90020-9

[47] Cartea, Á., Drissi, F., & Monga, M. (2025). Decentralised finance and automated market making: Execution and speculation. *Journal of Economic Dynamics and Control*, *177*, 105134.

[48] Albers, M., Cucuringu, M., Howison, S., & Shestopaloff, A. Y. (2025). The market maker's dilemma: Navigating the fill probability vs. post-fill returns trade-off. *arXiv preprint arXiv:2502.18625*. https://arxiv.org/abs/2502.18625 https://doi.org/10.48550/arXiv.2502.18625

[49] Malamud, S. (2017). Decentralized exchange. *American Economic Review*, *107*(5), 282–287.

[50] Xu, J., Paruch, K., Cousaert, S., & Feng, Y. (2023). SoK: Decentralized exchanges (DEX) with automated market maker (AMM) protocols. *ACM Computing Surveys*,

55(11), 1–50. https://doi.org/10.1145/3570639

[51] Mohan, V. (2022). Automated market makers and decentralized exchanges: A DeFi primer. *Financial Innovation*, *8*(1), 20. https://doi.org/10.1186/s40854-021-00314-5

[52] Krekel, W. P. (2025). *Tokenized, decentralized, democratized? Market microstructure and exchange innovation in digital asset trading* [Doctoral thesis, Macquarie University].

[53] Bruegel. (2023). *Decentralised finance: Good technology, bad finance.* Bruegel Policy Brief.

[54] Addy, W. A., Ajayi-Nifise, A. O., Bello, B. G., Tula, S. T., Odeyemi, O., & Falaiye, T. (2024). Algorithmic trading and AI: A review of strategies and market impact. *World Journal of Advanced Engineering Technology and Sciences*, *11*(1), 258–267.

[55] Dou, W. W., Goldstein, I., & Ji, Y. (2025). AI-powered trading, algorithmic collusion, and price efficiency. *NBER Working Paper No. 34054.* https://www.nber.org/papers/w34054 https://doi.org/10.3386/w34054

[56] Han, Y., Wang, C., Wang, H., Yang, Y., & Wang, X. (2024). A study of blockchain-based liquidity cross-chain model. *PLOS ONE*, *19*(6), Article e0302145. https://doi.org/10.1371/journal.pone.0302145

[57] CoinDesk. (2025). *Hyperliquid delists JELLYJELLY after vault squeezed in $13M tussle.* Accessed June 2026. https://www.coindesk.com/markets/2025/03/26/hyperliquid-delists-jellyjelly-after-vault-squeezed-in-usd13m-tussle

[58] CoinDesk. (2025). *Peak degen warfare: Alleged POPCAT manipulation hits Hyperliquid with $4.9M loss.* Accessed June 2026. https://www.coindesk.com/markets/2025/11/13/peak-degen-warfare-alleged-popcat-manipulation-hits-hyperliquid-with-usd4-9m-loss

[59] Longbridge. (2026). *Hyperliquid vault HLP profited approximately $15 million in the liquidation of the "1011 insider whale".* Accessed June 2026. https://longbridge.com/en/news/274395379

Part VIII
Volatility

Core question: Where does the extreme volatility of perpetual futures markets come from? How does it differ fundamentally from volatility in traditional markets, and can it be predicted, managed, or exploited?

A Four-Source Decomposition of Volatility and Excess Volatility

On March 5, 2024, Bitcoin fell more than 10% within hours from an all-time high of approximately $69,000, and derivative positions worth more than $1.1 billion across the market were forcibly liquidated, of which long positions accounted for roughly $870 million [1]. On the surface, this appears to be nothing more than a single episode of asset-price repricing. Behind the aggregate magnitude of the decline, however, lies a complex microstructure driven jointly by different market mechanisms and participant behaviors. Part of the 10% drop originated in the unexpected release of macroeconomic data, which led investors to reprice risk assets; this portion of the price movement reflected the market's absorption of new information and was therefore permanent in character. Another part originated in market makers withdrawing liquidity as volatility rose, so that subsequent sell orders produced a larger price impact on a shallow order book—a component that tends to revert once liquidity is restored. More strikingly, the initial price decline triggered the margin liquidation of a large number of highly leveraged long positions; the liquidation engine executed forced-closure orders at market prices, generating additional price movement on an already-thin order book that carried no fundamental information whatsoever [2]. At the same time, this sharp local volatility was transmitted through cross-asset linkages to Ethereum and other crypto assets, further intensifying selling pressure across the entire market.

This phenomenon raises the central question of this chapter: price volatility in perpetual futures markets is not a homogeneous stochastic process but a mixture of sources with fundamentally different natures. What, then, are these sources? How much volatility does each contribute? And which portion of volatility is efficient volatility that reflects the arrival of new information, while which portion is inefficient volatility manufactured by structural frictions and institutional mechanisms?

To answer this question, this chapter proposes a four-source decomposition framework for volatility, systematically decomposing the price volatility of perpetual futures into informational volatility, liquidity volatility, mechanism-induced volatility, and transmitted volatility. Informational volatility arises from the arrival of new information such as macroeconomic events and on-chain data, and it represents the healthy process of price discovery. Liquidity volatility is caused by insufficient liquidity supply or by fluctuations in liquidity provision, manifesting as temporary, information-free price shocks. Mechanism-induced volatility is the feature that distinguishes perpetual futures

markets from traditional markets; it is manufactured by institutional design elements such as liquidation cascades, funding-rate settlement, and mark-price deviations, and this type of volatility is almost nonexistent, or of negligible magnitude, in traditional financial markets [3]. Transmitted volatility reflects the spillover of volatility across different assets and markets. Building on this, the chapter introduces the concept of a volatility budget, arguing that total volatility can be viewed as a budget allocated across different sources, and that this allocation depends heavily on the market regime. In calm periods, informational volatility dominates; under market stress or crisis, the noise generated by liquidity frictions and mechanism design takes over most of the volatility budget. This regime-switching mechanism explains why, under extreme market conditions, price movements systematically overreact and are subsequently followed by predictable reversion.

Starting from the theoretical foundations of volatility decomposition, this chapter analyzes in turn the leverage-amplification effect and round-the-clock continuity of informational volatility in perpetual futures, and examines the time structure of liquidity volatility and its gradient distribution between mainstream and long-tail assets. It then analyzes in depth how liquidation cascades and funding-rate settlement act as independent sources of mechanism-induced volatility that inject periodic or explosive price noise into the market, and explores the transmission paths of volatility across markets and across assets. By tracing the interactions among these four volatility sources, the chapter reveals the microstructural foundations of excess volatility in perpetual futures markets and its welfare implications. After reading this chapter, readers will be able to move beyond the traditional view of volatility as a single metric, master quantitative methods for identifying and measuring different volatility sources, understand how the specific institutional design of perpetual futures endogenously manufactures price noise, and build a solid microstructural analytical framework for the subsequent chapters' explanation of volatility clustering and for the construction of volatility forecasts and trading strategies.

22.1 Theoretical foundations of volatility decomposition

Volatility is usually treated as a single risk metric, but this aggregate view obscures the fundamentally different mechanisms that drive price movements. This section first argues for the necessity of volatility decomposition, then reviews the decomposition methods of traditional finance and their limitations in crypto-derivatives markets. On this basis, it proposes a four-source decomposition framework suited to perpetual futures and introduces the analytical concept of a volatility budget to characterize the regime-dependent shares of each source.

22.1.1 The necessity of volatility decomposition

In the quantitative analysis of financial markets, undecomposed aggregate volatility is a composite metric that entails substantial information loss. When an asset's annualized volatility is observed to be 60%, that figure alone reveals nothing about the internal structure of the price movement, nor does it indicate what share represents a healthy price-discovery process and what share arises from noise generated by institutional

frictions. Treating volatility as a single random-walk process conceals the multiple parallel dynamic mechanisms within market microstructure. In perpetual futures—a market with distinctive leverage and liquidation features—the composition of volatility is more complex than in traditional spot markets, so decomposing it precisely is of central theoretical and practical importance.

The value of volatility decomposition is first evident in academic research, where it provides a quantitative basis for distinguishing information-driven efficient volatility from friction-driven inefficient volatility. Under the efficient market hypothesis (EMH), price changes should reflect only the arrival of new information, and this informational volatility is a manifestation of market efficiency. In real markets, however, there are numerous non-informational price shocks, such as amplified slippage caused by drying-up liquidity. Separating the two is a precondition for a deeper understanding of asset-pricing efficiency and a central link connecting price-discovery theory with market-microstructure research. Through decomposition, researchers can accurately measure the net effect of a specific trading mechanism on market quality, thereby enabling cross-sectional comparisons of efficiency across different asset classes or trading venues.

For traders and risk managers, volatility decomposition has direct strategic value. Volatility from different sources exhibits markedly different predictability and mean-reversion characteristics over time. Price changes triggered by fundamental information are typically permanent, representing a revaluation of an asset's intrinsic value, and traders should adjust their positions in line with this trend. By contrast, volatility triggered by short-term liquidity shocks or large-scale leveraged liquidations tends to exhibit overreaction, with prices tending to revert toward their initial level once the shock passes. If a risk-management model cannot distinguish between these two types of volatility, it may trigger unnecessary stop-losses during a false liquidity crisis, or mistakenly increase risk exposure during a genuine revaluation. Decomposing aggregate volatility into subcomponents with different dynamic characteristics is therefore a necessary step in building robust trading strategies and precise value-at-risk (VaR) models.

From the perspective of market designers, volatility decomposition is a quantitative guide for optimizing trading rules and improving protocol mechanisms. A central goal of market design is to minimize the non-essential volatility artificially manufactured by institutional frictions and structural deficiencies. By identifying the specific components of aggregate volatility that can be eliminated through institutional improvements, designers can adjust parameters in a targeted way. For example, if the decomposition shows that most of a particular asset's volatility stems from periodic arbitrage around funding-rate settlement, protocol developers can consider introducing a smoothing mechanism or a continuous-settlement scheme to reduce this mechanism-induced noise. This diagnostic approach, grounded in a decomposition framework, moves decentralized finance (DeFi) governance from an intuition-driven to a data-driven basis, providing solid empirical support for a wide range of parameter adjustments.

22.1.2 A review of traditional decomposition methods

Before examining the distinctive volatility structure of perpetual futures, it is worth reviewing the classic methods that traditional finance uses to address volatility. Early volatility research relied mainly on daily or lower-frequency return data, and this low-frequency sampling severely limited the observation of intraday price dynamics. As high-frequency trading data proliferated, realized volatility came into wide use as a non-parametric estimation method [4]. By summing the squares of high-frequency returns, this method provides a more precise measure of intraday volatility. Realized volatility solves the precision problem of low-frequency data and can effectively capture sharp short-term market movements, but it remains, in essence, an aggregate measure of total volatility and cannot distinguish whether that volatility stems from the release of macroeconomic information, the retreat of market makers, or the endogenous responses of market mechanisms.

To capture the time-varying features of volatility, GARCH models and their many variants have dominated time-series analysis [5]. This family of models has successfully described the volatility clustering that is pervasive in financial markets—large movements tend to be followed by large movements, and small movements by small movements. These models, however, focus on the statistical description of the volatility process rather than on explaining its economic origins. They lump all factors that drive price changes together as random shocks and cannot dissect the nature of these shocks at the microstructural level. When confronted with structural volatility triggered by specific institutional design, purely statistical models struggle to capture its generating mechanism, because they assume that shocks are homogeneous across market regimes.

Market-microstructure theory provides finer analytical tools for understanding the intrinsic mechanisms of price movements. Within a framework of information asymmetry, scholars have proposed a series of decomposition models to isolate different types of price shocks. Among them, the information-share method is widely used to measure the contribution to price discovery in multi-market settings [6] (for a methodological review, see [7]). Based on the common efficient-price random-walk component in a cointegration system, this method separates out the long-run equilibrium price of related assets, thereby quantifying each market's relative contribution to incorporating new information into the price process. In parallel, microstructure research has sought to decompose the price change caused by a single trade into a permanent component and a transitory component. The permanent component reflects the private information embedded in the trade, whereas the transitory component is treated as liquidity friction arising from inventory risk and order-processing costs. In addition, estimation methods based on the bid-ask spread offer a simple way to infer transaction costs and non-informational volatility [8].

Although these traditional methods have achieved notable success in equity and foreign-exchange markets, applying them directly to crypto-asset perpetual futures markets encounters clear limitations. Most classic models divide volatility into only two broad categories—information and noise, or permanent and transitory. This dichotomy may suffice for spot markets that lack complex derivative features, but it is too coarse for perpetual futures. Transitory volatility in perpetual futures is in fact a complex

composite: it contains not only market makers' liquidity frictions but also, deeply, the institutional price movements triggered by funding-rate settlement and leveraged liquidation cascades. Without further separating these subcategories of fundamentally different nature, one cannot accurately assess the true operating efficiency of perpetual futures markets. More importantly, traditional methods are usually built on the implicit assumption of a friction-free institutional environment and fail to adequately account for the unique features of crypto markets—round-the-clock trading, the prevalence of high leverage, and intense cross-asset contagion—which requires us to build a more comprehensive and fine-grained decomposition framework on the existing theoretical foundation.

22.1.3 The four-source decomposition framework

To fully parse the intrinsic structure of price movements in perpetual futures markets, we propose a four-source volatility decomposition framework. The framework decomposes aggregate realized volatility into four subcomponents of markedly different natures and driving mechanisms: informational volatility, liquidity volatility, mechanism-induced volatility, and transmitted volatility. This conceptual identity is not merely a statistical variance decomposition but an attribution model in the microeconomic sense, revealing how different types of shocks are ultimately converted into observable price changes through specific market channels.

Informational volatility is the permanent price adjustment caused by the arrival of new fundamental information, macroeconomic data releases, or idiosyncratic on-chain events. This volatility is a direct expression of the market's price-discovery function and represents a reasonable revaluation of an asset's intrinsic value. Informational volatility is irreversible: once new information is absorbed, the price stabilizes at a new equilibrium level rather than undergoing systematic mean reversion. In any efficient financial market, informational volatility cannot be eliminated; far from being a market defect, it is a sign of healthy market functioning. In perpetual futures markets, because of leverage, information of a given magnitude may trigger larger-scale position adjustments and thus amplify the magnitude of informational volatility in the short run—but in essence the price is still simply reflecting the new market environment.

Liquidity volatility arises from temporary price shocks produced by insufficient market depth or by adjustments in market makers' quoting strategies. When a large order is executed on an order book that lacks sufficient supporting liquidity, the price slips significantly; this deviation is not due to any change in the asset's fundamentals but is pure trading friction. The core feature of liquidity volatility is its temporariness: after the shock, as liquidity providers re-enter and arbitrageurs act, the price tends to revert toward the pre-shock equilibrium level. In perpetual futures markets, liquidity volatility often exhibits positive-feedback characteristics—an initial rise in volatility causes market makers to widen spreads and retreat, which in turn causes subsequent orders to produce larger price impacts. This volatility is information-free noise and can, in principle, be reduced by introducing more efficient market-making mechanisms or by deepening the market.

Mechanism-induced volatility is the core volatility source that distinguishes perpetual

futures markets from traditional markets, and its negative impact on price stability is especially pronounced. It is generated entirely by institutional design features unique to this derivative, including the forced-liquidation cascades of leveraged positions, the periodic settlement pulses of the funding rate, and the computational deviation between the mark price and the last traded price. This type of volatility contains no new information about an asset's true value, yet it can manufacture extreme price movements [2]. For example, when the price touches a key liquidation line, the liquidation engine dumps the liquidated positions at market prices, causing significant price deviation on an order book that lacks depth; such a decline, triggered purely by the mechanical execution of an institutional mechanism, is a textbook case of mechanism-induced volatility. Like liquidity volatility, mechanism-induced volatility usually reverts to some degree after the fact, but if the scale of liquidation is too large, it can also trigger systemic liquidity exhaustion, causing prices to deviate from fundamentals for an extended period. Mechanism-induced volatility is a direct cost of the institutional design of perpetual futures and can, in principle, be substantially reduced by optimizing liquidation algorithms and fee structures.

Transmitted volatility describes the spillover effect that spreads in from the violent volatility of other markets or related assets through cross-market arbitrage and cross-margin mechanisms. In a highly interconnected network of crypto assets, sharp volatility in the Bitcoin spot market is rapidly transmitted to its perpetual futures, and liquidation events in mainstream assets often trigger chain reactions in altcoins through balance-sheet effects. The nature of transmitted volatility depends on the character of its source: if the source is a macroeconomic information release, the transmitted volatility carries some information content; if the source is a local liquidity collapse, the transmitted volatility is purely contagious noise. The continuous trading and high fragmentation of perpetual futures markets make this cross-market, cross-asset transmission exceptionally rapid and large.

These four volatility sources differ fundamentally in nature, duration, information content, and policy implications. Informational volatility is a high-information-content permanent adjustment; liquidity volatility is an information-free temporary friction; mechanism-induced volatility is noise that is endogenous to institutional design and can be substantially reduced but not entirely eliminated; and transmitted volatility is an external input from cross-market linkages.

At the formal level, the four-source decomposition framework can be expressed as an attribution decomposition of aggregate realized variance. Let the total realized variance of an asset's log-price process p_t over the time window $[0, T]$ be $RV_T = \sum_{i=1}^{N} r_i^2$, where r_i is the i-th high-frequency return observation. The four-source decomposition treats each micro price change r_i as the superposition of four types of shocks: $r_i = r_i^{info} + r_i^{liq} + r_i^{mech} + r_i^{cond}$. The total variance can then be decomposed as:

$$RV_T = \sum_{i=1}^{N} (r_i^{info})^2 + \sum_{i=1}^{N} (r_i^{liq})^2 + \sum_{i=1}^{N} (r_i^{mech})^2 + \sum_{i=1}^{N} (r_i^{cond})^2 + \sum_{i=1}^{N} \sum_{j \neq k} r_i^j r_i^k$$

The first four terms on the right-hand side correspond to the direct contributions of the four sources, and the last term is the interaction term. In calm market periods, the

shocks from each source are approximately uncorrelated in time, the interaction term approaches zero, and the total variance is approximately additive. In crisis periods, however, liquidity exhaustion and liquidation cascades occur simultaneously within the same time window, the interaction term is significantly positive, and the total variance exceeds the sum of the independent contributions of the individual sources—which is precisely the mathematical basis for the superlinear amplification of volatility under extreme market conditions.

The linear additive decomposition above is essentially a first-order approximation, whose validity depends on the assumption that the shocks from each source are approximately independent in time. In calm periods this assumption broadly holds, the interaction term is negligible, and the total variance is approximately additive. In crisis periods, however, the four sources are highly entangled: a single liquidation-cascade event simultaneously produces a mechanism-induced shock (liquidation-engine execution), a liquidity shock (order-book penetration), and a transmitted shock (cross-asset contagion), and the interaction term may account for a substantial share of total variance. In such cases, the four-source decomposition should be regarded as a conceptual attribution tool rather than a precise statistical decomposition; its core value lies in revealing the multiple generating mechanisms of volatility rather than in providing precisely additive figures. Future research could introduce regime-switching structural models or nonlinear realized-covariance methods to handle the interaction effects of crisis periods more precisely.

At the level of empirical identification, the approximate separation of the four sources can draw on the following strategies. Mechanism-induced volatility can be identified via an event-window method, because liquidation timestamps are precisely known and funding-rate settlement times are fixed. Transmitted volatility can be separated along the cross-asset dimension via Granger causality tests or the Diebold-Yilmaz (DY) spillover index. Informational volatility can be estimated along the time-series dimension via permanent price impact or the information-share method. Liquidity volatility is then treated as a residual category, obtained approximately by subtracting the identifiable portions of the other three from total volatility. This identification strategy works well in normal markets, where the sources are well separated in time, but in crisis periods, when the sources fire simultaneously, the boundaries inevitably become blurred. Treating liquidity volatility as a residual category carries two costs. First, the fact that the four sources sum to exactly 100% is an artifact of this residual definition rather than a result of the framework being empirically tested; whether an unidentified fifth source exists (such as market-maker reflexivity or cross-exchange oracle-latency arbitrage) must be corroborated by examining the time-scale structure of the residual, and will not automatically reveal itself as a residual. Second, the reversion time scale of the narrative-driven volatility mentioned in Section 22.2 (several days to several weeks) is significantly longer than that of pure liquidity shocks (on the order of minutes), and under the residual method it would be merged into the liquidity bucket; a separate robustness check is therefore advisable for the medium- to long-horizon reversion component.

This decomposition framework further refines the permanent/transitory dichotomy proposed by Hasbrouck [9]: the “transitory component” of the traditional framework is

here disaggregated into three subcategories—liquidity volatility, mechanism-induced volatility, and transmitted volatility—each with a different generating mechanism and policy implication. Figure 22-1 compares the core characteristics of the four volatility sources along four dimensions: information content, duration, eliminability, and policy implications.

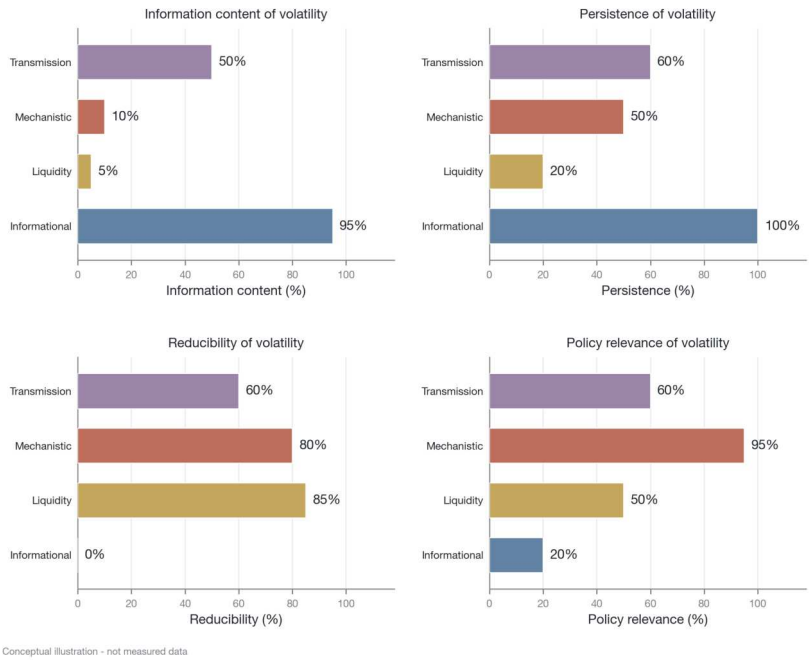


Figure 22-1. Comparison of the characteristics of the four volatility sources (drawn by the author)

As shown in Figure 22-1, mechanism-induced volatility is both information-free and highly eliminable, which makes it the primary target for institutional optimization. The core significance of this taxonomy is that it provides a unified framework for assessing the economic meaning of volatility: not all volatility is *bad*, and the key is to distinguish which volatility stems from price discovery and which from institutional friction.

22.1.4 The concept of a volatility budget

To understand the internal structural dynamics of aggregate volatility more intuitively, we introduce the analytical concept of a volatility budget. Building on the variance decomposition above, the volatility budget treats the total variance observed within a given time window as a 100% budget pool, in which the four sources occupy shares corresponding to their respective variance contributions (with a reasonable allocation of the interaction term). We adopt the convention of allocating the interaction term symmetrically and in full to both of the sources involved—that is, each pairwise interaction product is counted once in full within each of the two sources it involves. Other allocation schemes would change the absolute share of each source but would not alter the qualitative conclusion, described below, that the volatility budget switches with

the market regime. Formally, the budget share of the j-th volatility source is defined as:

$$w_j = \frac{\sum_{i=1}^{N}(r_i^j)^2 + \sum_{i=1}^{N}\sum_{k\neq j} r_i^j r_i^k}{RV_T}, \quad j \in \{info, liq, mech, cond\}$$

Here w_j is a signed share of the variance contribution: it satisfies the normalization $\sum_j w_j = 1$, but individual terms need not fall within the interval $[0, 1]$—when the interaction term between one source and the others is net negative, w_j may arithmetically be less than zero or greater than one.

This relative-share perspective allows the analyst to look beyond the surface of absolute volatility figures and discern the microstructural health of the market under different operating regimes (note that when the interaction term is negative—for example, when informational and liquidity shocks are negatively correlated—the budget share w_j of an individual source may arithmetically be less than zero, in which case the "100% budget pool" is merely a heuristic expression). The allocation of the volatility budget is not fixed; it undergoes a systematic structural switch as the market evolves from calm to crisis, and its detailed mechanism and empirical features are analyzed in Section 22.6. Figure 22-2 uses a stacked area chart to present the structural evolution of the volatility budget from calm periods to crisis periods.

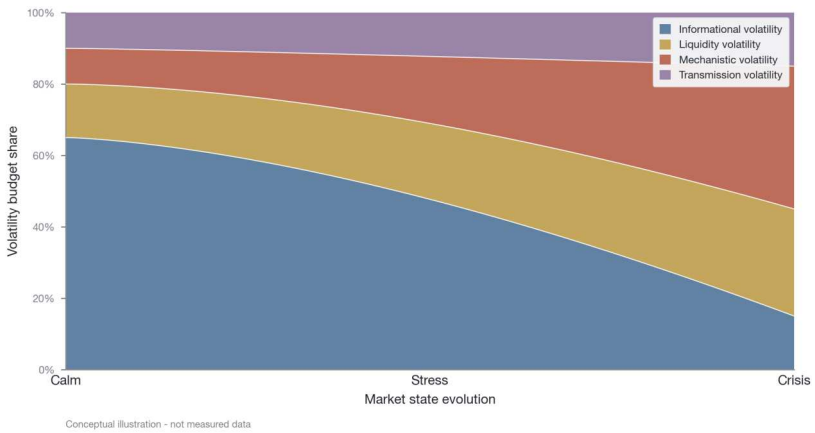


Figure 22-2. State dependence of the volatility budget (illustrative volatility budget: drawn by the author, with proportions referencing Nguyen (2024) [10] and Dimpfl & Peter (2021) [11]; not recomputed from measured data)

As shown in Figure 22-2, the budget share of informational volatility is sharply compressed as the market moves from calm to crisis, while liquidity and mechanism-induced volatility expand dramatically until they cover almost the entire budget pool. This structural reversal conveys the central thesis of this chapter: under extreme market conditions, the economic meaning of price movements undergoes a fundamental change.

22.2 Informational volatility

As previewed by the framework in Section 22.1.3, informational volatility is the price change—triggered by the arrival of new information and its digestion by the market—that represents a permanent revaluation of an asset's fundamental value, in contrast to the temporary volatility caused by liquidity frictions. In the four-source framework, "information" is defined broadly, consistent with the EMH: any signal that alters market participants' price expectations—macroeconomic data, on-chain anomalies, and even social-media narratives and opinion-leader statements—falls into this category, so that price changes driven by retail fear of missing out (FOMO) or by herding also count as informational volatility. The persistence of narrative-driven price changes, however, lies between that of fundamental information shocks (highly permanent) and pure liquidity shocks (reversion on the order of minutes): such changes may revert significantly within several days to several weeks, constituting a known gray area of the four-source taxonomy. In perpetual futures markets, round-the-clock trading and the nonlinear amplification of information shocks by high leverage mean that informational volatility often takes on a sharper and more rapid pattern of response.

22.2.1 Macro information shocks

As crypto assets have been absorbed into global macro asset allocation as an alternative investment, the pricing of Bitcoin and other major crypto assets has fallen increasingly under the influence of traditional macroeconomic variables. Releases of core macroeconomic indicators—the interest-rate decisions of the Federal Open Market Committee (FOMC), the consumer price index (CPI), and nonfarm payroll data—constitute the largest and most frequent sources of external information shocks in the crypto perpetual futures market. By altering expectations for the risk-free rate and the state of U.S. dollar liquidity, these macro data directly reshape investors' required risk premium on high-risk assets. Empirical research shows that during the window in which the FOMC releases a monetary-policy statement, the realized volatility of Bitcoin perpetual futures rises sharply intraday in a pulse-like pattern [13]; research based on daily-frequency data, by contrast, finds that the overall effect of such macro announcements on the Bitcoin price is relatively limited [12], indicating that this pulse effect is concentrated mainly at the intraday high-frequency level. This sharp rise in volatility is not a random walk but a direct expression of the market repricing its expectations for future funding costs.

Under the shock of these macro events, the response mechanism of the perpetual futures market often exhibits extremely high sensitivity and immediacy. Because macro data releases usually follow a clear schedule, market participants make substantial position adjustments and engage in expectation games before the data are published. Once the actual figures deviate from the market's consensus expectation—producing a so-called macro "surprise"—the perpetual futures market generates enormous trading volume within a short time. At this point, as the stop-loss or take-profit orders of leveraged traders are triggered en masse, liquidity providers, seeking to avoid adverse-selection risk, tend to rapidly withdraw their limit orders, causing order-book depth to thin instantaneously. This deterioration in liquidity conditions, superimposed on the infor-

mation shock, jointly amplifies the instantaneous magnitude of the price movement. Research indicates that when the CPI comes in above expectations, the instantaneous volatility of the Bitcoin price can reach several times its normal level, and this volatility is largest in the first few minutes before gradually decaying [13].

Beyond traditional macroeconomic data, crypto markets also face a distinctive type of macro information shock: the evolution of regulatory policy and compliance events. The progress of the U.S. Securities and Exchange Commission (SEC) in approving spot exchange-traded funds (ETFs), and regulatory classification announcements on crypto assets by major jurisdictions, can trigger sharp swings in market sentiment. Unlike regularly scheduled economic data, the timing of such regulatory events is often unpredictable, and their impact on industry fundamentals is more far-reaching. For example, during the period in which expectations for the approval of a spot Bitcoin ETF gradually accumulated and finally materialized, the volatility of the crypto market exhibited complex dynamics, encompassing both expectation-driven trending volatility and violent "sell-the-fact" swings after the event materialized [14]. Unlike the mainly pulse-like shocks of regularly scheduled macro data, the informational volatility triggered by regulatory policy is often not a discrete event but a persistent state of uncertainty: even after an ETF has been approved, the SEC's ongoing regulatory authority and its cautious stance toward future products still constitute a long-standing, unresolved risk. Figure 22-3 shows the typical dynamic pattern of intraday realized volatility for BTC perpetual futures before and after a macro information shock.

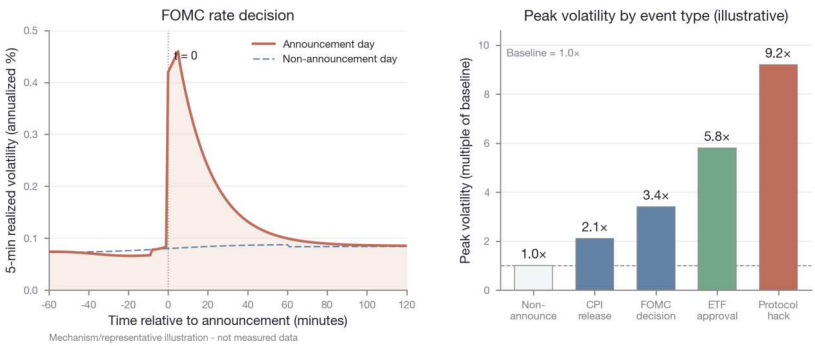


Figure 22-3. Intraday realized volatility of BTC perpetual futures before and after a macro information shock (mechanism illustration: depicting the impulse-decay pattern of 5-minute realized volatility in the ± 2 hours around macro events such as FOMC meetings; a synthetic representative curve, not measured Binance data)

As shown in Figure 22-3, realized volatility jumps in a pulse at the instant the macro data are released, peaks within a few minutes, and then gradually decays. This "shock-decay" profile is consistent with the characteristics of informational volatility: the rise in volatility is a necessary accompaniment to the price-discovery process in which the market absorbs new information and completes its repricing.

22.2.2 On-chain information and idiosyncratic events

Unlike traditional financial markets, which rely on corporate earnings reports and audited disclosures, the crypto-asset market possesses a unique public information layer: the blockchain ledger. On-chain data—large token transfers, changes in the number of active addresses, shifts in miner balances, and changes in total value locked (TVL) in decentralized finance protocols—constitute a form of "quasi-public inside information" unique to crypto markets. Although these data are publicly visible to all market participants, the barriers and latency of data interpretation mean that investors able to monitor and interpret these on-chain anomalies in real time often gain an informational edge. When a whale address transfers a large amount of Bitcoin into a centralized exchange (CEX), the market typically reads this as a potential selling signal, triggering defensive short-selling and downward price volatility [15]. Volatility triggered by such on-chain anomalies is, in essence, the market digesting the supply-demand imbalance risk implied by this "quasi-public information."

Although this book focuses on the order-book markets of centralized exchanges, protocol-level data changes in DeFi are also transmitted through arbitrage channels to the CEX perpetual futures market and have indirect but significant explanatory power for the price volatility of the relevant governance tokens. As a core metric of protocol liquidity and user trust, TVL exhibits a strong dynamic causal relationship between its trend and token prices. Research finds that when a protocol's TVL breaks through a particular threshold or shows accelerating growth, it often signals strengthening market confidence in that protocol, which in turn attracts more capital inflows and drives up the token price [16]. Conversely, if a protocol suffers a security exploit or large-scale capital outflows, a sharp drop in TVL will trigger panic selling. In these extreme scenarios, the perpetual futures market often becomes the primary venue for risk hedging. Traders use the short-selling mechanism of perpetual futures to hedge their spot exposure or to speculate on the short side, which sharply amplifies, in the derivatives market, the volatility triggered by project-idiosyncratic events.

For small- and mid-cap altcoins, project-idiosyncratic events—such as a mainnet upgrade, a change to the token-economic model, or turnover among core team members—are often the central drivers of price volatility. Because these assets have relatively poor liquidity and lack a mature market-maker ecosystem, any material fundamental information can cause a violent revaluation of the price. In the perpetual futures market, the funding rates of these assets often deviate to extremes before a major event, reflecting the market's strong directional expectations. When the event finally materializes, if the outcome fails to meet the market's high expectations, it often triggers a so-called short squeeze or a concentrated closing of long positions, causing an enormous price swing in a very short time. This volatility, driven by project-idiosyncratic events, reveals how crypto markets price assets under information asymmetry and liquidity stratification.

22.2.3 Informed trading and volatility decomposition

The micro-mechanism of informational volatility must be parsed at the level of trading behavior. As discussed in Chapter 13, price discovery is a core function of market

microstructure, and informed traders play a central role in this process. In the four-source decomposition framework, informed trading is the main micro-mechanism that generates informational volatility: the strategic trading behavior of informed traders incorporates private information into prices and drives prices toward fundamental value, and the price changes they contribute constitute the permanent component of informational volatility. According to the classic Kyle model, informed traders trade strategically using their informational advantage, a process that inevitably triggers price changes [17]. Complementing this, the probability of informed trading (PIN) model proposed by Easley and colleagues [18] provides a direct tool for quantifying the intensity of informed trading within a given period. A rise in volatility is therefore not always a sign of market inefficiency; in many cases it is a necessary stage in which prices reflect new information and achieve more accurate pricing.

In perpetual futures markets, the price-impact effect of informed trading is significantly amplified by leverage. When informed traders anticipate that an asset's price will change, they tend to use the high-leverage feature of perpetual futures to maximize the returns to their informational advantage. This leveraged informed trading produces enormous directional order flow that rapidly consumes liquidity on the order book. Facing such strong one-sided order flow, market makers realize that they may be trading against counterparties with an informational advantage—that is, they face severe adverse-selection risk. To protect themselves, market makers quickly adjust their quotes, widen the bid-ask spread, and reduce the number of resting orders [19]. This defensive behavior at the microstructural level further intensifies price volatility. Informational volatility in perpetual futures markets therefore contains not only the price revaluation brought by the information itself but also the liquidity premium caused by adverse selection.

From the perspective of volatility decomposition, total volatility can be split into permanent volatility driven by information and transitory volatility driven by microstructure noise. Informed trading contributes mainly to the permanent-volatility component, because these trades reflect changes in an asset's fundamental value, and this price change persists over the long run once the information is fully digested by the market. By contrast, volatility caused by liquidity shocks, order-book imbalances, or irrational speculation usually mean-reverts within a short time and belongs to transitory volatility. In crypto perpetual futures markets, because market sentiment is extremely sensitive and speculative forces are active, the share of transitory volatility tends to be high. When key information is released or a major event occurs, however, the share of permanent volatility rises significantly, indicating that the price-discovery process is dominating market dynamics.

22.2.4 Leverage amplification and round-the-clock continuity

Informational volatility in the crypto-asset perpetual futures market has two salient features: a leverage-amplification effect and the absolute continuity of information arrival. First, the leverage mechanism has both an accelerating and an amplifying effect on informational volatility. Leverage itself does not directly create new information; what it amplifies is the magnitude by which an information shock is converted into a

price change in the market. But leverage also indirectly affects the incentive structure for information production: when an informational edge can be amplified into higher returns through high leverage, the returns to resources invested in information-discovery activities—such as on-chain data monitoring and macro forecasting—are higher, which attracts more informed traders into the market and indirectly raises the level of informational volatility. When information such as macro data or on-chain anomalies arrives, the presence of highly leveraged traders enables the same amount of capital to generate order flow several or even dozens of times larger. When this amplified order flow hits the order book, it causes deeper price slippage. More importantly, the initial price movement can easily touch the liquidation lines of highly leveraged positions, triggering chain forced liquidations. The passive market orders produced by this forced-liquidation mechanism form a positive-feedback loop with the initial information-driven orders, causing extreme price movements that far exceed the fundamental change implied by the information itself [20].

Second, the round-the-clock trading feature of crypto markets fundamentally changes the temporal distribution of information arrival and digestion. In traditional stock or futures markets, fixed trading sessions and market closures mean that information accumulated during non-trading hours is often released in a concentrated fashion at the open, in the form of gap-ups or gap-downs, causing abnormally high volatility in the early open. The round-the-clock nature of the crypto perpetual futures market allows information to be priced in instantly at any moment. Although this continuity eliminates overnight-gap risk, it also means that market liquidity varies significantly across the hours of the day. Empirical observation shows that intraday volatility in crypto markets exhibits a pronounced geographic time-zone pattern and intraday seasonal structure, with the handover points among the major Asian, European, and North American trading sessions often being peaks of volatility and trading volume [21][13]. When major information arrives during a period of relatively thin market liquidity (such as North American weekends or late at night), even a small information shock can trigger enormous price volatility for lack of sufficient absorbing orders.

Moreover, the degree of information asymmetry in perpetual futures markets is, in many dimensions, higher than in traditional financial markets. Although on-chain data provide unprecedented transparency, this transparency is structural. Institutional investors and quantitative market makers possess powerful data infrastructure and algorithmic capabilities that let them parse on-chain anomalies, exchange fund inflows and outflows, and minute changes in macro data within milliseconds, and quickly establish positions in the perpetual futures market. Retail investors, by contrast, are often at the tail end of information reception and processing. This asymmetry in technology and resources enables informed traders to extract information rents more efficiently in perpetual futures markets, while retail investors more readily become passive providers of liquidity. This deep information asymmetry further reinforces the perpetual futures market's tendency toward one-sided volatility under information shocks and the steepness of its price discovery.

22.2.5 Measuring informational volatility

To isolate and measure informational volatility quantitatively, financial econometrics offers a set of rigorous analytical tools, the most central concept being permanent price impact. Permanent price impact measures the net change in price relative to the pre-trade level after a sufficiently long time has elapsed following a trade. This measure rests on a basic assumption: if a trade contains undisclosed private information, that information will ultimately be fully absorbed by the market, causing a permanent change in the asset's fundamental value; conversely, if the trade is driven merely by liquidity needs or noise trading, the price change it triggers will ultimately mean-revert. Within the framework of a vector autoregression (VAR) model, researchers can jointly model trade flow and quote updates to isolate the long-run cumulative impact of trade innovations on price, thereby precisely quantifying the informational-volatility component in the market [9].

In an environment where the same asset trades in multiple markets, assessing the contribution of different markets to price discovery and informational volatility is a key microstructure question. The Hasbrouck information-share model provides a classic solution to this problem. Based on the common efficient-price random-walk component in a cointegration system, the model defines each market's information share by computing the proportion of the variance of that common stochastic trend contributed by each market's innovations [6]. In the crypto-asset space, mainstream assets such as Bitcoin trade simultaneously on multiple spot exchanges and derivatives platforms. Studies applying the information-share model consistently show that lightly regulated offshore perpetual futures markets occupy a dominant position in Bitcoin price discovery. With their high leverage, low trading frictions, and vast liquidity, these platforms attract large numbers of informed traders, and their price innovations often lead compliant spot markets and regulated traditional futures markets [22].

Beyond high-frequency measures at the microstructural level, the event-study method is also a commonly used macro tool: by defining a specific information event (such as an FOMC decision or a protocol being hacked) and examining the abnormal returns and abnormal volatility within the event window, one can assess the market impact of the information shock. In the crypto perpetual futures market, macro monetary-policy tightening typically triggers a systematic rise in market-wide volatility, whereas a negative security event targeting a specific decentralized finance protocol triggers local extreme volatility in that token and may generate cross-asset volatility premia through nested capital pools. In addition, as mentioned in Section 22.2.3, the PIN model [18] provides an independent measurement perspective on the share of informational volatility by modeling the arrival rates of buy and sell orders. Combining permanent price impact, information share, PIN estimation, and event studies can fully describe the path by which information is converted into price volatility in perpetual futures markets. Figure 22-4 presents, using actual data, the decomposition of a price shock in BTC perpetual futures into its permanent and transitory components.

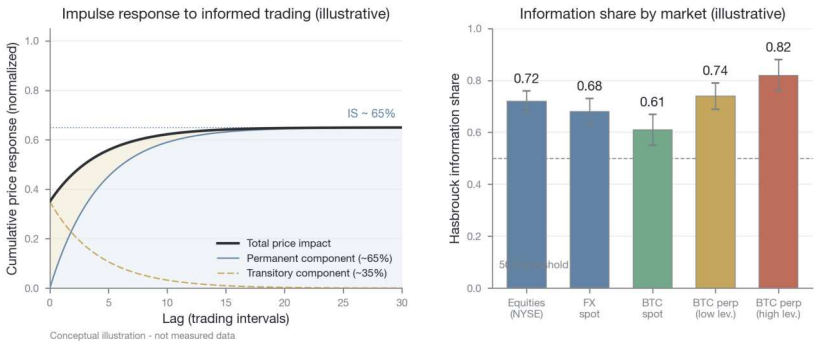


Figure 22-4. Decomposition of a BTC perpetual futures price shock into permanent and transitory components (mechanism illustration: a representative pattern based on the vector-autoregression (VAR) decomposition logic of Hasbrouck (1991) [9], not recomputed from measured data)

As shown in Figure 22-4, the permanent component (informational volatility) persists over the long run after the shock, while the transitory component rapidly decays back toward zero. The separation of the two shows that a single price change in perpetual futures contains both information and noise, which is precisely the microstructural empirical basis for this chapter's four-source framework.

22.3 Liquidity volatility

In contrast to the permanent nature of informational volatility, and as previewed in Section 22.1.3, liquidity volatility is a temporary price shock produced by insufficient market depth or by adjustments in market makers' quotes: it contains no information about an asset's true value, and the price tends to revert after deviating, simply because the order size exceeds the immediate liquidity the market can provide at the time. Even in calm moments when no new information arrives, executing an order on an illiquid order book produces such a shock. In perpetual futures markets, because of the endogenous nature of liquidity, this volatility manufactured by insufficient liquidity has a strong self-amplifying effect.

22.3.1 Illiquidity and spurious volatility

In an ideal frictionless market, price changes should fully reflect the arrival of new information. In the real crypto-trading environment, however, prices frequently jump violently because of pure supply-demand imbalances. This price change, which is not information-driven, is termed liquidity-induced volatility—that is, non-informational temporary volatility—in market-microstructure theory. Understanding this type of volatility hinges on recognizing the nonlinear relationship between price impact and order-book depth.

Consider a stylized example (actual slippage varies with the asset, the time of day, and the state of the order book; for BTC perpetual futures, the slippage of a $1 million market order is typically on the order of single-digit basis points during peak-liquidity hours, but can reach tens of basis points during the low-liquidity hours of late night). When a $1 million sell order enters the market, the price impact it causes depends

entirely on the thickness of the order book at that moment. If this order lands on an order book with $10 million of bid depth, it may consume only the top few price levels, causing a price decline of about 20 basis points. But if the same sell order is executed on an order book with only $1 million of total depth, it will penetrate many price levels, causing a decline of 80 basis points or more. In both scenarios, the act of selling itself conveys no different fundamental information; solely because of the difference in absorptive capacity, the price behavior differs sharply. The core parameter λ of the Kyle model (the price-impact coefficient) precisely characterizes this mapping between order flow and price change [17].

In the crypto-asset perpetual futures market, such spurious volatility is often more pronounced than in the spot market. The high-leverage feature of perpetual futures means that a small amount of initial capital can drive an enormous notional trading volume, which in practice amplifies the impact of a single order on the order book. At the same time, perpetual futures market makers face more complex risk-management constraints when providing liquidity. When the market moves in one direction or expected volatility rises, market makers, to protect their own positions, tend to rapidly withdraw resting orders or significantly widen the bid-ask spread, causing the order book to become abnormally thin precisely when liquidity is most needed. This instantaneous withdrawal of liquidity depth means that even a normal-sized market order can easily trigger sharp price slippage, producing large outliers in the price series.

The core criterion for distinguishing this liquidity volatility from genuine informational volatility lies in the trajectory of the price after the shock. Price changes driven by genuine information are often permanent: the price jumps or drops to a new equilibrium level and stabilizes there, because market participants have re-assessed the asset's value in light of the new information. By contrast, spurious volatility caused by insufficient liquidity is usually temporary. Once that large, liquidity-exhausting order has been executed, if no follow-on orders appear, market makers and other arbitrageurs will quickly notice that the price has deviated from fair value, and will therefore re-supply liquidity and push the price back toward its pre-shock level. This "shock-reversion" pattern is the typical signature of liquidity volatility. Figure 22-5 illustrates this mechanism along two dimensions: the decisive influence of order-book depth on the magnitude of price impact, and the fundamental difference in the post-shock trajectories of informational and liquidity shocks.

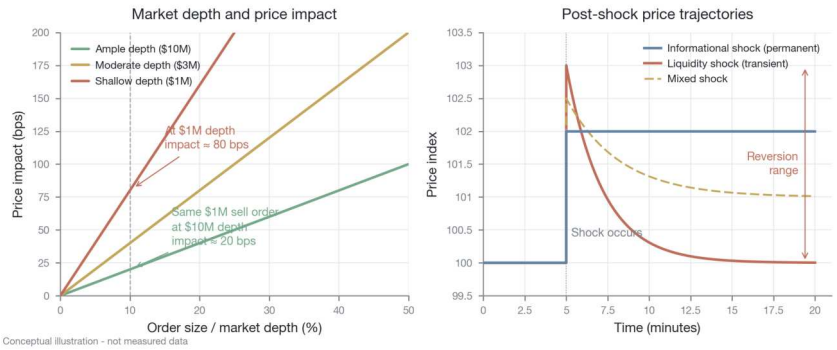


Figure 22-5. The relationship between order-book depth and price impact (representative mechanism illustration: depicting the relationship between market-order impact and order-book depth for BTC perpetual futures under different liquidity conditions; not measured Kaiko snapshot data)

As shown in Figure 22-5, when order-book depth is insufficient, the impact of the same order flow is amplified nonlinearly; an informational shock follows an "L-shaped" permanent path, whereas a liquidity shock follows a "V-shaped" temporary reversion path. This difference in post-shock trajectory provides an operable identification strategy for empirically distinguishing the two types of volatility.

22.3.2 Market-maker retreat and the volatility positive-feedback loop

Liquidity volatility does not always appear in an isolated, random form. In extreme crypto-market conditions, liquidity volatility often exhibits a highly structured self-amplifying mechanism. The root of this mechanism lies in the endogenous nature of market liquidity: liquidity is not only a result of price volatility but also a cause of further price volatility. In Chapter 20, we examined the formation of the liquidity spiral in detail; here we need to shift our analytical perspective and re-examine this positive-feedback loop from the dimension of volatility.

In the traditional finance framework, volatility is usually treated as an exogenous risk parameter. In crypto markets, however, volatility is to a large extent an endogenous variable. When the market experiences a slight price movement because of some initial random shock, this movement is captured by market makers' risk-management models. To control inventory risk and adverse-selection costs, the rational response of market makers is to widen the bid-ask spread and reduce the number of resting orders at each price level. This collective retreat by market makers directly causes the order book to thin.

Once the order book thins, the market's capacity to absorb order shocks declines significantly. At this point, even normal-sized trade flow triggers larger price slippage than usual on the thin order book. These price jumps produced by slippage statistically manifest as a further rise in realized volatility. This rise then feeds back to market makers' risk models as a new input signal, prompting them to cut liquidity supply further. This loop—in which rising volatility causes liquidity to dry up, and drying-up liquidity in turn produces greater volatility—is the core mechanism of self-fulfilling non-informational volatility [23][24].

In this positive-feedback loop, the liquidation mechanism unique to perpetual futures acts as an accelerator. When the price slides violently on a thin order book, it can easily touch the liquidation lines of highly leveraged positions. The liquidation engine is forced to throw unconditional market orders onto the market, and these forced orders continue to penetrate multiple price levels on an order book that has already lost the protection of market makers, triggering an even larger price decline. In the market turmoil of October 2025, this volatility spiral, woven together by liquidity exhaustion and cascading liquidations, triggered more than $19 billion (about $19.37 billion) in visible forced-liquidation notional within roughly 24 hours, constituting the largest single-day deleveraging event in history (the most violent shock was concentrated within a single trading day, while the full deleveraging window lasted about 36 hours; see Section 22.6.4 for details) [25]. In this process, the overwhelming majority of the price decline was not a reflection of deteriorating fundamentals but a direct reflection of a liquidity crisis in the form of price movement.

22.3.3 Time structure

Crypto markets trade around the clock without interruption, but liquidity depth and the capacity to absorb shocks are not equal at every moment. In fact, the liquidity volatility of crypto assets has a highly regular time structure. This time structure is shaped mainly by the alternating activity of traders in different global time zones, the operating cycles of algorithmic trading programs, and the specific settlement mechanisms of derivatives markets.

Empirical research shows that trading activity and liquidity depth in crypto markets exhibit a pronounced bell-shaped curve over the course of a day. Around 16:00–17:00 UTC—the overlap between the tail of the European session and the early New York session—market liquidity reaches its daily peak. At this time, the world's major institutional investors and high-frequency trading firms are all actively participating in the market, order-book depth reaches its highest level of the day, the market's capacity to absorb large orders is strongest, and spurious volatility caused by insufficient liquidity falls to its lowest point [26].

During the midnight-to-early-morning hours in UTC, and especially late at night in the Asian session, market liquidity falls into a deep trough. Within this liquidity-trough window, the activity of mainstream European and American market makers drops sharply, and the depth of the order book may be only one-third of its peak level or even less. It is precisely this periodic liquidity exhaustion that makes crypto markets exceptionally fragile at particular times. Historical data reveal a striking phenomenon: most historical crypto-market flash crashes have occurred in this late-night window when liquidity depth is at its daily low—though, as participation by Asian markets (particularly South Korea and Hong Kong) increases, the time-zone distribution of flash crashes is evolving; for example, the crypto flash crash triggered by the unwinding of the yen carry trade in August 2024 occurred during the active hours of the Asian session. At these moments, even a sell order of no extreme size is enough to form a significant price gap on an order book that lacks depth, producing extremely high instantaneous volatility.

Beyond this intraday pattern created by time-zone alternation, derivatives settlement mechanisms also manufacture predictable price pulses at specific moments: in the run-up to a funding-rate settlement or the expiry of a dated futures contract, arbitrageurs and directional traders reposition en masse to secure a favorable settlement price or to avoid a funding payment, triggering sawtooth-shaped volatility around settlement. Such pulses, in essence, are mechanism-induced volatility rather than liquidity volatility, and their full generating mechanism is detailed in Section 22.4.2. Figure 22-6 uses a two-panel chart to present the empirical features of the above intraday periodic pattern.

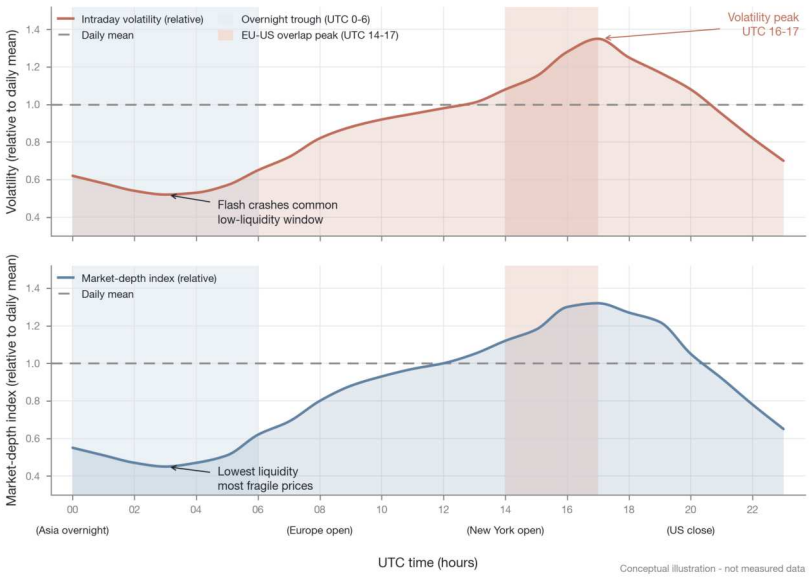


Figure 22-6. Intraday liquidity and volatility patterns in the crypto market (intraday-pattern illustration: representative hourly patterns of realized volatility and order-book depth, with the pattern referencing Brauneis et al. (2025) [26]; not recomputed hour by hour from measured data)

As shown in Figure 22-6, realized volatility peaks at UTC 16:00–17:00 (the European-American session overlap), while order-book depth falls to its lowest from UTC midnight to early morning. This key asymmetry—that the volatility peak and the liquidity trough do not coincide—means that a unit shock in the late-night hours causes a larger price deviation, making this window a high-incidence period for flash-crash events.

22.3.4 The asset dimension gradient

The degree of liquidity volatility not only exhibits regular fluctuations along the time axis but also varies enormously and structurally across assets. If we rank the entire crypto market by market capitalization and liquidity, we can clearly observe a liquidity gradient extending from absolutely mainstream assets such as Bitcoin all the way to tiny-cap altcoins. This liquidity gradient corresponds to a volatility gradient.

As the market's foundational assets, Bitcoin and Ethereum possess the deepest liquidity pools, the most active market-maker communities, and the most complete derivatives ecosystems. On the order books of these assets, orders on the order of tens of millions of dollars often produce only tiny price movements. In the aggregate volatility of main-

stream assets, therefore, informational volatility driven by fundamental information, macroeconomic expectations, and overall market sentiment dominates, while spurious volatility caused purely by insufficient liquidity accounts for a relatively small share.

As we move down the market-cap ladder into the range of mid-cap and even tiny-cap altcoins, however, the situation reverses completely. The order books of these assets are usually noticeably thin, with wide bid-ask spreads and no professional market makers providing continuous liquidity support. In these markets, insufficient liquidity is no longer an occasional extreme event but a normalized market feature. These small-cap tokens therefore exhibit volatility levels far higher than those of mainstream assets.

More critically, in these high-volatility altcoins, spurious volatility dominates absolutely. Because liquidity is extremely scarce, any buy or sell order of even modest size—regardless of whether it carries a genuine informational advantage—produces a large price impact on the order book, triggering sharp price jumps or declines. This extreme liquidity sensitivity fills the price trajectory of small-cap tokens with noise and random-walk characteristics. For these assets, their high volatility is to a large extent a risk compensation for their poor liquidity conditions—or, put differently, a premium the market demands for bearing their illiquidity friction [27]. Figure 22-7 shows the correspondence between the liquidity gradient and the volatility gradient from high-cap to low-cap assets.

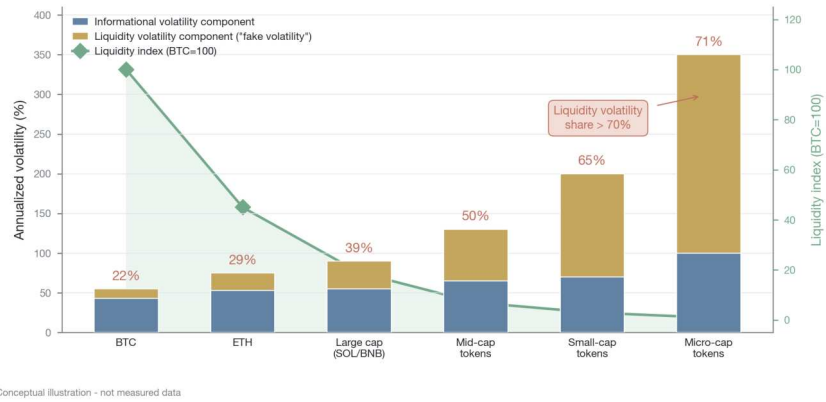


Figure 22-7. Asset liquidity gradient and volatility gradient (gradient values are a representative illustration, not actually pulled from CoinGecko: market capitalization and the Amihud illiquidity ratio [28] reference the grouping of the top 50 crypto assets by CoinGecko market-cap ranking; the informational/liquidity volatility split shown in the figure is an illustrative estimate based on the four-source framework, not produced directly from these data)

As shown in Figure 22-7, aggregate volatility rises steeply as assets move from mainstream coins to mid- and low-cap altcoins; its composition (split illustratively in the figure, conveying only a qualitative gradient rather than precise shares) reverses accordingly: mainstream assets are dominated by informational volatility, whereas the share of liquidity volatility for low-cap assets expands dramatically, even exceeding informational volatility. This means that the economic meaning carried by the volatility of different market-cap tiers differs fundamentally: the volatility of high-cap assets mainly reflects information, while the volatility of low-cap assets mainly reflects friction.

22.3.5 Measurement methods

To handle liquidity volatility effectively in quantitative trading and risk management, we need a rigorous measurement system that isolates this spurious volatility from aggregate price movements. In the market-microstructure literature, scholars have developed a variety of metrics to capture and quantify this volatility arising from liquidity friction.

The most direct measurement method is to quantify the transitory price impact. This method tracks the high-frequency price trajectory after a large trade and decomposes the initial price deviation into a permanent component and a transitory component. The permanent component represents the information content contained in the trade, whereas the transitory component—the portion of the price that mean-reverts after the shock—precisely measures the magnitude of the spurious volatility caused by insufficient liquidity. By computing the proportion of this transitory shock in the aggregate price movement, and the time required for the price to complete its mean reversion (that is, the reversion rate of volatility), we can clearly depict the liquidity resilience of a specific asset in a specific period.

Over more macro daily or longer horizons, the Amihud illiquidity ratio is a widely adopted classic metric. This ratio measures the magnitude of price movement induced per unit of trading volume by computing the ratio of an asset's absolute return to its dollar trading volume [28]. In crypto markets, a high Amihud ratio means that very little trading volume can trigger violent price movement, which strongly suggests that the asset's volatility is driven mainly by scarce liquidity. Empirical research confirms that, after controlling for other risk factors, the expected returns of crypto assets are significantly positively correlated with their Amihud illiquidity ratio, indicating that the market does indeed demand risk compensation for bearing this liquidity-induced volatility [29].

In addition, from the micro perspective of market makers, the realized spread is another key metric for measuring liquidity volatility. Unlike the quoted spread displayed on screen, the realized spread measures the profit that a market maker can actually capture after providing liquidity and holding the position for a period. In an environment of violent liquidity volatility, because of frequent price jumps and increased adverse-selection risk, market makers often find that their realized spread is far below the quoted spread, or even negative. This compression of the realized spread is the most direct signal that market makers perceive liquidity-volatility risk, and it is also the micro trigger that prompts them to withdraw liquidity and thereby set off the volatility positive-feedback loop.

With these fine-grained measurement tools, we can isolate, from aggregate price movements, the spurious volatility manufactured purely by market-structure friction. This identification capability is a precondition for building robust trading strategies and risk-management models, and it helps distinguish information-driven permanent price adjustments from temporary price shocks caused by insufficient liquidity. Liquidity-measurement methods tailored to the high-frequency crypto environment have continued to evolve in recent years; methods such as high-frequency estimation of Kyle's lambda from tick-by-tick trade data and limit-order-book resilience metrics can capture the dynamic features of liquidity volatility on shorter time scales, providing finer

tools for the empirical application of the four-source framework.

22.4 Mechanism-induced volatility

Price volatility in perpetual futures markets stems not only from updates to fundamental information or temporary shortages of liquidity but, more profoundly, from the driving force of its underlying institutional design. As previewed in Section 22.1.3, mechanism-induced volatility is the price movement created by the market rules themselves and unrelated to new information; its distinctive sources—the leveraged liquidation engine, the funding-rate settlement cycle, and the mark-price smoothing mechanism—while maintaining system solvency and anchoring to the spot price, inevitably inject additional noise into the market, making it the core source of the "excess volatility" of perpetual futures. This section focuses on institutional design at the level of algorithms and rules; it does not cover volatility triggered by operational-risk events such as exchange outages, security attacks, or technical failures. Although such events can also manufacture extreme price movements unrelated to fundamentals, their suddenness and unpredictability make them better discussed within a risk-management framework than within a volatility-decomposition one.

22.4.1 Liquidation cascades

A liquidation cascade is not merely an amplifier of existing volatility; it is itself an independent source of volatility. When the market is relatively calm, a large directional trade may push the price only a slight initial distance—a movement that belongs to informational or liquidity volatility. If this tiny price movement touches the liquidation lines of highly leveraged positions, however, the liquidation engine automatically takes over these positions and enters the forced-liquidation process. Mainstream CEXs typically use a multi-tier liquidation mechanism rather than a simple market dump: the first tier attempts to liquidate via limit orders backed by the insurance fund, and only when limit orders cannot fill and the insurance fund is exhausted is auto-deleveraging (ADL) triggered. This multi-tier design effectively mitigates the market impact of liquidation in normal markets, but under extreme conditions it can still degrade into quasi-market forced execution. These liquidation orders exert further downward pressure on an already-weak order book, which in turn triggers deeper liquidations, forming a positive-feedback loop. In this process, the price moves violently without any new information arriving, and this additional decline is manufactured purely by the liquidation mechanism.

Liquidation-driven volatility exhibits a highly characteristic "V-shaped" or spike pattern in the price series, in sharp contrast to the "L-shaped" permanent adjustment of informational volatility. During a liquidation cascade, the price usually overshoots significantly and then, once the liquidation pressure is exhausted, quickly rebounds toward a level closer to fundamentals. This shape reflects the cost-insensitive nature of liquidation orders and their extreme consumption of short-term liquidity. In traditional futures markets, because leverage ratios are strictly limited, there is a margin-call buffer period, and price-limit (circuit-breaker) mechanisms provide protection, the scale

and frequency of liquidation cascades are greatly compressed. By contrast, crypto perpetual futures markets generally allow extremely high leverage multiples and enforce instant forced liquidation, which makes liquidation cascades a normalized market phenomenon. Figure 22-8 compares the standardized price paths of liquidation-driven and information-driven events, presenting the fundamental difference in the post-event trajectories of the two types of volatility.

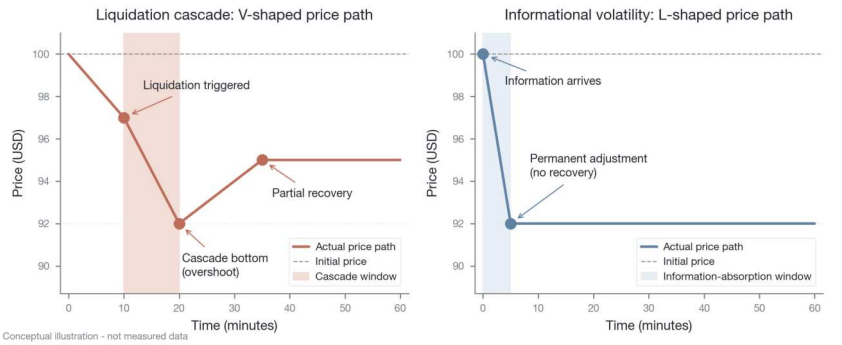


Figure 22-8. Comparison of the liquidation-driven V-shaped price path and the information-driven L-shaped price path (mechanism-comparison illustration: representative standardized price paths in the ± 30 minutes around liquidation events and information events; not measured Coinglass data)

As shown in Figure 22-8, the "V-shaped" path of a liquidation cascade dives rapidly under the shock to an extreme position far below equilibrium, then partially rebounds toward a level closer to fundamentals; the "L-shaped" path of an informational shock, by contrast, stabilizes over the long run after the jump and does not undergo systematic mean reversion. The core of this difference is that liquidation orders contain no information about an asset's true value; their price deviation is purely a product of liquidity consumption and must therefore be corrected once liquidity is restored.

Extreme market events in history have repeatedly verified the destructive power of liquidation cascades as an independent source of volatility. In the collapse of the Terra ecosystem in May 2022, the depegging of the algorithmic stablecoin UST triggered panic across the entire crypto market. As the price of the Luna token plummeted, a large number of leveraged positions that relied on Luna as collateral were liquidated, and these liquidation orders further accelerated the price collapse, ultimately driving Luna's market capitalization to near zero within just a few days [30]. Even in periods of relatively stable macro conditions, crypto markets frequently experience violent volatility triggered by liquidation. For example, in the deleveraging event of October 2025 (whose scale and window are detailed in Sections 22.3.2 and 22.6.4), large-scale visible forced liquidations drove the prices of multiple assets into a deep decline before they rapidly rebounded [25]. These cases show that when leveraged positions are highly concentrated within a particular price range, even a tiny initial trigger is enough to set off a massive liquidation cascade, thereby manufacturing extreme volatility completely disconnected from fundamentals. In assessing the *eliminability* of mechanism-induced volatility, however, the fundamental purpose of the liquidation mechanism—maintaining system solvency—must be kept in view. Reducing the market impact of liquidation (for example, by adopting gradual liquidation or auction-based liquidation

[31]) comes at the cost of increased liquidation latency, which under extreme conditions may leave the protocol with bad debt. Mechanism-induced volatility therefore cannot be entirely eliminated, and the goal of institutional optimization is to find the Pareto optimum between minimizing market impact and maximizing solvency.

22.4.2 Funding-rate settlement pulses

The funding-rate mechanism is a core design by which perpetual futures maintain their anchor to the spot price, but its periodic settlement process also injects predictable periodic volatility into the market. To push the perpetual futures price back toward the spot price, longs and shorts must periodically exchange the funding rate. Under a typical 8-hour settlement cycle, when the funding rate is positive and high, holders of long positions face pressure to pay a high fee at the settlement moment. To avoid this cost, a large number of long traders tend to close or reduce positions before settlement, thereby forming systematic selling pressure ahead of the settlement window. That said, the magnitude of the settlement pulse depends not only on the level of the funding rate but also on the degree of consensus among market participants about the future price direction. In a strongly one-sided market, even if the rate becomes extreme, the paying side may prefer to pay the rate rather than close out (because it expects the price to keep moving in its favor), in which case the settlement pulse is instead smaller. This concentrated trading, driven by the desire to avoid funding payments, produces a predictable downward price movement when the market direction is uncertain.

As the settlement moment approaches, this directional pressure gradually accumulates, causing the contract price to deviate from the spot price. Once settlement is complete, however, the motive to avoid the funding payment disappears, the selling pressure dissipates, and the price often rebounds quickly to the pre-settlement level. This price volatility triggered by funding-rate settlement contains no fundamental information whatsoever; it is purely pulse-like noise produced by institutional design. In periods when the funding rate is at extreme levels, the magnitude of this periodic pulse is significantly amplified, becoming a key force dominating the short-term market microstructure. Figure 22-9 uses a time-series chart to present the periodic pulse pattern of basis and volatility around funding-rate settlement.

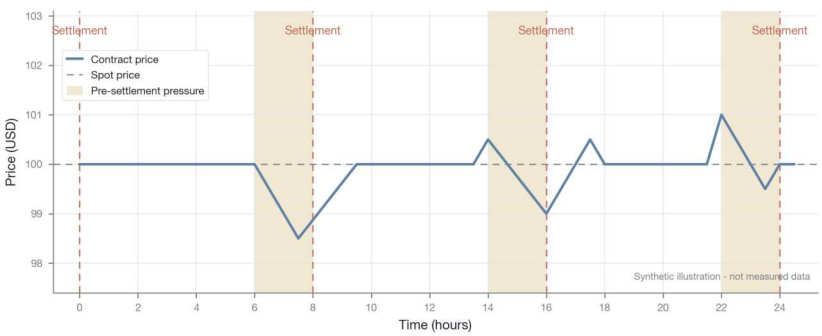


Figure 22-9. Volatility pulses under the 8-hour funding-rate settlement cycle (periodic-mechanism illustration: depicting the representative pulse pattern of basis and realized volatility in the ±1 hour around each 8-hour settlement; a synthetic curve, not measured Binance data)

As shown in Figure 22-9, as each 8-hour settlement moment approaches, the concentrated position-closing by the high-rate paying side generates regular downward pressure, manifested as a systematic rise in basis and realized volatility before settlement and a return to normal within a few minutes after settlement. This highly predictable periodic pulse demonstrates the existence of the funding-rate settlement mechanism as an independent source of volatility—it periodically injects noise into the market without any change in fundamental information.

Cross-platform comparison across exchanges with different settlement frequencies can more clearly isolate the contribution of the funding-rate mechanism to volatility. On platforms that use the traditional 8-hour settlement cycle, price volatility around settlement exhibits a clear clustering feature. Some emerging decentralized exchanges (DEXs), or platforms that adopt continuous-settlement mechanisms, effectively smooth traders' avoidance behavior by dispersing funding payments into every block or into a finer time granularity. Theoretical models show that the funding rate is an algorithmic feedback rule rather than a passive transfer, and that its design parameters—such as settlement frequency and rate caps—affect the mean reversion of the basis and its tail risk [32]. Continuous settlement fundamentally eliminates the concentrated-trading motive brought by discrete settlement, thereby significantly reducing this type of mechanism-induced volatility. But continuous settlement may also introduce new problems in an on-chain environment: the tiny rate exchange in each block increases the opportunity for maximal extractable value (MEV) extraction and may generate mechanism-induced noise of a different form. Even so, funding-rate settlement pulses remain the source of mechanism-induced volatility with the greatest room for institutional optimization.

22.4.3 The lag effect of the mark-price mechanism

To guard against erroneous liquidations triggered by market manipulation and anomalous trades, perpetual futures platforms generally introduce a mark-price mechanism as the benchmark for computing unrealized profit and loss and triggering forced liquidation. The calculation of the mark price usually involves two steps. First, a price index is aggregated as a weighted composite of component prices from multiple external spot exchanges, with single-exchange prices that deviate too far from the median excluded. Second, a time smoothing such as an exponential moving average (EMA) is applied on top of the price index, together with a reasonable basis-adjustment term, so that the mark price can smoothly track the theoretical fair value of the perpetual futures in normal markets (see Chapter 14 for the full governance implications of the mark-price mechanism). This two-layer design makes the volatility of the mark price lower than that of the last traded price, effectively filtering short-term noise and reducing unnecessary liquidation events. In extreme conditions where the market moves rapidly in one direction, however, the exponential-moving-average nature of the mark price can produce lag risk.

When the true market price changes sharply, the smoothing algorithm causes the mark price to respond noticeably more slowly than the last traded price. This lag protects, in the short run, some positions that should have been liquidated, but it also allows systemic risk to build up unnoticed. Over time, when the mark price finally catches up

with the true market price, those loss-making leveraged positions accumulated during the lag period face concentrated forced-liquidation risk. This dynamic of accumulation and release effectively compresses liquidation events that should have occurred in a dispersed fashion into an extremely short time window, thereby producing a larger and more forceful liquidation pulse when the lag ends.

Moreover, the mark-price mechanism itself constitutes a transmission channel for cross-market volatility. Because the mark price synthesizes data from multiple external exchanges, when one of the spot exchanges experiences an extreme price deviation due to liquidity exhaustion or a technical failure, this deviation is partially passed on to the perpetual futures market through the mark-price formula. If this deviation is enough to trigger the perpetual futures liquidation engine, it will manufacture, in the derivatives market, a spurious liquidation unrelated to that market's own fundamentals and liquidity conditions. Notably, the lag effect of the mark price takes on different forms in decentralized perpetual futures exchanges. The design of DEX perpetual futures has already diverged significantly: the mark price of oracle-based DEXs (such as GMX) relies entirely on oracle price feeds, traders act as price takers, and their lag risk comes from the oracle's update frequency, so that even a tiny deviation in the mark price can trigger a significant chain reaction [33]; the mark-price mechanism of order-book DEXs (such as Hyperliquid and dYdX v4) is closer to that of a CEX, but their lag risk comes mainly from sequencer latency rather than oracle latency. By comparison, a CEX's mark price is based on a weighted index of multiple spot exchanges plus a reasonable basis adjustment, and its degree of lag varies with the setting of the smoothing parameters.

22.4.4 Attribution of volatility differences

To quantify the scale of mechanism-induced volatility accurately, one can compare the volatility of the same underlying asset under different market structures. Take Bitcoin as an example: its spot market, dated futures market, and perpetual futures market share the same macro information inputs and fundamental expectations, yet the volatility characteristics of the three differ significantly. The volatility of the spot market consists mainly of informational volatility and a small amount of liquidity volatility, and can be regarded as the benchmark for an asset's intrinsic volatility. The dated futures market adds, on this basis, volatility from basis changes and a small amount of liquidation volatility under limited leverage. The perpetual futures market not only contains all the volatility components of the spot market but also superimposes more severe liquidity volatility and a large amount of mechanism-induced volatility.

The literature and market observation consistently show that the realized volatility of perpetual futures tends to be higher than that of contemporaneous spot and dated futures. This volatility premium—the excess volatility that Section 22.7 will analyze systematically—is caused mainly by the unique institutional design of perpetual futures, and its specific source attribution is detailed in Section 22.7.2. Figure 22-10 presents the source of this excess volatility by comparing the volatility levels and their composition for the same underlying asset under three different market structures.

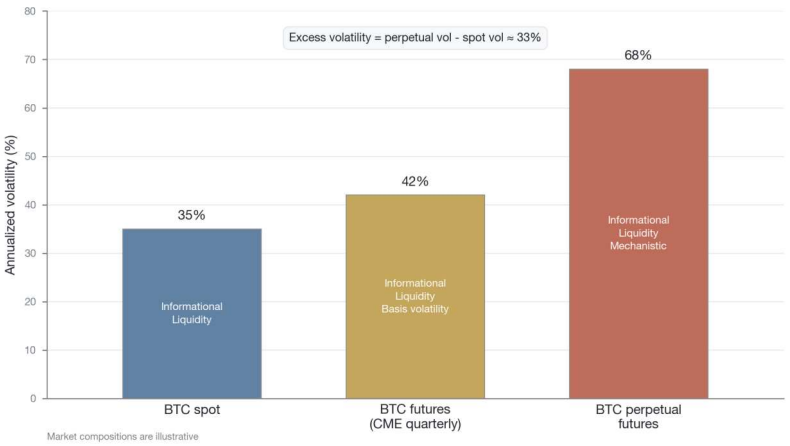


Figure 22-10. Comparison of the volatility composition of the same underlying asset in the spot, dated futures, and perpetual futures markets (the RV bar heights for each market are a representative illustration, not recomputed from measured data, referencing the contemporaneous daily realized volatility of BTC on Coinbase spot, CME dated futures, and Binance perpetual futures, 2023–2025; the four-source composition split within each market is an illustrative estimate based on this chapter's framework)

As shown in Figure 22-10, volatility rises successively from spot (dominated by the informational component, the lowest benchmark), to dated futures (adding limited basis and a small amount of liquidation volatility), to perpetual futures (superimposing the mechanism-induced volatility of liquidation cascades and funding-rate settlement). Because the three markets share the same underlying asset and information set, this systematic difference can only be attributed to institutional design, providing natural-experiment-style evidence for quantifying the net contribution of mechanism-induced volatility.

This cross-market comparison is not only of theoretical significance; it also provides a natural-experiment method for isolating and measuring mechanism-induced volatility. When the markets face the same information shock, the perpetual futures market often exhibits a deeper decline and higher short-term volatility, a difference that cannot be explained by differences in fundamental information but can only be attributed to institutional friction. A horizontal comparison across platforms with different liquidation mechanisms and settlement frequencies can further confirm the significant role of institutional design in generating non-informational volatility. For example, platforms that adopt gradual liquidation rather than cliff-edge liquidation exhibit a markedly weaker volatility-amplification effect under extreme conditions.

22.4.5 Measuring mechanism-induced volatility

The core feature of mechanism-induced volatility is its high degree of rule dependence, which makes it largely predictable. To measure this portion of volatility accurately in empirical research, researchers usually adopt the event-window method. In the liquidation event-window method, isolating the excess volatility in the few minutes before and after a large-scale liquidation makes it possible to strip out the price movement caused

purely by the liquidation engine's dumping. Similarly, funding-settlement event studies focus on the short window around the settlement moment, computing the volatility pulse that deviates from the norm within that period and classifying it as mechanism-induced volatility.

Beyond the event-window method, building a no-liquidation counterfactual model is another effective measurement approach. By simulating the theoretical price path in the absence of a forced-liquidation mechanism and comparing it with the actually observed price path, the difference between the two represents the net contribution of the liquidation mechanism to volatility. Together, these measurement methods reveal one fact: under extreme market conditions, mechanism-induced volatility not only accounts for a significant share of total volatility but may even become the decisive force dominating short-term price movements.

The predictability of mechanism-induced volatility has given rise to trading strategies that arbitrage institutional friction specifically. The certainty of funding-rate settlement times and the visibility of liquidation cascades on the order book allow some high-frequency traders to position themselves in advance. They profit from institutionally manufactured noise by providing liquidity before settlement to absorb the funding-avoidance selling pressure, or by buying against the trend when a liquidation cascade causes a price overshoot. In actual execution, however, this arbitrage faces significant constraints: during a liquidation cascade the exchange may trigger overload protection or restrict new position openings; the actual fill price of a contrarian buy may be far worse than expected because of slippage; and if the liquidation pressure does not subside as expected but continues to intensify, the losses on the contrarian position may far exceed the initial budget. Moreover, while this arbitrage dampens some mechanism-induced volatility, it may also, in extreme cases, accelerate the triggering of liquidations. From the perspective of market design, clearly distinguishing and quantifying mechanism-induced volatility carries important guidance for optimizing the liquidation logic of perpetual futures, improving the frequency of funding-rate settlement, and enhancing the robustness of the mark price.

22.5 Transmitted volatility

The crypto-asset perpetual futures market is not a closed, independent system; its volatility stems not only from internal microstructure or supply-demand imbalances but also, and more so, from the continuous influx of external shocks. As previewed in Section 22.1.3, transmitted volatility is the spillover effect that spreads in through cross-market arbitrage and cross-margin mechanisms; in essence it is a form of second-hand volatility, highly contagious and strongly cross-correlated, with that correlation often approaching one when a crisis erupts. This volatility includes both inter-asset spillover within the crypto market and the macro transmission of fundamental changes in traditional financial markets.

22.5.1 Spillover effects within the crypto market

Within the crypto market, the transmission of volatility exhibits a highly structured and hierarchical network feature. When the market encounters a sudden shock, volatility is not evenly dispersed across all assets but follows specific transmission paths. As the crypto asset with the largest market capitalization and deepest liquidity, Bitcoin is usually the core hub of this volatility-transmission network. Research shows that Bitcoin is, in most periods, a major net transmitter of volatility connectedness, passing its own volatility shocks to Ethereum and other smaller-cap crypto assets through trading pairs and arbitrage mechanisms [34]. This transmission is especially rapid in the perpetual futures market, often completing cross-asset volatility co-movement within minutes. Under a unified-margin system, moreover, the transmission path is not one-directional. When a trader holds altcoin perpetual futures positions with BTC as collateral, the liquidation of the altcoin triggers the forced sale of the BTC collateral, thereby transmitting the altcoin's volatility back to BTC. This reverse-transmission path means that BTC can also become a net receiver of volatility in extreme crises, challenging its positioning as a pure "net transmitter."

Ethereum acts as a relay node in this network. As DeFi and the non-fungible token ecosystem have flourished, Ethereum's role in volatility transmission has grown, and in certain periods it has even surpassed the spillover of Bitcoin [35]. From Bitcoin to Ethereum, and from Ethereum on to numerous altcoins, this hierarchical transmission constitutes the basic skeleton of internal volatility spillover in the crypto market. Figure 22-11 uses a directed network diagram to visualize the volatility-spillover relationships among major crypto assets.

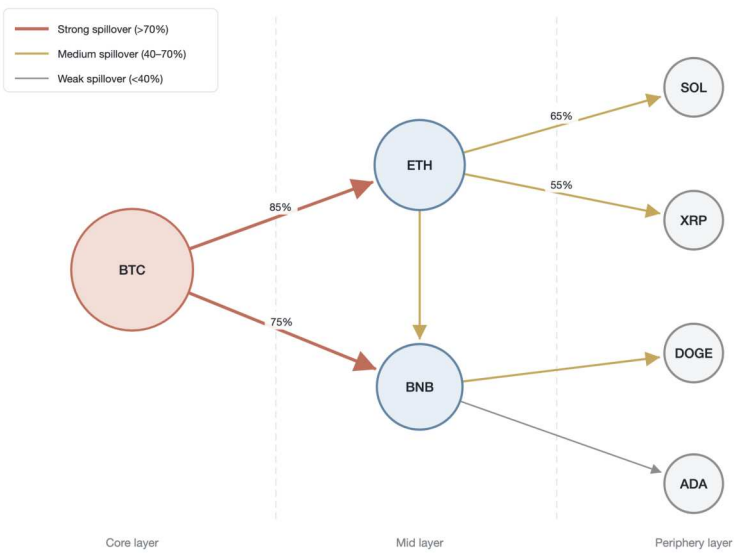


Figure 22-11. The crypto-market volatility-spillover network (the network is a conceptual illustration, not a computed DY matrix: the method references the DY spillover-network framework of Yi et al. (2018) [34] and Diebold & Yilmaz (2012) [36], depicting the representative spillover structure among BTC, ETH, and the top 10 altcoins by market capitalization)

As shown in Figure 22-11, Bitcoin sits at the core hub, and the thickness of the arrows it emits (spillover intensity) is the greatest, confirming its status as a major net transmitter of volatility (note that Yi et al. (2018) [34] point out that although Bitcoin is a major net transmitter, it is not the sole dominant one, and some small-cap coins can also become significant net transmitters in particular periods); Ethereum is a secondary hub, while the peripheral mid- and low-cap assets mainly play the role of receivers, with most of their volatility changes traceable to spillover from the core assets. This star-hierarchical topology shows that volatility transmission is not evenly diffused but has clear directionality and hierarchy.

In calm market periods, the connection strength of this transmission network is relatively moderate, and each asset can still maintain some degree of independent pricing. When a crisis occurs or an extreme volatility event appears, however, the transmission strength increases sharply and nonlinearly. Empirical analysis confirms that under a high-volatility regime, the spillover effect among crypto assets is significantly enhanced, exhibiting strong regime dependence [37]. In a panic sell-off, the high-leverage feature of the perpetual futures market accelerates the cascading liquidation of margin, causing the volatility of all crypto assets to spike in sync; a once-diversified portfolio instantly loses its protective value, and the entire market moves in lockstep.

22.5.2 BTC dominance and beta volatility

In assessing the relative volatility of crypto assets, the market often borrows the beta-coefficient concept from traditional finance. Altcoins typically exhibit extremely high-beta characteristics relative to major assets, meaning they are exceptionally sensitive to price changes in the core assets. When Bitcoin's volatility rises moderately, altcoins often experience much more violent price swings—that is, a small rise in Bitcoin's volatility is significantly amplified in altcoins.

This high-beta feature is not fixed; it is significantly time-varying and deeply influenced by the market cycle. In the exuberant phase of a bull market, as liquidity is ample and risk appetite runs high, the beta coefficients of altcoins tend to inflate, and investors tend to hold high-beta assets to capture excess returns. Conversely, in the capital-flight phase of a bear market, although altcoins still decline significantly, the absolute value of their beta coefficient may change structurally because of liquidity exhaustion and trading stagnation.

The widespread use of perpetual futures has greatly amplified this transmission effect. Because the spot liquidity of altcoins is usually poor, the perpetual futures market becomes the primary venue for speculators to express their views. When Bitcoin faces a macro shock or an internal structural sell-off, arbitrageurs and market makers rapidly adjust the risk exposure of the entire crypto-asset matrix. Highly leveraged positions in altcoin perpetual futures are extremely fragile in the face of such transmitted shocks, and even a slight movement in the benchmark asset is enough to trigger large-scale forced liquidation of altcoin contracts, producing transmitted volatility several times the size of the initial shock.

22.5.3 Arbitrage transmission between spot and perpetual

The price linkage between perpetual futures and the spot market is another core path of volatility transmission. These two markets are tightly linked through the funding-rate mechanism and the uninterrupted operation of arbitrageurs. In an ideal state, the spot market and the perpetual futures market should maintain price consistency, but in reality microstructural differences and the asymmetry of information inflow often cause a price gap—the basis—between them.

In the process of closing the basis, arbitrageurs inadvertently serve as an intermediary channel for volatility transmission. When the spot market experiences violent volatility because of external news or a large order, the rapid movement of the spot price widens the basis against the perpetual futures. At this point, statistical arbitrageurs and basis traders quickly step in, buying in the undervalued market and selling in the overvalued one. Although this cross-market high-frequency trading helps price discovery, it also effectively replicates and transmits the spot market's volatility to the perpetual futures market. Academic research indicates that the spot market is usually the main transmission source of volatility spillover and has a significant causal driving relationship with the derivatives market [38].

This transmission, however, is not always instantaneous and frictionless. Under extreme market conditions, structural arbitrage constraints cause transmission to lag. Research by the Bank for International Settlements (BIS) shows that the basis in crypto markets can sometimes reach extremely high levels, with annualized returns even exceeding 40%, reflecting that during extreme volatility the deployment of arbitrage capital is severely constrained by multiple execution frictions: blockchain confirmation delays (on-chain BTC confirmation takes 10 to 60 minutes) prevent cross-exchange fund transfers from completing instantly; exchanges freezing or restricting withdrawals under extreme conditions further block capital rebalancing; and margin calls sharply reduce the capital available to arbitrageurs [39]. When the arbitrage pipeline is blocked, the basis widens abnormally, and the spot market's volatility may take longer to be fully reflected in the perpetual futures. In certain information-driven events, however, such as a macro data release or the finalization of a regulatory policy, the perpetual futures market, with its higher leverage and lower trading costs, sometimes leads the spot market in completing price discovery and becomes the origin of volatility [40].

22.5.4 Macro transmission from traditional finance

The crypto market was once viewed as a safe-haven asset or "digital gold" independent of the traditional financial system. Several forces have since eroded this isolation, however: institutional investors allocating directly to crypto assets, the common global-liquidity factor created by the Federal Reserve's quantitative easing, and the rise of retail cross-asset speculation in a zero-interest-rate environment. The correlation between the crypto market and traditional financial assets has since changed markedly, and macroeconomic volatility can now be transmitted efficiently to the crypto perpetual futures market.

Before 2020, the correlation between Bitcoin and traditional equity assets such as the S&P 500 index was extremely low, and in some periods even negative. The outbreak

of COVID-19 became a historic turning point, however. A report by the International Monetary Fund notes that since the outbreak of the pandemic, the correlation between the volatility of the Bitcoin price and the volatility of equity indices has increased significantly, with Bitcoin's spillover effect accounting for 14% to 18% of the change in global equity-market volatility [41]. Figure 22-12 uses a time-series chart to show the long-term evolution of the dynamic correlation between Bitcoin and the S&P 500 index.

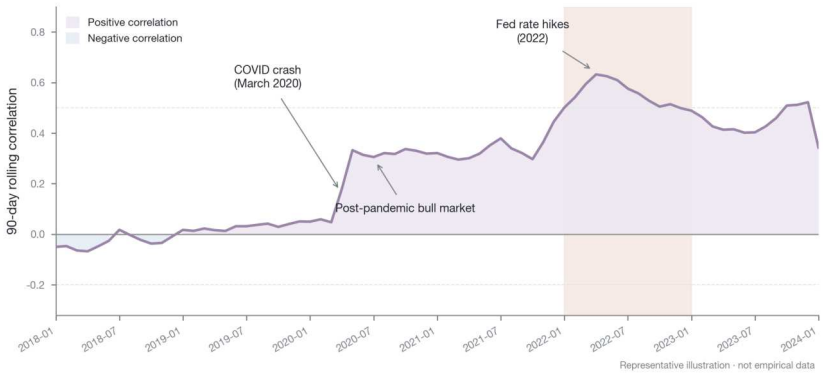


Figure 22-12. Time-varying correlation between BTC and the S&P 500 (representative illustration: the 90-day rolling correlation between BTC and the S&P 500 (2018–2024), consistent with the empirical findings of Iyer (2022) [41] but not actually computed)

As shown in Figure 22-12, before 2020 the rolling correlation coefficient between Bitcoin and the S&P 500 was for a long time close to zero or even occasionally negative, supporting the "digital gold" narrative; after the COVID-19 outbreak in 2020 it jumped structurally, once climbing above 0.5 during the macro tightening of 2022. This indicates that, as institutions have become deeply involved, crypto assets have been incorporated into the pricing system for global macro risk assets, so that volatility shocks in traditional markets are effectively transmitted to the perpetual futures market.

The core driving force of this macro transmission is the cyclical shift in global monetary policy. When the Federal Reserve implements extremely loose monetary policy, the excess liquidity simultaneously pushes up the prices of traditional risk assets and crypto assets. When monetary policy turns to tightening, the rise in the risk-free rate rapidly compresses the valuations of all risk assets. Research finds that when the shadow policy rate turns negative (that is, monetary policy is extremely loose), the spillover effect of monetary policy on the crypto market is especially large [42].

This transmission forms a clear chain: the Federal Reserve's policy decisions first shock the U.S. Treasury and equity markets; traditional financial institutions then, to balance the risk of their overall portfolios, adjust their crypto-asset exposure; this adjustment is transmitted through the Bitcoin spot market to Bitcoin perpetual futures, and finally spills over through the internal network to the entire altcoin perpetual futures market. During typical macro-tightening crises such as that of 2022, the correlation between crypto assets and traditional risk assets rose sharply toward one, and the traditional portfolio-diversification strategy failed significantly in the face of this systemic transmitted volatility.

22.5.5 Measuring transmitted volatility

To quantify and monitor this complex transmitted volatility precisely, financial econometrics offers a set of advanced analytical tools. The most widely adopted is the spillover-index model proposed by Diebold and Yilmaz. Based on a VAR system and the generalized forecast-error variance decomposition, this model can strip out what proportion of an asset's volatility is caused by shocks from other assets in the system [36].

By constructing a dynamic spillover-index matrix, researchers can clearly compute whether each market is a net receiver or a net transmitter of volatility. In the crypto perpetual futures market, this matrix usually exhibits a highly asymmetric structure. The diagonal elements represent an asset's own idiosyncratic volatility, while all the off-diagonal elements quantify the intensity of cross-market spillover. Figure 22-13 uses a heat map to present the directional-spillover matrix among the perpetual futures of major crypto assets.

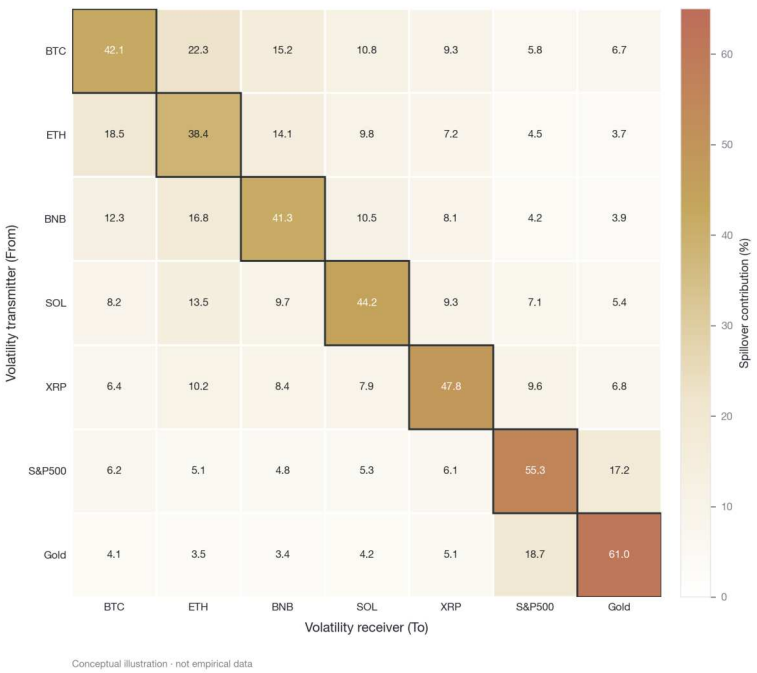


Figure 22-13. DY spillover-index heat map (conceptual illustration, not a computed matrix: based on the framework of Diebold & Yilmaz (2012) [36], depicting the directional-spillover structure among major crypto assets such as BTC, ETH, BNB, SOL, and XRP and the S&P 500 and gold)

As shown in Figure 22-13, the matrix exhibits a highly asymmetric structure: Bitcoin's spillover to other assets (the corresponding row) is significantly higher than the spillover it receives (the corresponding column), further confirming its status as a major net transmitter of volatility. The diagonal elements of small-cap assets (the proportion of their own idiosyncratic volatility explained) tend to be low, meaning their volatility is driven to a large extent by external contagion rather than by changes in their own fundamentals.

Beyond the spillover index, multivariate GARCH models, especially the BEKK specification, are also a core tool for capturing volatility transmission. This model can not only estimate the mean-reversion property of an asset's volatility but, more importantly, dynamically track the conditional covariance among different crypto assets. By introducing a time variable, researchers have confirmed that the shock transmission and volatility-spillover effects among crypto assets have significant two-way features, and that their conditional correlation evolves dynamically over time [43]. Both the DY spillover index and BEKK-MGARCH are linear frameworks, however, with known limitations in handling extreme tail events and nonlinear contagion. In recent years, nonlinear transmission-measurement methods based on Copula-DCC and quantile regression have been applied in the crypto-market literature and can more accurately capture nonlinear features such as sudden correlation shifts and enhanced tail dependence during crises.

Further, combining network-topology theory, these econometric results can be visualized as complex directed weighted graphs. In these network topologies, centrality metrics such as eigenvector centrality or betweenness centrality are used to identify which perpetual futures occupy a systemically important position in volatility transmission. This network-perspective analysis not only reveals the direction of volatility transmission but also provides market regulators and risk managers with an early-warning system, helping them identify the key nodes in the network with high systemic-risk concentration before local volatility evolves into a global crisis.

22.6 Four-source interaction and regime switching

The four volatility sources do not operate independently of one another; among them there exist systematic nonlinear interactions and feedback-amplification mechanisms. More critically, the share that these sources occupy in total volatility (the volatility budget) is not fixed but undergoes a violent structural reorganization as the market regime switches. Understanding this regime dependence is central to revealing the microstructural foundations of volatility clustering.

22.6.1 The nonlinear interaction of the four sources

The four volatility sources exhibit approximate independence in calm market periods, but in a high-volatility environment their interaction terms amplify sharply, forming a complex feedback network. This interaction is first evident in the one-way triggering of liquidity volatility by informational volatility. When a new macro shock or fundamental information arrives, the degree of information asymmetry rises, and the adverse-selection risk faced by market makers intensifies. To compensate for this risk, market makers rationally widen the bid-ask spread and reduce order-book depth. At this point, subsequent order flow executes on thinner liquidity, producing a larger temporary price impact. Informational volatility therefore endogenously triggers liquidity volatility by changing microstructure parameters [23].

The most destructive link in the four-source interaction is the two-way positive feedback between liquidity volatility and mechanism-induced volatility. When insufficient liquid-

ity causes a relatively large temporary price decline, this decline may itself touch the liquidation threshold of long positions. Once forced liquidation is triggered, the liquidation engine submits market sell orders to the market, producing mechanism-induced volatility. These liquidation orders further consume the already-fragile liquidity, causing the price to continue falling and thereby triggering more liquidations. This spiraling feedback loop drives the price far from the equilibrium level implied by the initial information.

Transmitted volatility plays a cross-market amplifying role in this process. In a highly interconnected crypto-asset market, the volatility of a single asset is often rapidly contagious through market makers' cross-asset risk management and cross-margin mechanisms. When Bitcoin experiences liquidity contraction and liquidation, market makers, to control the risk exposure of their overall portfolio, synchronously withdraw their liquidity quotes in the Ethereum and altcoin markets. At the same time, the decline in the Bitcoin price lowers the portfolio collateral ratio of multi-asset investors, which in turn triggers chain liquidations of altcoin positions. This cross-asset contagion allows transmitted volatility to directly activate mechanism-induced and liquidity volatility in other markets, forming a nonlinear cascade effect.

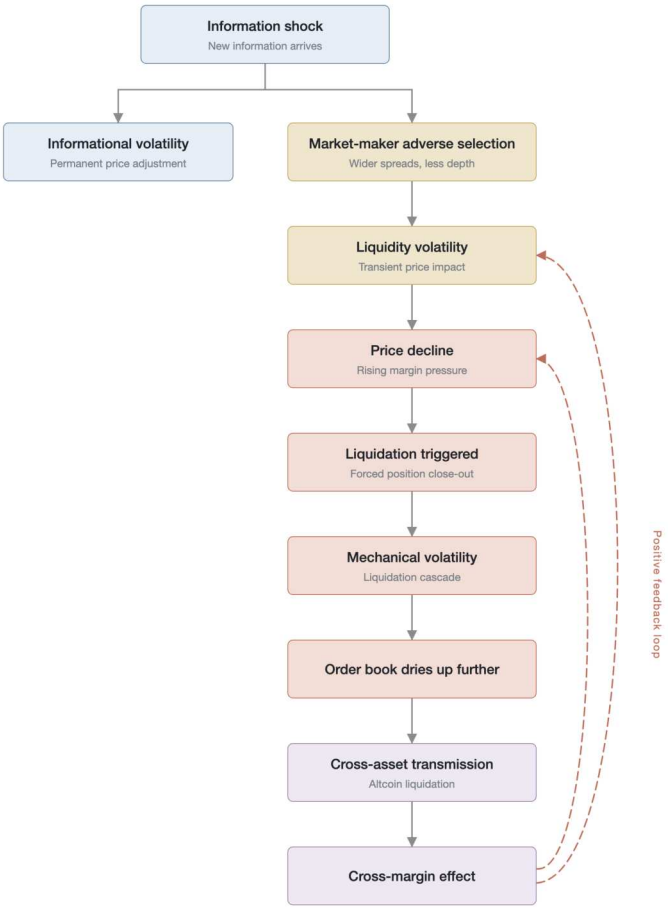


Figure 22-14. The nonlinear feedback loop of four-source interaction (drawn by the author, referencing the liquidity-spiral framework of Brunnermeier & Pedersen (2009) [23])

As shown in Figure 22-14, the feedback loop begins with an information shock: market makers retreat to avoid adverse selection, triggering the expansion of liquidity volatility; a temporary decline on the thin order book crosses the liquidation threshold, activating the liquidation engine and triggering mechanism-induced volatility; the liquidation sell orders then further consume liquidity, forming a two-way positive feedback. Ultimately, the extreme volatility spills over to other markets through cross-asset arbitrage and cross-margin, producing transmitted volatility, which may in turn re-trigger the above spiral in the target market, forming a global cascade effect.

22.6.2 The state dependence of the volatility budget

As introduced in Section 22.1.4, total volatility can be viewed as a budget jointly composed of four sources: informational, liquidity, mechanism-induced, and transmitted. The allocation of this budget depends heavily on the market regime, exhibiting a systematic structural switch from calm periods to crisis periods.

When the market is in a normal or calm period, the volatility budget is dominated mainly by informational volatility. At this time, market liquidity is ample, market makers compete fully, and the bid-ask spread stays at a low level. New fundamental information or macro data can be smoothly incorporated into prices, and the temporary price impact produced by large trades is small and quickly absorbed. Dimpfl and Peter's study of cross-spot-exchange price-discovery efficiency shows that the signal-to-noise ratio in crypto markets is about 36%, whereas that of the NYSE and NASDAQ is about 90% [11]. Although this metric measures the noise in cross-exchange price discovery (rather than the volatility composition in this chapter's four-source framework), it provides a useful reference benchmark: even in calm periods, the share of non-informational components in crypto markets is significantly higher than in traditional stock markets. This signal-to-noise difference partly reflects the structural features of crypto markets (24/7 trading, market fragmentation) rather than pure *inefficiency*, and the high signal-to-noise ratio of traditional markets partly stems from institutional noise-suppression mechanisms such as price limits and market-maker obligations. In the crypto perpetual futures market, informational volatility remains the largest of the four sources in calm periods. Mechanism-induced volatility contributes only limitedly at this stage, manifesting only as tiny arbitrage pulses around funding-rate settlement, while transmitted volatility also stays within the normal range of cross-market linkage.

When the market enters a tense period (for example, facing the uncertainty expectations of a major macro event or the eve of a key technical upgrade), the structure of the volatility budget begins to shift. As uncertainty rises, market makers, to avoid adverse-selection risk, actively widen quote spreads and reduce order-book depth. The deterioration in liquidity conditions causes order flow of the same size to produce larger price slippage, driving up the share of liquidity volatility in the total budget. At the same time, as the price range widens, some highly leveraged positions begin to face margin-call pressure, and small-scale liquidation events occur sporadically, so that the share of mechanism-induced volatility also begins to expand. The absolute amount of informational volatility may not decrease, but because of the rise in the other volatility sources, its relative share of the total budget declines, and more and more frictional noise mixes into market prices.

In an extreme crisis period, the structure of the volatility budget reverses violently. When the initial negative shock breaks through the market's tolerance threshold, sharp liquidity contraction and large-scale liquidation cascades occur simultaneously and reinforce each other. Market makers retreat on a large scale, and liquidity volatility expands sharply; the liquidation engine indiscriminately fires market orders, and mechanism-induced volatility dominates; panic and cross-asset sell-offs push transmitted volatility to its peak. Nguyen's research shows that the contribution of the noise component to the total variance of crypto prices reaches as high as 40%, far higher than the 21% of traditional stocks and the 18% of the foreign-exchange market [10] (note that this 40% is the unconditional mean for the full 2015–2024 sample, not the share under crisis conditions). In a crisis period, the share of non-informational components is very likely to rise significantly relative to calm periods and become dominant, but its precise value awaits empirical testing on a crisis subsample. In this regime, a considerable portion of the observed aggregate volatility no longer reflects fundamental information but is

manufactured by the collapse of market microstructure and institutional friction. This means that price movements in a crisis exhibit systematic overreaction, with a large number of assets sold off to levels far below their fundamental value.

This regime switch indicates that under extreme market conditions, the high volatility observed is mainly a representation of market-structure failure rather than a violent change in fundamental information. At the same time, this structural overreaction also provides a theoretical basis for strategists: since most of the volatility in a crisis period is information-free temporary noise, prices will undergo a systematic mean-reversion process after liquidity is restored and liquidation pressure subsides.

Taken together (for the compositional evolution of the volatility budget across the three states of calm, tense, and crisis, see Figure 22-2), under extreme market conditions the overwhelming majority of observable price movements have lost their price-discovery function and are driven purely by institutional friction and the liquidity spiral.

22.6.3 Triggers and early warning of regime switching

The switch of the market from a calm period to a crisis period is not a gradual, linear process but has a distinct critical-point effect. When external shocks or the accumulation of internal frictions break through a certain threshold, the market microstructure undergoes a sudden change. This nonlinear phase transition can be given early warning through specific microstructure indicators.

An abnormal decline in order-book depth is usually the earliest signal of a regime switch. Before the price moves substantially, market makers have often already sensed the rise in adverse-selection risk through the imbalance in micro order flow, and have withdrawn their limit orders in advance. When the depth at the top of the order book shrinks to below one standard deviation of its historical mean, the market's capacity to absorb shocks declines significantly, raising the probability of a liquidity-volatility outbreak.

The extremization of the funding rate and the widening of the cross-exchange basis are clear signs that mechanism-induced volatility is about to take over the market. When the funding rate of perpetual futures continuously deviates from the neutral level, and arbitrage capital cannot effectively close the basis between spot and contract, it indicates that the balance-sheet constraints of arbitrageurs have already tightened. In this state, the market is extremely vulnerable to the shock of a one-sided move, which in turn triggers large-scale liquidation. An initial rise in liquidation volume is the direct catalyst for the full eruption of the crisis state: once the selling pressure of the liquidation engine exceeds the residual liquidity-bearing capacity, the system irreversibly enters a cascade-liquidation state.

22.6.4 The liquidity-volatility spiral in the crisis state

In the crisis state, the exponential amplification of volatility is essentially the mapping of a liquidity tragedy of the commons onto the price dimension. As described in Chapter 20, liquidity in crypto markets has the property of a quasi-public good (more precisely, a common-pool resource). When volatility rises, the optimal rational choice for an

individual market maker is to widen spreads or suspend quoting to protect its own capital. When all market makers adopt this defensive strategy simultaneously, however, the liquidity of the entire market rapidly disappears.

This collective liquidity retreat directly triggers the liquidity-volatility spiral described in Sections 22.3.2 and 22.6.1: the extreme thinness of the order book causes stop-loss orders or liquidation orders to produce enormous slippage and extreme realized volatility; this spurious volatility is then captured by risk-management models, pushing up volatility forecasts and prompting market makers to exit even more decisively.

In this spiral, mechanism-induced volatility further intensifies the process. The mechanical execution of the liquidation mechanism is not price-sensitive; it indiscriminately submits market sell orders at the moment of worst liquidity, converting a temporary price shock into deeper balance-sheet destruction. In the liquidation-cascade event of October 2025 (whose deleveraging scale and window are detailed in Section 22.3.2), the most violent shock was concentrated within about 40 minutes, during which the intensity of cross-asset contagion was about 20% stronger than the spillover effect of the 2018 trade war [44]. This shows clearly how, in a crisis, the collapse of microstructure amplifies a limited information shock into a systemic, sharp rise in volatility.

22.6.5 Practical implications

The regime-switching framework of the volatility budget provides a new perspective for traders, risk managers, and market designers. For traders, identifying the core source of current volatility is a precondition for formulating a strategy. If one judges that the current high volatility is driven mainly by liquidity and mechanism-induced factors, this means the price movement contains a large amount of temporary shocks and overreaction, and a mean-reversion strategy may have a relatively high expected return. This judgment requires two preconditions, however. First, the shock must indeed contain no fundamental information; if it stems from a fundamentally destructive event such as a protocol being attacked or a project team absconding, the price will not revert (Luna in 2022 is a textbook counterexample). Second, the strategy must be actually executable, whereas under extreme conditions the exchange may restrict new position openings, throttle the API, or freeze withdrawals, making the theoretical mean-reversion opportunity impossible to capture. Making this ex ante judgment is precisely the task for which this chapter's four-source decomposition is designed. Conversely, if volatility is dominated by informational factors, contrarian trading will face enormous adverse-selection risk.

From a risk-management perspective, the traditional VaR model usually assumes that volatility is a homogeneous stochastic process, which is a fatal flaw in the perpetual futures market. Regime-aware risk management requires not being overconfident in calm periods and being able to dynamically adjust parameters in stress periods to adapt to the nonlinear eruption of mechanism-induced volatility. The funding rate, as an endogenous feedback rule, has an update frequency and cap whose settings directly affect the tail risk of basis volatility, and a reasonable mechanism design can effectively suppress the generation of an extreme basis [32]. The four-source framework provides a conceptual basis for building regime-aware risk management: by monitoring

microstructure indicators such as order-book depth, funding-rate extremity, and liquidation volume in real time, and dynamically adjusting the tail parameters of the VaR model (for example, switching from a normal distribution to a fat-tailed one) when the combination of indicators triggers a signal of a switch from calm to tense, one embeds the regime-switching logic of the volatility budget into the risk-management process.

For market designers, the key to reducing excess volatility in crisis periods lies in interrupting the vicious interaction among the four sources. Introducing a dynamic margin-buffer mechanism that appropriately relaxes liquidation conditions when volatility spikes can slow the occurrence of liquidation cascades. At the same time, optimizing cross-exchange circuit-breaker mechanisms to prevent the liquidity exhaustion of a single platform from rapidly spreading network-wide through oracles and arbitrageurs is a necessary direction of institutional evolution for enhancing the microstructural resilience of the crypto derivatives market. In addition, traditional financial markets solve the collective-action problem of liquidity retreat in crisis periods through a market-maker-obligation system; whether the market-maker incentive programs of CEXs can effectively maintain liquidity supply in a crisis is worthy of further empirical testing. In terms of leverage management, major jurisdictions have introduced differentiated leverage-cap policies (for example, Japan limits retail crypto margin-trading leverage to 2x, and the EU, under the MiFID II framework, has limited retail crypto contracts for difference (CFDs) to 2:1, with crypto derivatives not falling under the jurisdiction of MiCA), while offshore exchanges still allow leverage of more than 100x. This regulatory fragmentation is itself one of the institutional sources of excess volatility: the concentrated migration of traders toward high-leverage platforms intensifies the liquidation-cascade risk of these platforms.

22.7 The excess volatility of perpetual futures

Synthesizing the analysis of the preceding six sections on the microstructure of perpetual futures, the funding-rate mechanism, and the liquidation engine, this section directly answers a question of central importance in crypto finance: why the volatility of perpetual futures is systematically higher than that of the spot market. This phenomenon is termed *excess volatility* in the traditional finance literature, but in the crypto derivatives market, because of frictionless access to leverage and round-the-clock trading, it exhibits a distinctive generating mechanism and distinctive welfare-distribution consequences. This section disaggregates the sources of this excess volatility and assesses its welfare impact on different market participants.

22.7.1 Definition and measurement

The theoretical origin of excess volatility can be traced back to Shiller's (1981) classic study of the stock market, which showed that stock-price volatility systematically exceeds the level that dividend fundamentals can explain [45]. In the crypto derivatives market, excess volatility is defined as the difference between an asset's actual volatility in the derivatives market and the volatility that should prevail in the spot market under the same information set. This is a theoretical benchmark definition: the subtrahend—"the spot volatility that should prevail under the same information set"—is not directly

observable, nor can the observed spot volatility simply be substituted for it (the spot market itself also contains a noise component of about 40%); empirically, it can only be approximately identified with the help of the natural experiments discussed below [46] and the permanent/transitory variance decomposition [10], so the credibility of this benchmark is limited by the credibility of the spot-noise decomposition. Unlike the stock excess volatility Shiller found, which stems mainly from investor sentiment and discount-rate volatility, the excess volatility of perpetual futures is driven to a large extent by the mechanism-induced design and liquidity frictions analyzed in this chapter, which is an important feature distinguishing the crypto derivatives market from traditional asset markets (this chapter does not claim that the excess volatility of perpetuals can be explained by Shiller's discount-rate channel; it borrows only his measurement structure of "actual volatility exceeding the volatility that fundamentals should support"). In a fully efficient market, the price change of a derivative should reflect only the change in the discounted value of the underlying asset's future cash flows or updates to macro fundamental information. In crypto markets, however, the price-discovery process is often accompanied by significant microstructure noise. As cited in Section 22.6.2, the signal-to-noise ratio of the crypto market is significantly lower than that of traditional stock markets such as the NYSE and NASDAQ [11]. This high proportion of noise trading means that most price volatility is driven not by fundamental information but by the trading behavior of market participants itself.

The core challenge in measuring the excess volatility of perpetual futures is how to control for information differences. Because perpetual futures and the spot market share the same underlying asset, their price-discovery processes are highly intertwined, and a simple subtraction of volatilities cannot accurately isolate the common volatility driven by information flow. To isolate pure excess volatility, financial economists usually adopt a natural-experiment design. For example, by observing the changes in market structure before and after the approval of U.S. spot Bitcoin ETFs, researchers found that spot-market volatility declined significantly after the introduction of a more information-efficient investment channel, which supports the stabilization hypothesis that over-speculation in the derivatives market causes the spot market to bear additional volatility [46].

Another measurement framework separates the permanent component from the transitory component through variance decomposition of high-frequency data. The permanent component represents the information-driven price random walk, while the transitory component captures the mean-reverting volatility caused by liquidity shocks and market frictions. As cited in Section 22.6.2, the share of the noise component in the total variance of crypto prices is significantly higher than in traditional stock and foreign-exchange markets [10]. This systematically high noise share provides a benchmark for quantifying the microstructure frictions unique to perpetual futures.

22.7.2 Source attribution

The excess volatility of perpetual futures is not caused by a single factor but is the joint result of its unique market-microstructure features. One categorical point must be clarified first: by definition, excess volatility has already removed the informational

component common to perpetuals and spot (the two cancel each other out in the subtraction), so the source decomposition in this section focuses on three sources unique to perpetuals—mechanism-induced, liquidity, and transmitted. The "informational" item appearing below and in Figure 22-15 refers specifically to the information-overshoot differential that perpetuals produce relative to spot because of high leverage (its amplification mechanism is classified under mechanism-induced volatility), not to informational volatility itself. Figure 22-15 illustratively presents the relative contribution of these sources to the excess volatility of perpetual futures.

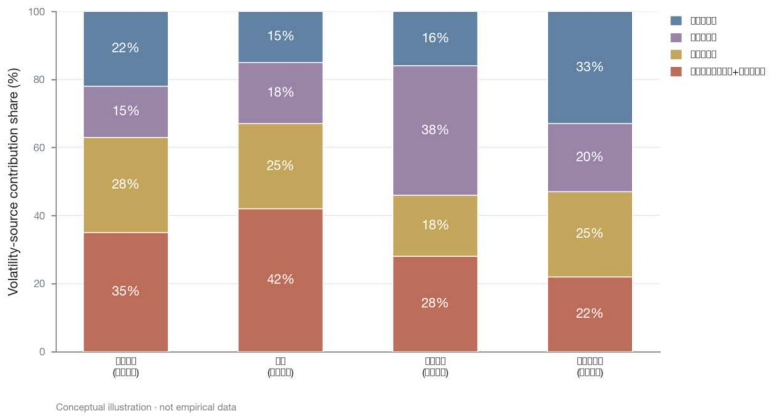


Figure 22-15. Source-attribution decomposition of the excess volatility of perpetual futures (drawn by the author: the relative share of each source is an illustrative estimate based on the logic of event-window/variance-ratio decomposition and the four-source conceptual framework; the four-source decomposition is a first-order approximation and liquidity is a residual category, see Section 22.1.3; not a precise measurement based on a single empirical sample)

As shown in Figure 22-15 (the proportions in the figure are an illustrative ordering, not a precise statistical decomposition), mechanism-induced volatility usually accounts for a larger share of the excess volatility (its core being the periodic noise of liquidation cascades and funding-rate settlement), followed by liquidity volatility, with transmitted volatility being especially prominent in crisis events; the information-overshoot differential, though a relatively small share, nevertheless shows by its existence that even a *good* price-discovery process generates some excess volatility through overshoot in a high-leverage environment.

Based on the analysis in Sections 22.4 and 22.3, the contribution of each source to excess volatility can be summarized as follows (the following is a conceptual attribution ordering based on mechanism analysis, not a precise measurement of variance shares). Mechanism-induced volatility is very likely the main contributor to excess volatility. As argued in Sections 22.4.1 and 22.4.2, liquidation cascades and the funding-rate settlement cycle are its core driving forces. Empirical evidence further shows that during the every-8-hour funding-rate settlement window of perpetual futures, adverse-selection risk in the spot market also increases significantly, causing quote spreads to widen and market depth to deteriorate [20], which means that the impact of mechanism-induced volatility has already spilled over into the microstructure of the spot market.

Liquidity volatility is another main contributor to excess volatility, rooted in the frag-

mentation of crypto-market depth and the constraints of market makers' inventory management. On high-frequency time scales, the temporary price deviation triggered by a large order and its subsequent partial reversal are a typical manifestation of liquidity friction [47]. Transmitted volatility contributes prominently in crisis events, as the extreme volatility of large-cap assets is rapidly transmitted to less liquid altcoin markets [48]. Although informational volatility is theoretically an efficient price-discovery process, because the high leverage of perpetual futures naturally attracts informed traders, the process of information being incorporated into prices is often accompanied by overshoot, and this overshoot itself also constitutes a component of excess volatility.

22.7.3 The heterogeneity of excess volatility

Excess volatility exhibits significant heterogeneity across different types of crypto assets and different stages of market development. From the asset dimension, Bitcoin's excess volatility is systematically lower than that of Ethereum and other mid-cap altcoins. This difference is mainly attributable to the fact that the Bitcoin spot market has the deepest liquidity and the broadest institutional participation, and its perpetual futures market has the highest arbitrage efficiency. By comparison, the altcoin spot market has poorer liquidity and lacks an effective short-selling mechanism, causing its perpetual futures to often accumulate a large basis in one-sided markets, which in turn triggers more violent funding-rate arbitrage volatility and liquidation shocks. Figure 22-16 tracks the evolution over time of the excess volatility of different asset classes.

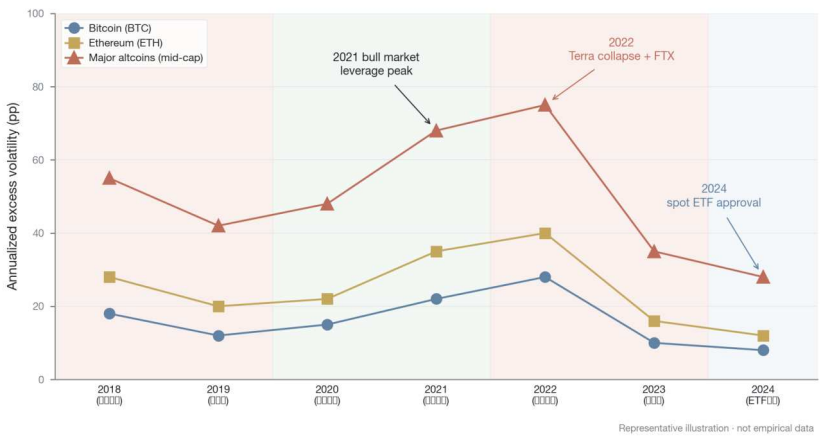


Figure 22-16. Time evolution of the excess volatility of perpetual futures across asset classes (representative illustration: the evolution trend, across the market cycle, of the annualized excess volatility of the perpetual futures of BTC, ETH, and altcoins ranked 10–30 by market capitalization relative to contemporaneous spot; a trend illustration, not measured monthly RV differences)

As shown in Figure 22-16, the excess volatility of altcoins is consistently higher than that of Bitcoin and Ethereum, and jumps by a larger amount when volatility intensifies; Bitcoin, by contrast, is relatively moderate and stable, reflecting its mature market-making ecosystem and efficient cross-market arbitrage. The long-term downward trend of the three curves suggests that the crypto derivatives market is gradually moving toward institutional maturity.

Examined from the time dimension, excess volatility exhibits a strong procyclical feature. In the bull-market phase, driven by retail FOMO and strong demand for leverage, market-wide open interest and leverage rise sharply. This high-leverage environment greatly amplifies the sensitivity of prices to small shocks, making excess volatility during a bull market significantly higher than during a bear or range-bound market. When the observation horizon is extended to the 10-year period from 2015 to 2025, however, one finds that the price-formation mechanism of the crypto market is undergoing a structural transformation. Since 2017, the noise component of the market's total variance has shown an overall downward trend, but with cyclical rebounds during speculative frenzies—for example, temporary jumps in the noise share can be observed during the 2020 DeFi boom and the 2024 memecoin wave. At the same time, the component reflecting overall market information has risen steadily from a long-term perspective.

This long-term downward trend reflects the gradual maturation of the crypto market. As market regulation has tightened, derivatives infrastructure has matured, and institutional investors have entered in depth, the market's pricing efficiency has improved substantially. For example, empirical analysis of successive Bitcoin halving events shows that, as the market becomes more standardized, the abnormal returns and excess volatility around halving events are both declining significantly [49]. The abundance of institutional arbitrage capital allows the calendar and cross-market basis to be closed more quickly, thereby suppressing the extreme price volatility triggered by local liquidity exhaustion.

22.7.4 Welfare implications

Quantifying excess volatility ultimately requires answering its welfare-distribution impact on market participants. The additional volatility created by perpetual futures is not evenly distributed among all participants but profoundly reshapes the market's wealth-transfer pattern. For profit-maximizing informed traders and high-frequency arbitrage institutions, high volatility is directly equivalent to high-frequency trading opportunities and rich arbitrage space. Using advanced algorithms and low-latency infrastructure, they extract profits from the periodic swings of the funding rate and the instantaneous mispricing triggered by liquidation events.

Market makers' attitude toward excess volatility, however, is more complex. Moderate volatility can bring considerable bid-ask spread income, but when volatility breaks through a certain threshold, the inventory risk faced by market makers rises sharply. Empirical data from the options market show that when market makers' aggregate gamma inventory is imbalanced, they significantly widen quote spreads to hedge risk, causing overall market liquidity to dry up [50]. This liquidity retreat often occurs precisely when the market most needs depth, further intensifying the spiraling price decline. Figure 22-17 uses a tiered framework diagram to summarize the differentiated welfare impact of excess volatility on different groups of market participants.

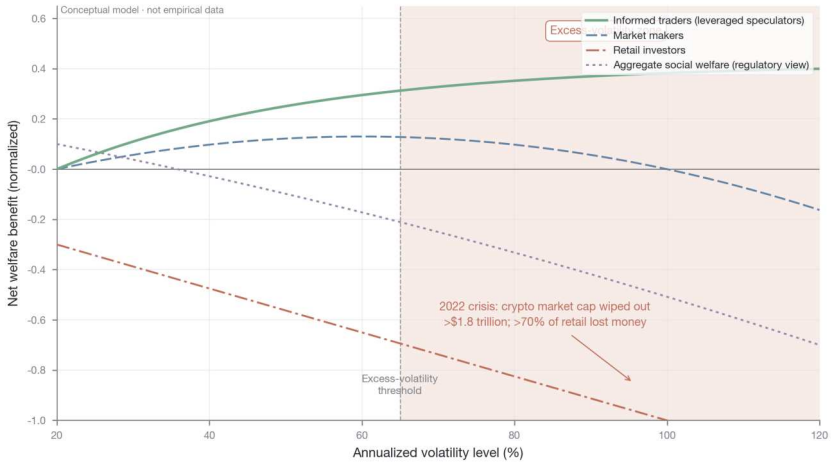


Figure 22-17. The welfare impact of excess volatility on different market participants (Data source: drawn by the author, supplemented with data on retail loss rates from Cornelli et al. (2023) [51])

As shown in Figure 22-17, the welfare effect of excess volatility is markedly asymmetric: informed traders and high-frequency arbitrage institutions sit at the positive end, benefiting from higher volatility and more frequent mispricing opportunities; market makers are in the middle, with moderate volatility bringing spread income but extreme volatility bringing inventory losses; retail investors are at the negative end, lacking an informational advantage and risk-control tools, more easily triggering forced liquidation and thereby becoming the bearers of a systemic wealth transfer.

Near the positive end of the welfare distribution, there is also a class of size-driven participants who win by position size rather than analytical capability—namely, large holders. Although these participants do not rely on an informational or analytical advantage, their large positions give them the ability to influence the timing of a liquidation cascade through strategic selling or accumulation. In excess volatility, large holders play a dual role: they may profit by intensifying mechanism-induced volatility through selling, or they may become a key force for market stability when liquidity is exhausted. This internal heterogeneity suggests that the three-way framework of Figure 22-17 still admits further stratification at the informed-trader end.

Retail investors are the main losers in excess volatility. Lacking sophisticated risk-management tools and a deep understanding of market microstructure, retail investors are extremely prone to triggering forced liquidation during excess volatility. A BIS research report reveals this striking wealth transfer: during the crypto-market crisis of 2022 (during which market-wide capitalization shrank by more than $1.8 trillion), large investors successfully reduced their holdings before the crisis erupted, while a large number of retail investors added to positions against the trend during the decline, ultimately causing most crypto-app users to suffer substantial losses [51]. This systematic transfer of wealth to institutions indicates that the high-leverage mechanism of perpetual futures has a significant negative welfare effect in an environment lacking investor protection.

From the macro perspective of regulators, persistent excess volatility is not only a

signal of market manipulation and institutional deficiency but also a manifestation of financial-resource misallocation. If the price movements of a market are driven mainly by endogenous mechanism friction and speculative noise rather than reflecting fundamental value discovery, then that market cannot effectively perform the core function of capital allocation. How to optimize the design of the liquidation mechanism—for example, by introducing auction-based liquidation to reduce price impact [31]—and how to regulate leverage caps will therefore be key challenges for the crypto derivatives market as it moves toward maturity and ultimately integrates into the mainstream financial system.

22.8 Chapter summary

Price volatility in perpetual futures markets is not a homogeneous random-walk process but a complex product jointly shaped by four mechanisms of fundamentally different natures. Informational volatility reflects how the market prices new fundamentals, macro events, or on-chain data, and is an indispensable price-discovery mechanism of a healthy market. Liquidity volatility arises from temporary price shocks when order-book depth is insufficient and is a friction cost of market microstructure. Mechanism-induced volatility is driven entirely by the institutional design unique to perpetual futures, including the forced liquidations triggered by liquidation cascades and the predictable pulses brought by the funding-rate settlement cycle. Transmitted volatility reveals the linkage across assets and markets, allowing an isolated event to spread globally through arbitrage channels and margin-cross effects.

The relative shares of these four sources under different market regimes constitute the dynamic switching mechanism of the volatility budget. In a normal market, informational volatility dominates, and price movements mainly reflect the genuine arrival of information. When the market faces stress or falls into crisis, however, sharp liquidity contraction and mechanism-induced liquidation rapidly take over the volatility budget. At this point, most of the price movement observed in the market no longer contains new information but is pure structural noise. This exit of informational volatility and sharp expansion of frictional volatility under extreme conditions is central to understanding the fragility of the crypto market.

A direct corollary of the four-source decomposition framework is that perpetual futures markets exhibit significant excess volatility. Comparison with the spot market and traditional financial derivatives shows that the total volatility of perpetual futures is systematically higher than the level its information content should support. This excess portion is contributed mainly by liquidity frictions and mechanism-induced deficiencies and is a hidden cost that the current institutional design imposes on all market participants. The existence of excess volatility not only increases market makers' inventory risk but also causes retail traders to trigger stop-losses more frequently, systematically harming market welfare. More critically, because mechanism-induced volatility such as liquidation cascades and funding-rate settlement is largely endogenous to the rules, this portion of excess volatility can in principle be substantially reduced through institutional optimization (such as introducing gradual liquidation or a continuous funding rate); but because there is a fundamental trade-off between liquidation efficiency and

system solvency, eliminating mechanism-induced volatility entirely is neither realistic nor desirable.

Decomposing the sources of volatility not only clarifies the multiple causes behind a single price movement but also lays the foundation for understanding the evolution of volatility over time. The four volatility sources do not occur in isolation along the time dimension but reinforce one another through systematic interaction effects. When an initial information shock triggers a price adjustment, the ensuing adverse-selection risk causes market makers to retreat, thereby activating liquidity volatility. The enlarged temporary price shock further touches the bankruptcy price of leveraged traders, setting off a liquidation cascade and causing mechanism-induced volatility to rise sharply. Ultimately, this extreme local volatility spills outward through cross-asset arbitrage and cross-margin mechanisms, forming transmitted volatility, which in turn further intensifies the global liquidity contraction.

This nonlinear positive-feedback loop—triggered by information, amplified by liquidity, detonated by mechanisms, and finally transmitted market-wide—explains why high volatility is often not a fleeting, isolated event. Once volatility breaks through a critical threshold, the mutual reinforcement among the sources makes it difficult for the system to return to calm in the short run. This leads to the next core topic: now that we have clarified the constituent elements of volatility and their interaction mechanisms, what macro features does this positive-feedback process exhibit over time? Why does volatility always exhibit strong time-variation and autocorrelation in the market? Chapter 23 will formally examine the dynamic mechanism of this volatility-clustering phenomenon, analyzing how slow variables, feedback loops, and state machines jointly determine the market's switching between low- and high-volatility states.

References

[1] Sandor, K. (2024, March 5). Bitcoin tumbles 10% after hitting record high; triggers $1B crypto liquidations. *CoinDesk*. https://www.coindesk.com/markets/2024/03/05/bitcoin-tumbles-6-after-hitting-record-high-as-mounting-sell-orders-cap-price-gains

[2] He, S., Manela, A., Ross, O., & von Wachter, V. (2022). Fundamentals of perpetual futures. *arXiv*. https://arxiv.org/abs/2212.06888

[3] Soska, K., Dong, J.-D., Khodaverdian, A., Zetlin-Jones, A., Routledge, B. R., & Christin, N. (2021). Towards understanding cryptocurrency derivatives: A case study of BitMEX. In *Proceedings of the Web Conference 2021* (pp. 45–57). ACM. https://doi.org/10.1145/3442381.3450059

[4] Andersen, T. G., Bollerslev, T., & Meddahi, N. (2005). Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica*, *73*(1), 279–296.

[5] Palm, F. C. (1996). GARCH models of volatility. In G. S. Maddala & C. R. Rao (Eds.), *Handbook of statistics*: Vol. 14. *Statistical methods in finance* (pp. 209–240). Elsevier. https://doi.org/10.1016/s0169-7161(96)14009-8

[6] Hasbrouck, J. (1995). One security, many markets: Determining the contributions

to price discovery. *The Journal of Finance*, *50*(4), 1175–1199. https://doi.org/10.1111/j.1540-6261.1995.tb04054.x

[7] Hasbrouck, J. (2002). Stalking the "efficient price" in market microstructure specifications: An overview. *Journal of Financial Markets*, *5*(3), 329–339. https://doi.org/10.1016/s1386-4181(02)00029-0

[8] Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, *39*(4), 1127–1139. https://doi.org/10.2307/2327617

[9] Hasbrouck, J. (1991). Measuring the information content of stock trades. *The Journal of Finance*, *46*(1), 179–207. https://doi.org/10.1111/j.1540-6261.1991.tb03749.x

[10] Nguyen, K. Q. (2024). *Why are cryptocurrencies so volatile?* [Working paper]. Auckland University of Technology.

[11] Dimpfl, T., & Peter, F. J. (2021). Nothing but noise? Price discovery across cryptocurrency exchanges. *Journal of Financial Markets*, *54*, Article 100584. https://doi.org/10.1016/j.finmar.2020.100584

[12] Pyo, S., & Lee, J. (2020). Do FOMC and macroeconomic announcements affect Bitcoin prices? *Finance Research Letters*, *37*, Article 101386. https://doi.org/10.1016/j.frl.2019.101386

[13] Ben Omrane, W., Dabbou, H., Saadi, S., Savaser, T., & Sebai, S. (2025). Exploring volatility reactions in cryptocurrency markets using intraday macroeconomic news analysis. *International Review of Economics & Finance*, *103*, Article 104509. https://doi.org/10.1016/j.iref.2025.104509

[14] Mehdian, S., Gherghina, Ș. C., & Stoica, O. (2025). The reaction of cryptocurrencies to the approval of spot Bitcoin and Ethereum ETFs: An intraday event study. *Borsa Istanbul Review*. Advance online publication.

[15] Magner, N., & Sanhueza, A. (2025). The Moby Dick effect: Contagious Bitcoin whales in the crypto market. *Finance Research Letters*, *85*, Article 107906. https://doi.org/10.1016/j.frl.2025.107906

[16] Grande, M., & Borondo, J. (2025). Trust as a driver in the DeFi market: Leveraging TVL/MCAP bands as confidence indicators to anticipate price movements. *Finance Research Letters*, *75*, Article 106705. https://doi.org/10.1016/j.frl.2024.106705

[17] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1336. https://doi.org/10.2307/1913210

[18] Easley, D., Kiefer, N. M., O'Hara, M., & Paperman, J. B. (1996). Liquidity, information, and infrequently traded stocks. *The Journal of Finance*, *51*(4), 1405–1436. https://doi.org/10.1111/j.1540-6261.1996.tb04074.x

[19] Tiniç, M., Sensoy, A., Akyildirim, E., & Corbet, S. (2023). Adverse selection in cryptocurrency markets. *The Journal of Financial Research*, *46*(2), 497–546. https://doi.org/10.1111/jfir.12317

[20] Ruan, Q., & Streltsov, A. (2022). *Perpetual futures contracts and cryptocurrency*

market quality (SSRN Working Paper No. 4218907). https://doi.org/10.2139/ssrn.4218907

[21] Ben Omrane, W., Houidi, F., & Savaser, T. (2024). Macroeconomic news and intraday seasonal volatility in the cryptocurrency markets. *Applied Economics*, *56*(38), 4594–4610. https://doi.org/10.1080/00036846.2023.2212970

[22] Alexander, C., & Heck, D. F. (2020). Price discovery in Bitcoin: The impact of unregulated markets. *Journal of Financial Stability*, *50*, Article 100776. https://doi.org/10.1016/j.jfs.2020.100776

[23] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[24] Forino, A., & Morelli, G. (2024). *Liquidity and volatility nexus in cryptocurrency markets* (SSRN Working Paper No. 5128381). https://doi.org/10.2139/ssrn.5128381

[25] FTI Consulting. (2025, December 24). *The crypto crash of October 2025: When leverage met liquidity*. https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity

[26] Brauneis, A., Mestel, R., & Theissen, E. (2025). The crypto world trades at tea time: Intraday evidence from centralized exchanges across the globe. *Review of Quantitative Finance and Accounting*, *64*(1), 275–304. https://doi.org/10.1007/s11156-024-01304-1

[27] Al-Yahyaee, K. H., Mensi, W., Ko, H. U., Yoon, S. M., & Kang, S. H. (2020). Why cryptocurrency markets are inefficient: The impact of liquidity and volatility. *The North American Journal of Economics and Finance*, *52*, Article 101168. https://doi.org/10.1016/j.najef.2020.101168

[28] Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56. https://doi.org/10.1016/S1386-4181(01)00024-6

[29] Zhang, W., & Li, Y. (2023). Liquidity risk and expected cryptocurrency returns. *International Journal of Finance & Economics*, *28*(1), 472–492. https://doi.org/10.1002/ijfe.2431

[30] Fasanara Digital. (2022, May 21). May 2022 Terraquake. The demise of Luna. *Medium*. https://fasanaradigital.medium.com/terraquake-3080f4b5e5cf

[31] Tian, P., & Zhu, Y. (2025). *Liquidation mechanisms and price impacts in DeFi* (Staff Working Paper 2025-12). Bank of Canada. https://www.bankofcanada.ca/2025/03/staff-working-paper-2025-12/

[32] Zhang, T. (2026). *Funding rate mechanism in perpetual futures* (SSRN Working Paper No. 6185958). https://doi.org/10.2139/ssrn.6185958

[33] Chen, E., Ma, M., & Nie, Z. (2024). Exploring the impact: How decentralized exchange designs shape traders' behavior on perpetual future contracts. *arXiv*. https://arxiv.org/abs/2402.03953

[34] Yi, S., Xu, Z., & Wang, G.-J. (2018). Volatility connectedness in the cryptocurrency

market: Is Bitcoin a dominant cryptocurrency? *International Review of Financial Analysis*, *60*, 98–114. https://doi.org/10.1016/j.irfa.2018.08.012

[35] Korkusuz, B. (2025). Volatility transmission in digital assets: Ethereum's rising influence. *Journal of Risk and Financial Management*, *18*(3), 111. https://doi.org/10.3390/jrfm18030111

[36] Diebold, F. X., & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting*, *28*(1), 57–66. https://doi.org/10.1016/j.ijforecast.2011.02.006

[37] Shahzad, S. J. H., Bouri, E., Kang, S. H., & Saeed, T. (2021). Regime specific spillover across cryptocurrencies and the role of COVID-19. *Financial Innovation*, *7*, Article 5. https://doi.org/10.1186/s40854-020-00210-4

[38] Apostolakis, G. N. (2024). Bitcoin price volatility transmission between spot and futures markets. *International Review of Financial Analysis*, *94*, Article 103251. https://doi.org/10.1016/j.irfa.2024.103251

[39] Schmeling, M., Schrimpf, A., & Todorov, K. (2023). *Crypto carry* (BIS Working Papers No. 1087). Bank for International Settlements. https://www.bis.org/publ/work1087.htm

[40] Fassas, A. P., Papadamou, S., & Koulis, A. (2020). Price discovery in bitcoin futures. *Research in International Business and Finance*, *52*, Article 101116. https://doi.org/10.1016/j.ribaf.2019.101116

[41] Iyer, T. (2022). *Cryptic connections: Spillovers between crypto and equity markets* (Global Financial Stability Notes No. 2022/01). International Monetary Fund.

[42] Elsayed, A. H., & Sousa, R. M. (2024). International monetary policy and cryptocurrency markets: Dynamic and spillover effects. *The European Journal of Finance*, *30*(16), 1855–1875. https://doi.org/10.1080/1351847X.2022.2068375

[43] Katsiampa, P., Corbet, S., & Lucey, B. (2019). Volatility spillover effects in leading cryptocurrencies: A BEKK-MGARCH analysis. *Finance Research Letters*, *29*, 68–74. https://doi.org/10.1016/j.frl.2019.03.009

[44] Ali, Z. (2025). *Anatomy of the Oct 10–11, 2025 crypto liquidation cascade: Macroeconomic triggers, market microstructure, and systemic risk lessons* (SSRN Working Paper No. 5611392). https://doi.org/10.2139/ssrn.5611392

[45] Shiller, R. J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review*, *71*(3), 421–436.

[46] Babalos, V., Bouri, E., & Gupta, R. (2025). Does the introduction of US spot Bitcoin ETFs affect spot returns and volatility of major cryptocurrencies? *The Quarterly Review of Economics and Finance*, *102*, Article 102006. https://doi.org/10.1016/j.qref.2025.102006

[47] Morin, A., & Moore, T. (2024). *The impact of cryptocurrency spot price volatility on derivative markets* (SSRN Working Paper No. 4729600). https://doi.org/10.2139/ssrn.4729600

[48] Korkusuz, B., & Sahiner, M. (2025). Coin impact on cross-crypto realized volatility

and dynamic cryptocurrency volatility connectedness. *Financial Innovation*, *11*. https://doi.org/10.1186/s40854-025-00881-x

[49] Veloso, V., Gatsios, R. C., Magnani, V. M., & Lima, F. G. (2025). Is Bitcoin's market maturing? Cumulative abnormal returns and volatility in the 2024 halving and past cycles. *Journal of Risk and Financial Management*, *18*(5), 242. https://doi.org/10.3390/jrfm18050242

[50] Atanasova, C., Miao, T., Segarra, I., & Willeboordse, F. (2025). Aggregate illiquidity and crypto option returns. *Finance Research Letters*, *85*(C), Article 108003. https://doi.org/10.1016/j.frl.2025.108003

[51] Cornelli, G., Doerr, S., Frost, J., & Gambacorta, L. (2023). *Crypto shocks and retail losses* (BIS Bulletin No. 69). Bank for International Settlements. https://www.bis.org/publ/bisbull69.htm

The Microstructural Mechanisms of Volatility Clustering

The volatility of Bitcoin perpetual futures follows a pattern no statistician can ignore: high volatility tends to follow high volatility, and low volatility tends to follow low volatility. In cryptocurrency markets, this volatility clustering is more pronounced and more persistent than in traditional financial assets. Empirical research shows that the probability of volatility clustering in crypto assets—high volatility persistently followed by high volatility—is significantly higher than in traditional financial assets [1]. This characteristic implies that volatility is not pure white noise but possesses deep long-memory properties. Generalized autoregressive conditional heteroskedasticity (GARCH) models and their variants (detailed in Section 23.1.1) can capture this clustering at the statistical level [2]. However, the successful fit of a statistical model is merely a mathematical description of the phenomenon; it does not answer a more fundamental economic question: why does the current conditional variance depend on past shocks? What exactly is the microstructural mechanism that drives this volatility memory and clustering?

This question marks the starting point for a deeper investigation into volatility dynamics. Statistical description tells us that high volatility is more likely to follow high volatility, but understanding and predicting market behavior requires uncovering the causal chain behind the data. In perpetual futures markets, volatility is not determined independently by a single random event; rather, it is shaped jointly by a series of interwoven market frictions, institutional designs, and participant behaviors. When an exogenous information shock breaks the market's calm, it alters not only the current price but also the market's liquidity depth, the level of systemic leverage, and market makers' risk appetite. Changes in these variables set new initial conditions for the next period's price movement, so that a new shock of the same magnitude can produce a volatility response far exceeding the norm under the new market state. The microstructural mechanism of volatility clustering therefore lies in how the market state, evolving on its own across different time scales, amplifies external shocks.

To explain this process systematically, this chapter constructs a three-layer explanatory framework composed of slow variables, feedback loops, and a state machine. At the lowest layer, the macroeconomic environment, the level of systemic leverage, and liquidity depth act as slow variables that evolve gradually over longer time scales, setting the baseline for market fragility. These slow variables do not directly generate violent

price swings, but they determine the system's capacity to absorb external shocks. At the intermediate layer, strong positive feedback loops form between volatility and liquidity, between volatility and leverage, and between volatility and participant behavior. Once the slow variables accumulate to a critical point, a minor triggering event can activate these feedback loops, causing shocks to be continuously amplified rather than dampened within the system. At the top layer, the interaction of these mechanisms manifests as nonlinear jumps of the market among three discrete states: calm, stressed, and crisis. These three layers of mechanism may exist in any leveraged market, but the institutional design specific to perpetual futures—the buffer-free transmission of liquidation cascades, the pulse-superposition effect of funding rates, and the around-the-clock trading environment with no market close—imposes systematic amplification on all three layers, explaining why the volatility clustering effect in crypto markets far exceeds that of traditional assets. This chapter also reveals that the accumulation rate of systemic risk often peaks during low-volatility periods, and it ties all the theoretical components together through a complete volatility life-cycle case study.

Dissecting these three layers of mechanism one by one moves the analysis beyond a purely statistical view of volatility toward one grounded in market microstructure and dynamics. Grasping how slow variables gradually accumulate fragility, how feedback loops amplify shocks, and how the state machine drives regime transitions not only explains why high-volatility events cluster in the time series; it also provides a solid theoretical foundation and a structured framework for building forward-looking volatility forecasting models and dynamic risk management strategies in the next chapter.

23.1 The empirical facts of volatility clustering

Financial asset time series commonly exhibit a statistical feature known as *volatility clustering*: large price movements tend to be followed by large price movements, while small price movements tend to be followed by small price movements. In crypto assets, and especially in perpetual futures markets, this clustering effect is particularly severe. Using empirical data and econometric models, this section establishes the empirical facts of volatility clustering in perpetual futures markets and compares them quantitatively with traditional financial markets. It aims not only to describe the phenomenon precisely but also to motivate the inquiry into the underlying microstructural mechanisms that follows.

23.1.1 Definition and measurement

The essence of volatility clustering is the persistence over time of the variance of asset returns. Operationally, this persistence is most often measured through the GARCH model. Since Engle (1982) introduced the autoregressive conditional heteroskedasticity (ARCH) model and Bollerslev (1986) generalized it into the GARCH form, this framework has become the standard tool for characterizing volatility clustering in financial time series [3][4]. The model assumes that the current conditional variance depends not only on past squared returns (i.e., external shocks) but also on past conditional variances. The standard form of the GARCH(1,1) variance equation is:

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

where σ_t^2 is the conditional variance in period t, $\omega > 0$ is a constant term (which determines the long-run variance level), ϵ_{t-1} is the return residual from the previous period (i.e., the external shock), $\alpha \geq 0$ measures the immediate impact of new information shocks on volatility, and $\beta \geq 0$ measures the memory effect of historical volatility on current volatility. Their sum, $\alpha + \beta$, is called the volatility persistence parameter; the closer its value is to 1, the stronger the volatility clustering effect—that is, the longer high-volatility or low-volatility states persist [5] (when $\alpha + \beta < 1$, the process is covariance-stationary, and the half-life of a shock is $\ln 0.5 / \ln(\alpha + \beta)$).

In most traditional financial assets, this persistence parameter typically falls within the range of 0.95 to 0.99. This means that a major market shock triggers not only violent volatility on the day itself but also an effect that persists, in diminishing fashion, over the following weeks or even months. This slowly decaying process constitutes the volatility half-life—the time required for a shock's impact to halve. For assets with a high persistence parameter, the volatility half-life can extend to several weeks. Beyond parameter estimation, volatility clustering can also be observed directly through the autocorrelation function of realized volatility. In an idealized random-walk market that lacks any clustering effect, the autocorrelation coefficients of realized volatility should decay rapidly to zero; in real markets, however, these coefficients often remain significantly positive across many lag orders, confirming the self-perpetuating character of volatility states.

Figure 23-1 uses a GARCH(1,1) model to simulate the conditional volatility of BTC perpetual futures across a typical market cycle.

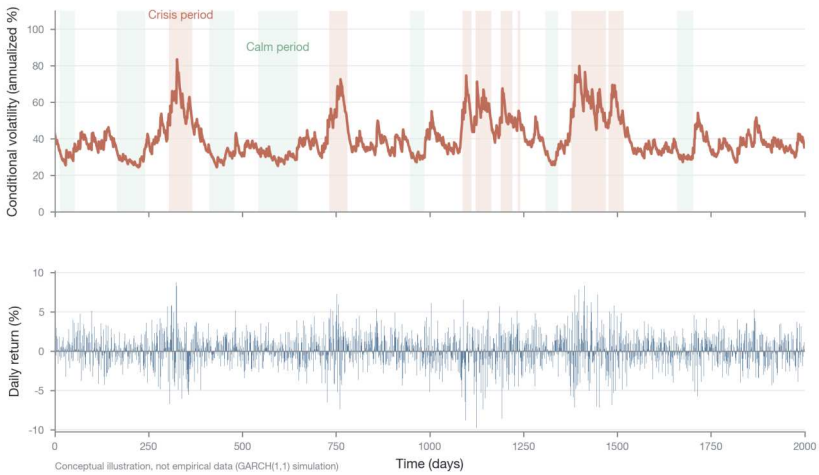


Figure 23-1. The clustering pattern of BTC perpetual futures volatility (Simulated by the author using a GARCH(1,1) model; parameters: ω=0.00001, α=0.08, β=0.90, sample length 2,000 days; illustrative simulation, not empirical data)

In Figure 23-1, volatility is not uniformly distributed but exhibits a clearly clustered structure: each high-volatility cluster contains multiple consecutive days of extreme

movement. When the market moves from a calm period into a stressed or crisis period, the volatility baseline rises markedly and sustains itself for a considerable time, until the intrinsic constraints of market mechanisms gradually exhaust it. This clustered distribution is precisely the intuitive manifestation of $\alpha + \beta$ approaching 1 in the GARCH model: past shocks, through the path dependence of the conditional variance, continuously transmit energy along the time series.

23.1.2 Cross-market quantitative comparison

Comparing crypto-asset perpetual futures with traditional financial markets more clearly delineates the distinctiveness of their volatility clustering. Empirical research shows that the returns of major crypto assets such as BTC and ETH not only exhibit the leptokurtic, fat-tailed features similar to those of traditional markets, but that the intensity of their volatility clustering is significantly higher than that of traditional assets across multiple dimensions [6]. This difference does not stem simply from the innate "high volatility" of crypto assets; rather, it is deeply rooted in their specific market microstructure and derivatives design.

To quantify the specific magnitude of this difference, Table 23-1 systematically compares BTC perpetual futures, BTC spot, and S&P 500 index futures across five dimensions: the persistence parameter, the autocorrelation coefficient, the half-life, the duration of clustering episodes, and average volatility.

Metric	BTC perpetual futures	CME BTC futures	BTC spot	S&P 500 futures
GARCH(1,1) $\alpha+\beta$	0.97–0.99	0.95–0.97	0.95–0.98	0.95–0.98
Daily autocorrelation of realized volatility	0.65–0.80	0.60–0.74	0.58–0.72	0.55–0.70
Volatility half-life (days)	15–40	12–28	12–30	10–25
Duration of extreme volatility clusters (days)	5–15	4–10	4–10	3–8
Average daily volatility (%)	3.5–5.2	3.3–4.9	3.2–4.8	1.2–1.8

Table 23-1. Cross-market quantitative comparison of volatility clustering intensity (Data source: the author's directional estimates synthesized from multiple sources; the specific ranges are not reported cell-by-cell from a single source [6][1])

The quantitative comparison reveals several key facts. The most diagnostically valuable contrast in the table is not the difference between BTC perpetual futures and the S&P 500 (the two differ fundamentally in the risk premium and volatility level of the underlying asset), but the difference between BTC perpetual futures and CME BTC futures, because the two share the same underlying asset yet differ significantly in institutional design (CME uses a traditional margin-call mechanism, imposes trading-session limits, and sets lower leverage caps). BTC perpetual futures are systematically higher than

CME BTC futures in both the persistence parameter and the half-life, a difference that points more cleanly to the institutional amplification effect of perpetual futures than to the volatility characteristics of the underlying asset itself. The $\alpha+\beta$ ranges of BTC perpetual futures and BTC spot do overlap (0.97–0.99 versus 0.95–0.98), so this difference may not be statistically significant, and the magnitude of the institutional amplification effect requires rigorous testing on larger samples. The high persistence of BTC perpetual futures may also be driven partly by macro factors, such as the volatility cycles of the VIX and the U.S. dollar index, rather than being entirely endogenous to the perpetual futures mechanism, and this motivates the slow-variable analysis in Section 23.2. The duration of extreme volatility clusters in perpetual futures is significantly longer than in traditional markets or CME BTC futures, and this prolonged high-pressure state places extremely high demands on market makers' inventory management and risk control.

23.1.3 Nonlinear clustering features

Traditional autoregressive models tend to depict volatility clustering as a smooth, symmetric, slowly decaying process. In perpetual futures markets, however, the clustering phenomenon exhibits more complex nonlinear features. Cont (2007) [7], using an agent-based simulation model, showed that when trend followers and fundamental traders interact dynamically in the market, volatility clustering can arise spontaneously as an emergent property, without any exogenous GARCH structure. The analysis in this chapter is complementary to Cont's agent-based model: Cont's model focuses on how information diffusion and strategy imitation generate clustering in the cross-section, whereas this chapter's three-layer framework focuses on how leverage, liquidity, and liquidation mechanisms sustain and amplify clustering along the time series. Together they point to the multi-source origins of volatility clustering. These nonlinear features challenge simple statistical descriptions and call for a more refined mechanistic analysis.

Tail clustering is a prominent nonlinear feature of crypto-asset markets. Extreme volatility events—that is, observations in the tails of the return distribution—rarely occur in isolation; instead, they tend to erupt in bursts within a short period. This phenomenon is already evident in the distribution of extreme funding-rate values, and along the volatility dimension it manifests as a single 5-standard-deviation shock often rapidly inducing several subsequent shocks of 3 to 4 standard deviations [8]. Such bursts of extreme events indicate that the market's risk-absorption capacity is structurally damaged after the first shock, so that subsequent minor disturbances can also trigger violent price responses.

Jump clustering describes the discontinuous transition of volatility states. In perpetual futures, the rise in volatility is often not gradual but accompanied by jump-like discontinuities in price. After the market experiences a liquidation cascade triggered by leverage unwinding, the probability of a second or even third cascade occurring within a short period rises sharply and nonlinearly. This clustering of jumps reflects the asymmetric dynamics of market liquidity in a crisis state: sudden evaporation and slow recovery.

Directional clustering reveals the asymmetry of volatility clustering. Empirical data

show that volatility clustering triggered by price declines significantly exceeds clustering triggered by price increases in both intensity and duration. The root of this asymmetry is that the positive feedback loop of the liquidation mechanism is stronger in a declining market: falling prices erode long-position margin, triggering forced liquidations; liquidation orders act as market sell orders that push prices down further, forming a self-reinforcing downward spiral. By comparison, although short liquidations also occur in a rising market, their systemic positive feedback is typically weaker. This asymmetry stems from three structural factors. First, the share of long open interest in perpetual futures markets is typically higher than that of shorts, so the stock of positions available for liquidation is larger on the downside. Second, long liquidations destroy buy-side purchasing power at the system level (margin is confiscated), whereas the purchasing power released by short liquidations is dispersed across different accounts and is not necessarily reinvested in the market. Third, market makers withdraw liquidity more rapidly in response to downside volatility than to upside volatility, because downside risk typically poses a more direct threat to market makers' inventory.

These empirical facts—tail clustering, jump clustering, and directional clustering—demonstrate that although statistical models can precisely fit the persistence of volatility, they cannot answer the mechanistic question of why the conditional variance depends on past shocks. Moving from "description" to "explanation" requires introducing the causal chain of market microstructure and participant behavior. The next section begins with the outermost slow variables and gradually constructs this causal chain.

23.2 Slow variables and the fragility baseline

Beneath the surface appearance of volatility clustering lie complex microstructural dynamics. The framework faces an endogeneity problem at the outset: volatility is both the explained variable and a determinant of the state of the slow variables, since a low-volatility environment itself drives leverage accumulation. This circularity is a common challenge in the analysis of financial systems and can be partly mitigated by introducing indicators exogenous to volatility (such as the ratio of open interest to market capitalization or the number of active on-chain addresses). To understand why high-volatility periods tend to persist, one must decompose the factors that drive market volatility across different time scales. In this analytical framework, the most fundamental level consists of macro and systemic factors that change relatively slowly. These factors, termed "slow variables," do not directly generate instantaneous price jumps, but they constitute the underlying environment in which the market operates and determine how sensitive the market is to external shocks. The mechanism through which slow variables operate is to set the activation threshold of the feedback loops: within a safe range, even a large information shock can be smoothly absorbed by the market; within a dangerous range, a minor disturbance may trigger a violent liquidation cascade and liquidity exhaustion.

23.2.1 The macroeconomic environment

The macroeconomic environment is the outermost slow variable that determines the overall tone of crypto-asset markets. As a typical high-beta asset, Bitcoin and other

cryptocurrencies exhibit extremely high sensitivity of their volatility to changes in global liquidity conditions and investor risk appetite [9]. The interest rate cycle plays a central role in this transmission mechanism. When a central bank is in a rate-hiking cycle or maintains a high-interest-rate environment, the rise in risk-free yields inevitably depresses global investors' overall risk appetite. Against this macroeconomic backdrop, capital tends to withdraw from the high-risk, high-volatility crypto market, thinning overall market liquidity. A systemic decline in liquidity means that, under an order-flow shock of the same magnitude, asset prices deviate more sharply, thereby raising the baseline volatility level at the macro level [10]. The transmission strength of the interest rate channel to the crypto market is time-varying, however, and may be weaker than commonly assumed. During 2023–2024, BTC still rose sharply even as the Federal Reserve maintained the highest interest rate level in two decades. This counterexample suggests that crypto-native factors, such as ETF inflows and the halving narrative, can override interest rate transmission in particular periods. The macro variable that more directly affects crypto-market liquidity may be the total amount of USD liquidity—including the size of the Federal Reserve's balance sheet and the global M2 supply—rather than the policy rate level itself.

Beyond traditional macroeconomic indicators, the narrative cycles unique to the crypto market constitute another dimension of slow variable. In stages dominated by a bull-market narrative, optimistic market sentiment prompts traders to make widespread use of leverage. Although this sentiment-driven leverage accumulation manifests initially as steady price increases and positive volatility, it in fact accumulates structural fragility within the system. Conversely, in a bear-market narrative, persistent negative expectations lead market makers and arbitrage capital to scale back their operations, and market depth and resilience decline markedly. At this point, any negative news is more likely to trigger a disproportionate price decline in an environment lacking buy-side support, manifesting as strong negative volatility clustering [11].

Macroeconomic uncertainty—such as geopolitical conflict or a sudden shift in inflation expectations—further reinforces this slow-variable effect. When macroeconomic uncertainty rises, cross-market arbitrageurs and macro hedge funds typically reduce their risk exposure in the crypto market. This precautionary withdrawal of capital not only reduces the market's effective liquidity but also raises the risk aversion of all participants. Thus, although changes in the macroeconomic environment unfold slowly on a monthly or quarterly scale, they set the baseline for the low-frequency component of volatility and determine the market's initial resistance when facing sudden events.

23.2.2 The level of systemic leverage

In crypto derivatives markets, the level of systemic leverage is the key internal slow variable that determines the intensity of volatility clustering. As the earlier analysis of leverage endogeneity noted, the evolution of leverage over the market cycle is not random but follows a self-reinforcing logic. During a calm period of persistently rising prices, the margin value of existing positions increases accordingly, so the effective leverage ratio naturally declines. This superficial sense of safety prompts traders to establish new leveraged positions or to raise the leverage multiples of existing ones, so

that a massive volume of open interest slowly accumulates at the system level.

Figure 23-2 illustrates the nonlinear relationship between the level of systemic leverage and volatility amplification. Leverage does not directly generate volatility; its core role is to define the trigger distance of a liquidation cascade.

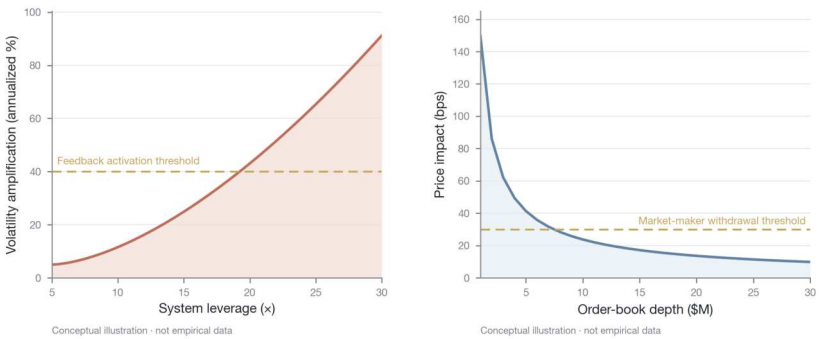


Figure 23-2. How slow variables set the activation threshold of feedback loops (Drawn by the author based on this chapter's analytical framework; conceptual illustration, not empirical data)

In a high-leverage environment, an initial price decline of merely 2% can reach the liquidation lines of a large number of highly leveraged long positions. Once the liquidation engine is activated, forced market sell orders indiscriminately hit the order book, driving prices down further and, in turn, triggering the liquidation of more positions that were originally in the safe zone [12]. This low trigger threshold, set by the accumulation of leverage, explains why, in certain market states, a trivial piece of external news can trigger volatility as high as 10% or even more extreme. The persistent climb of leverage during a bull market shortens the liquidation trigger distance, so that a minor price disturbance can set off the positive feedback loop of volatility clustering.

23.2.3 Liquidity depth

If the level of leverage determines the scale of potential selling pressure, then liquidity depth determines the market's capacity to absorb it. In the crypto market, the baseline level of liquidity is not constant but is constrained by the capital-deployment cycles of professional market makers, constituting a core slow variable. Market makers' capital-allocation decisions are typically based on a long-term risk-return assessment, including funding costs, investment in technical infrastructure, and expectations about overall market trading activity. This adjustment of capital deployment is a slow process, so the market's apparent depth remains relatively stable over weeks or even months.

Changes in liquidity depth directly alter the price impact function. At year-end, during holidays, or in prolonged market downturns, market makers tend to systematically reduce their capital deployment across exchanges, thinning the order book. Against this slow-variable backdrop of scarce liquidity, even routine-scale institutional rebalancing or a moderate information shock can produce far greater price slippage than usual, owing to the lack of sufficient counterparties, which manifests statistically as a rise in volatility [13].

A particularly destabilizing interaction also exists between liquidity depth and the level of systemic leverage. Figure 23-3 reveals how this interaction shapes the market's most fragile state: when high leverage (enormous potential liquidation selling pressure) and thin liquidity (extremely weak absorptive capacity) appear simultaneously, the market enters a highly unstable danger zone in which liquidation orders are more easily triggered and encounter enormous slippage during execution, accelerating the spread of the liquidation cascade [14].

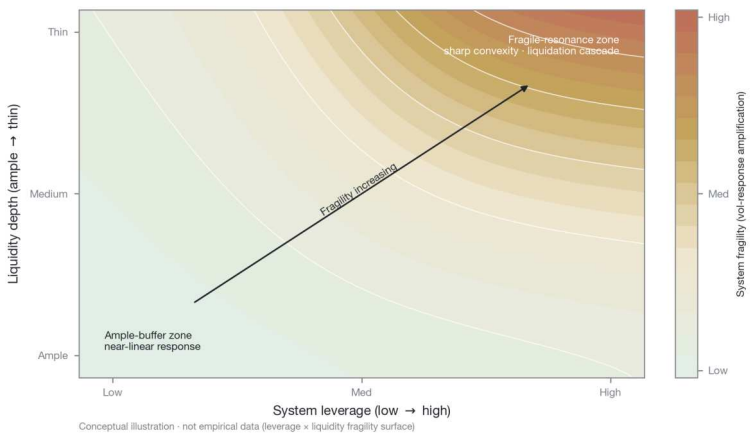


Figure 23-3. The two-dimensional fragility surface of the interaction between leverage level and liquidity depth: in the lower-left region (low leverage, ample liquidity) the volatility response is approximately linear, whereas in the upper-right region (high leverage, thin liquidity) the convexity increases sharply (Drawn by the author based on market microstructure theory; conceptual illustration, not empirical data)

The core message of Figure 23-3 is that systemic fragility is not determined linearly by any single slow variable but is driven nonlinearly by the interaction term between leverage level and liquidity depth. This interaction is closer to a threshold interaction than to simple multiplication: when the systemic leverage ratio is below a certain critical level, the effect of liquidity depth on fragility is approximately linear and mild; once the leverage ratio crosses that threshold, the sensitivity of fragility to liquidity increases convexly and sharply, because high leverage shortens the liquidation trigger distance, so that a liquidity gap of the same magnitude causes disproportionate price displacement and liquidation volume. This feature explains why monitoring risk indicators along a single dimension is often insufficient to capture the system's true fragility.

23.2.4 The pre-programming of fragility

Taking the macroeconomic environment, leverage level, and liquidity depth together yields a core insight about volatility clustering: slow variables do not directly generate short-term violent volatility; their mechanism is to "pre-program" the market's fragility, thereby determining how low the threshold for activating the feedback loops is. This mechanism can be described as a state-dependent nonlinear response.

During periods when the slow variables are in the safe range—that is, when the macroe-

conomic environment is accommodative, systemic leverage is moderate, and market-maker liquidity is ample—the market possesses a strong shock-absorbing capacity. At such times, even when major negative news breaks or a large-scale selling shock hits, ample liquidity can smoothly absorb these orders, and the low leverage ratio ensures that a short-term price decline does not trigger a chain of forced liquidations. In this state, the volatility induced by a shock is isolated, its energy decays rapidly, and volatility does not exhibit pronounced clustering.

However, as the slow variables gradually shift into the danger range, the market microstructure undergoes a qualitative change. Using the joint time series of the level of systemic leverage and conditional volatility, Figure 23-4 depicts the complete cycle in which the "fragility gap" widens continuously during the calm period and is then violently released once triggered.

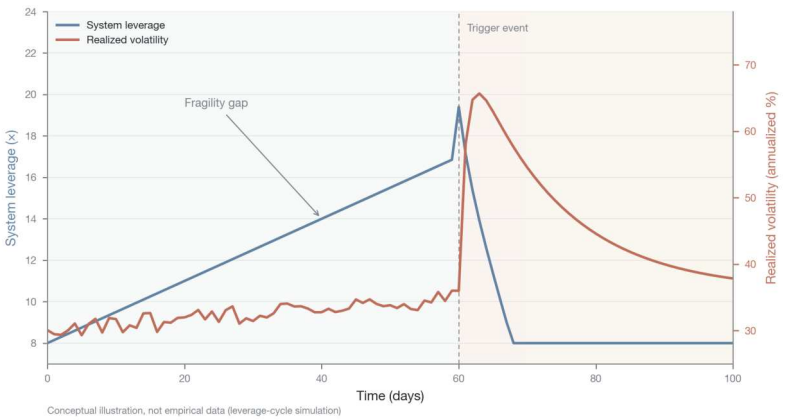


Figure 23-4. The time-series relationship between the level of systemic leverage and volatility (Simulated by the author based on leverage-cycle theory; core parameters: α=0.08, β=0.90, with systemic leverage in the figure ranging roughly from 8× to 19×; illustrative simulation, not empirical data)

In this highly fragile state, what triggers a crisis is often not a large-scale black-swan event but merely a routine information shock or a market order slightly larger than usual. This minor disturbance is enough to break through the feedback-activation threshold that the slow variables have already driven to an extremely low level, triggering the liquidation engine and driving market makers to retreat. Once the self-reinforcing cycle of liquidity withdrawal and liquidation selling begins to operate, volatility no longer depends on the input of external information but is self-perpetuated and amplified by the market's internal microstructural mechanisms. Volatility clustering is therefore not a purely random phenomenon but the inevitable result of slow variables gradually accumulating fragility along the time dimension and finally releasing it violently through feedback loops under a minor trigger. The specific path by which slow variables evolve toward the danger range during low-volatility periods is analyzed in detail in Section 23.6.

23.3 Positive feedback loops and shock amplification

After slow variables have set the market's fragility baseline, the key factor determining whether volatility clustering actually occurs is the activation of positive feedback loops. This section analyzes three positive feedback channels—between volatility and liquidity, between volatility and leverage, and between volatility and market behavior—to explain how an initial shock produces persistent volatility clustering along the time series through microstructural transmission.

23.3.1 The intertemporal transmission of shocks

The analysis in Chapter 22 identified four main sources of volatility: information shocks, liquidity frictions, microstructural mechanisms, and behavioral biases. This cross-sectional perspective explains, at any given moment, which factors jointly push up market price volatility. However, to understand why volatility exhibits pronounced clustering along the time series, one must shift the analytical perspective from a static decomposition of sources to a dynamic time-series evolution. The core of volatility clustering is that an initial external shock not only has an impact in the current period but also, through the market's internal transmission mechanisms, alters the market state of the next period, so that a subsequent shock of the same magnitude can trigger larger price movements.

This time-series self-amplification mechanism constitutes the microfoundation of volatility memory. When an information shock in period t causes violent price movement, market makers contract their liquidity supply because of the rise in inventory risk and adverse-selection risk (the mechanism is detailed in Section 23.3.2), pushing the market liquidity of period t+1 below the normal level. In a liquidity-scarce environment, even if only routine-scale order flow appears in period t+1, it produces a far larger price impact than usual owing to insufficient market depth, thereby realizing higher volatility [15]. This transmission chain clearly explains the microeconomic meaning of the β coefficient in the GARCH model: the conditional variance depends on past shocks because past shocks, by altering the behavior of market participants and the state of the system, substantially reduce the market's capacity to absorb new shocks.

23.3.2 The liquidity loop

Among all the mechanisms that amplify volatility, the feedback loop between volatility and liquidity has the shortest transmission path. This loop is essentially the manifestation, along the volatility dimension, of the reflexivity triangle discussed in Chapter 20. The classic model of Kyle (1985) [16] defines the price impact coefficient λ as the marginal effect of order flow on price, while the illiquidity measure proposed by Amihud (2002) [17]—the ratio of the absolute value of daily returns to trading value—provides a standardized measure for cross-market liquidity comparison. Both classic indicators require a revised reading in crypto perpetual futures markets, however. The Kyle lambda derives from an equilibrium model of a single risk-neutral market maker facing informed traders, which differs structurally from the crypto environment of multiple competing market makers who can cancel orders at any time; and the Amihud ratio is

highly sensitive to the choice of aggregation window (hourly, daily, weekly) in a 24/7 continuous-trading market. In practice, the decay rate of order-book weighted depth—for example, how the ratio of depth in the $\pm 0.5\%$ and $\pm 2\%$ price ranges changes with volatility—may be a liquidity measure more native to perpetual futures. Nevertheless, the core mechanism that the Kyle lambda captures—that the marginal price impact of order flow rises nonlinearly as liquidity declines—still holds in the crypto market. When market volatility rises, λ rises nonlinearly, and the cost structure faced by market makers changes nonlinearly. According to inventory models and information-asymmetry theory, market makers' cost function exhibits pronounced convexity when volatility rises [18]. Market makers' liquidity withdrawal unfolds along two causal paths. First, market makers actively widen their quotes and withdraw quoted depth because of the rise in inventory risk and adverse-selection risk; this is a rational response based on a risk-return trade-off. Second, market makers are forced to reduce their exposure because their own margin constraints tighten, and in extreme cases this even triggers the liquidation of the market makers' own positions. In crypto perpetual markets, because market makers bear no obligation to remain present, the response along the first path is extremely fast (millisecond-level order cancellation) and is usually the primary driver of liquidity evaporation. The common result of both paths is that market makers sharply reduce quoted depth across all levels of the order book.

The withdrawal of liquidity has a direct and significant effect on market microstructure. As the order book thins, the price impact coefficient faced by a market order of the same magnitude increases severalfold. During extreme volatility, the bid-ask spread can rapidly widen from a normal 3 basis points to more than 30 basis points, and the market impact coefficient can even be amplified tenfold [19][20]. This large price jump caused by liquidity exhaustion is statistically recorded as a further rise in realized volatility. This rise in secondary volatility contains no new fundamental information; it is purely a mechanical result of the deterioration in market microstructure and the decline in liquidity-bearing capacity. Because the process by which market makers reassess risk and restore liquidity supply is typically slow and cautious, this liquidity-scarce state persists, so the volatility amplification effect continues along the time series and ultimately manifests as the volatility clustering we observe in the data.

The foregoing analysis of the liquidity loop implicitly assumes a single exchange and a single asset, but the actual transmission network is far more complex. BTC perpetual futures trade simultaneously on more than a dozen exchanges, each of which uses a composite index incorporating prices from other platforms as its mark price. When a liquidation cascade on one exchange causes a sharp price drop, its price deviation is transmitted through the index price to other platforms, triggering their liquidations even if their own order books have not been directly hit. Under a unified margin account, losses on a BTC perpetual position can also directly cause margin shortfalls for positions in ETH, SOL, and other instruments within the same account, forcing traders to close out non-BTC instruments to raise liquidity. The cross-market liquidity spiral described by Brunnermeier and Pedersen (2009) [21] is therefore especially pronounced in crypto derivatives markets: during the May 2021 event, the daily correlation between BTC and ETH rose markedly (climbing above 0.9 in the crisis period), a direct manifestation of this cross-instrument contagion.

The liquidity loop does not operate in isolation but is nested with the leverage loop and the behavioral loop. Figure 23-5 integrates these three feedback channels into a unified conceptual network, showing how an initial shock is transmitted simultaneously along multiple paths and reinforced at each node.

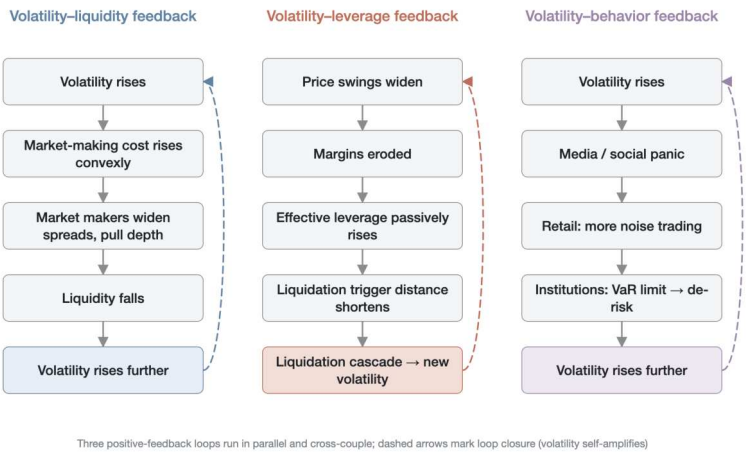


Figure 23-5. Conceptual diagram of volatility feedback loops (Drawn by the author based on this chapter's analytical framework; conceptual mechanism illustration, not empirical data)

The key feature of Figure 23-5 is that multiple cross-coupling points exist among the three loops. The withdrawal of liquidity simultaneously lowers the execution quality of liquidation orders (coupling with the leverage loop) and intensifies the visual impact of price jumps (coupling with the behavioral loop), while the panic selling of retail investors in the behavioral loop further consumes the already-thin liquidity. This multi-channel, cross-coupled network structure makes the volatility amplification effect far greater than any single loop can explain. The following two subsections analyze the independent transmission mechanisms of the leverage loop and the behavioral loop respectively.

23.3.3 The leverage loop

The liquidity feedback loop explains the slow decay of volatility, whereas the leverage feedback loop explains the violent short-term eruption of volatility. In perpetual futures markets, high leverage is not merely a tool of capital efficiency but a core source of systemic fragility. When an initial rise in volatility causes a large deviation in the asset price, the margin balances of traders holding highly leveraged positions are rapidly eroded. As net value declines, the effective leverage ratio of these positions rises passively, bringing them ever closer to the threshold that triggers forced liquidation [22].

This passive rise in effective leverage significantly increases the probability of liquidation across the entire system. Once the price reaches a key liquidation line, the exchange's liquidation engine automatically takes over the position and sends a large volume of market orders to the book. This forced, extremely price-insensitive order flow floods into an already fragile order book, instantly breaking through multiple layers of liquidity and driving a further plunge or spike in price. This violent price movement driven by

liquidation generates additional volatility, which in turn triggers deeper liquidations, forming the classic liquidation cascade [14]. This is a positive feedback mechanism whose operation requires no external information input and is driven entirely by the internal leverage-clearing demand of the system. Exchanges have designed multiple circuit-breaker mechanisms to curb this cascade, including tiered margin systems (large accounts face higher margin rates to suppress concentration), auto-deleveraging (ADL) mechanisms (which preferentially offset profitable counterparties rather than dumping everything at market price), and insurance funds (which absorb bankruptcy losses to avoid socialized loss allocation). In extreme conditions, however, these mechanisms often fail to curb the cascade effectively—because the insurance fund is exhausted within minutes, auto-deleveraging triggers panicked withdrawal by counterparties, and tiered margin adjusts with a lag amid successive price gaps—so that positive feedback ultimately still dominates market dynamics. This feedback loop does not terminate naturally until all these circuit-breaker mechanisms have failed, or until leverage has been fully released and fragile positions have been completely unwound. Because a large-scale deleveraging process often requires several consecutive days of liquidation activity to complete, and each day's liquidations create new volatility in the market, the duration of deleveraging overlaps highly, along the time series, with the period of extreme volatility clustering.

23.3.4 The behavioral loop

In crypto-asset markets, the high share of retail investors gives the emotional contagion and herd behavior studied in behavioral finance a more prominent role in volatility transmission. When market volatility rises sharply, it is often accompanied by a surge in media coverage and the rapid spread of panic on social networks. This deterioration of the information environment triggers investors' loss aversion and herd mentality, leading large numbers of noise traders lacking fundamental support to flood into the market and engage in panic selling or momentum chasing [23]. This irrational trading behavior not only fails to provide liquidity but consumes precious market depth, further pushing up the market's realized volatility. At the same time, not all retail investors choose to sell in a decline. "Buying the dip" is a deeply entrenched behavioral pattern in the crypto community, and some traders add leverage to buy in as prices fall. Research shows that the loss aversion and aggressive trading behavior of Bitcoin derivatives traders under liquidation risk systematically amplify the tail losses of hedged portfolios in extreme markets [24]. Such behavior can slow the pace of the rise in volatility in the short term, but it ultimately converts these newly established highly leveraged positions into fuel for the next wave of liquidations, making the delayed cascade larger in scale.

At the institutional-investor level, behavioral feedback takes a more institutionalized and mechanical form: the value-at-risk (VaR) spiral. Most institutional investors and market makers are subject to strict VaR risk-control models. When market volatility rises, the VaR value computed from historical data soars accordingly, causing the institution's risk exposure to rapidly approach or breach its internally set risk limit. To meet compliance requirements and internal risk-control standards, these institutions are forced to cut their risk exposure simultaneously—that is, to sell risk assets—precisely when market liquidity is worst. This procyclical de-risking behavior generates enormous

selling pressure, directly causing further price declines and a continued rise in volatility, which in turn triggers a new round of VaR breaches and forced liquidations [21]. In practice, only a few leading market makers and hedge funds in the crypto market are truly bound by strict VaR models. A larger share of trading volume comes from retail and semi-professional traders who use simple stop-losses or fixed position limits, and whose de-risking is driven more by a pain threshold than by a model. Nevertheless, whether it is retail panic selling or institutional mechanical de-risking, both reinforce the self-propagation of volatility along different dimensions.

23.3.5 Activation and termination conditions

Feedback loops are not activated in every market state; their activation threshold is determined by the slow-variable dynamics of the first layer of explanation (Section 23.2.4). When the slow variables are in the safe range, the threshold is high, and the market can absorb a large information shock without triggering systemic positive feedback; once a high leverage ratio or excessive liquidity concentration has caused fragility to accumulate sufficiently, the threshold drops markedly, and a routine-scale shock is enough to trigger the feedback loops.

Once a feedback loop is activated, its termination exhibits pronounced lag and asymmetry: the end of volatility clustering depends on the exhaustion of the momentum internal to the feedback mechanism, and this exhaustion is far slower than the activation (the bidirectional micro-mechanisms of entering and exiting the high-volatility regime are detailed in Section 23.4.3). The leverage loop must wait for large-scale liquidations to complete and the systemic leverage ratio to fall substantially; the liquidity loop must wait for market makers to return cautiously at a higher risk premium; and the termination of the behavioral loop lags especially. This lag arises not because the same participants gradually become desensitized, but because the price-sensitive seller cohort has already been flushed out during the earlier liquidations and panic. The remaining holders either have already accepted their paper losses (refusing to realize losses under the disposition effect) or are new entrants who built positions at lower prices, so the composition of the market crowd has undergone a structural change. Because capital rebuilding, confidence recovery, and the resetting of risk-control parameters all take time, the termination of a feedback loop is far more protracted than its activation. This explains the sharp-rise, slow-decay asymmetric shape of volatility clusters in real data. This asymmetry also reveals a structural blind spot of the standard GARCH model: its β coefficient imposes a symmetric decay structure on upward and downward shocks, which systematically underestimates the speed of the transition from calm to crisis while overestimating the speed of recovery from crisis to calm. Asymmetric GARCH variants—such as the GJR-GARCH proposed by Glosten, Jagannathan, and Runkle (1993) or the EGARCH—can better capture this dynamic [25], which will be discussed further in the selection of volatility forecasting models in Chapter 24.

23.4 The state machine and volatility regime transitions

The interaction between slow variables and feedback loops makes volatility change not continuously and smoothly but with discrete state jumps. Although continuous

GARCH-type models can fit the persistence of volatility, they cannot capture the market's abrupt shifts between calm and crisis. This section introduces a Markov regime-switching model that abstracts the volatility dynamics of perpetual futures markets into nonlinear transitions among three discrete states—calm, stressed, and crisis—and analyzes the asymmetry of state transitions and their interaction with slow variables.

23.4.1 The model framework

In traditional financial econometrics, volatility clustering is usually modeled as a continuously evolving process. The GARCH model and its variants form the foundation of this analytical paradigm, whose core assumption is that the conditional variance at the current moment is a continuous function of past return shocks and past variances. Within this framework, volatility slides smoothly between "high" and "low," and its response to new information takes the form of a gradual adjustment of variance. However, the empirical data of cryptocurrency perpetual futures markets reveal a markedly different dynamic character: changes in volatility are often not continuous but exhibit pronounced jumpiness and state-like behavior. The market may hold a low annualized volatility of 30% for weeks, then jump above 120% within hours on a single liquidation event, *sit* at this high level for several days, and then switch states again. This multimodal feature in the frequency distribution suggests that a single continuous process is insufficient to capture the abrupt changes in market microstructure; volatility in fact switches among several discrete regimes [7].

To describe this nonlinear regime switching precisely, the Markov regime-switching model provides a rigorous statistical framework. Hamilton (1989) [26] first introduced the Markov chain into the analysis of macroeconomic time series, proposing that the generating process of an observed variable depends on an unobservable discrete state variable. When this idea is extended to volatility modeling, it forms the Markov regime-switching conditional-heteroskedasticity model [27][28]. In this model, the market is assumed to be in one of a finite number of discrete states, each with its own mean, variance characteristics, and mechanism for responding to shocks. Let the unobservable state variable $S_t \in \{1, 2, 3\}$ correspond to the three states of calm, stressed, and crisis, respectively, with transitions among states governed by a 3×3 transition probability matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

where $p_{ij} = \Pr(S_t = j \mid S_{t-1} = i)$, and the elements of each row sum to 1. The conditional return distribution in each state follows a normal distribution with different parameters:

$$r_t \mid S_t = k \sim \mathcal{N}(\mu_k, \sigma_k^2), \quad k = 1, 2, 3$$

where $\sigma_1^2 < \sigma_2^2 < \sigma_3^2$, corresponding to the increasing conditional variances in the calm, stressed, and crisis states. Throughout this chapter, the volatility metric is realized volatility (returns aggregated intraday at a 5-minute frequency), annualized

by the crypto-market convention of continuous trading 365 days a year—that is, the daily volatility is multiplied by $\sqrt{365}$—rather than the traditional-market convention of $\sqrt{252}$. The current state depends only on the previous state, which is precisely the core embodiment of the Markov property. By introducing discrete states, the model can explain why in some periods the market has an extremely high capacity to absorb shocks, while in other periods a shock of the same magnitude triggers a violent surge in volatility.

Applying the Markov regime-switching model to perpetual futures markets rests intuitively on the nonlinear mapping between the continuous change of slow variables and the threshold activation of feedback loops. Slow variables such as macro liquidity depth and the systemic leverage ratio undergo slow and continuous evolution in daily trading, and these changes do not by themselves directly manifest as a rise in volatility. However, the feedback loops in the microstructure—particularly the liquidation cascade and the market-maker order-cancellation mechanism—have a distinct threshold-trigger character. When the accumulation of slow variables pushes systemic fragility past a certain critical point, an ordinary information shock that could originally have been smoothly absorbed instantly activates the positive feedback loop. Once the feedback loop is activated, the market microstructure undergoes a qualitative change, and the system "jumps" from one state into an entirely new one. In the new state, high-frequency liquidation events and widened bid-ask spreads reinforce each other, allowing the market to sustain itself within this high-volatility regime until the internal imbalance is fully digested. This discrete jump, driven by continuous slow variables and triggered via a threshold mechanism, constitutes the third layer of explanation for understanding volatility clustering in perpetual futures.

23.4.2 The transition matrix

Based on empirical estimation with a hidden Markov model (HMM), the volatility dynamics of perpetual futures markets can be abstracted into a three-state Markov machine: the calm state, the stressed state, and the crisis state. The three-state division is a deliberate simplification of the actual market continuum. The internal risk-control systems of major exchanges typically use a finer-grained classification of five to six market states (including intermediate states such as early warning and price-protection triggering); this chapter adopts three states to strike a balance between analytical simplicity and the ability to capture the phenomenon. These three states differ significantly not only in their statistical distributions but also in the relative weights of their underlying microscopic drivers. The calm state is the market's norm, occupying roughly 65% to 75% of trading time. In this state, annualized volatility typically remains in the range of 20% to 40%, and the return distribution is approximately normal with thin tails. The microstructure is characterized by extremely narrow bid-ask spreads, liquidation volume at a trough, and smoothly functioning cross-exchange arbitrage. At this point, volatility is driven mainly by exogenous information shocks, with information sources accounting for more than 60% of total volatility, and traditional continuous volatility models have good predictive power in this regime.

When market fragility accumulates and a specific shock strikes, the system jumps to

the stressed state, which accounts for about 20% to 30% of the time. The stressed state is typically characterized by annualized volatility rising to 50% to 80%, with the return distribution beginning to show pronounced fat tails and skewness. At the microscopic level, market makers perceive increased risk and begin to passively widen spreads and reduce order-book depth; at the same time, some highly leveraged positions approach the liquidation line, causing liquidation volume to rise initially. In this regime, the contributions of liquidity sources and mechanistic sources (such as funding-rate pulses) to volatility increase significantly, together accounting for 40% to 60%. If the shock is not relieved and liquidation volume breaks through a critical threshold, the market moves further into the crisis state. Although the crisis state accounts for only 3% to 8% of total time, its annualized volatility often soars above 100% to 200%, accompanied by extreme fat tails and frequent price jumps. At this point, the liquidation cascade erupts fully, liquidity is nearly exhausted, mechanistic sources completely dominate market dynamics, and the relative importance of information shocks falls to its lowest level.

Figure 23-6 integrates these three states and their transition logic into a complete state-machine schematic.

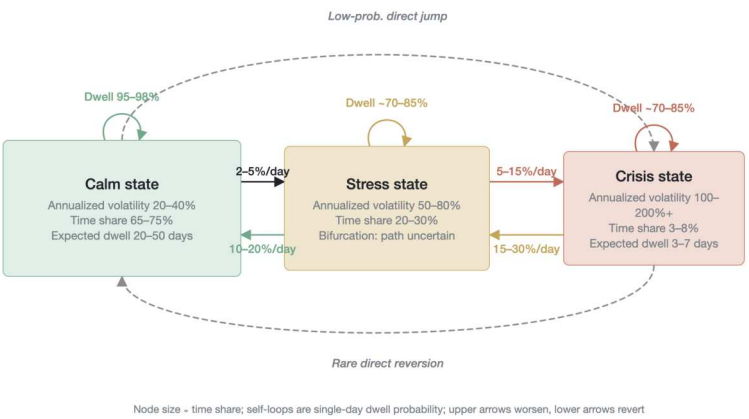


Figure 23-6. Schematic of the volatility state-transition machine for perpetual futures (Drawn by the author based on a Markov regime-switching model [26]; the transition probabilities and time shares are the author's HMM estimates on daily data for Binance BTC-USDT perpetual futures from 2019 to 2024, with wide confidence intervals, not empirically measured)

Two key structural features can be identified intuitively from Figure 23-6. First, the self-loop probability of the calm state (the probability of remaining in the current state) is the highest, indicating that the low-volatility regime has the strongest capacity for self-maintenance. Second, the probabilities of the two paths departing from the stressed state (regression toward calm and deterioration toward crisis) are not far apart, which means that the stressed state is a highly uncertain bifurcation point whose ultimate direction depends on the specific configuration of the slow variables at that moment.

Transitions between states are not a random walk but follow a specific set of empirical probability matrices. According to the HMM parameter estimates on historical high-frequency data (the following ranges are based on the same HMM parameter estimates

listed in the figure note above), the daily probability of the market jumping from the calm state to the stressed state is typically between 2% and 5%; once in the stressed state, the probability of further deterioration to the crisis state rises significantly, reaching 5% to 15%. This means that the stressed state is a highly unstable transitional period that either evolves rapidly into a full-blown crisis or returns to calm after turbulence. The probability of regression from the crisis state to the stressed state is relatively high, at about 15% to 30%, reflecting the fact that extreme liquidation events typically exhaust their momentum within a few days; however, the probability of complete regression from the stressed state to the calm state is only 10% to 20%. This parameter structure of the transition matrix indicates that the calm state has the strongest regime stickiness, with an expected dwell time of 20 to 50 days, whereas the crisis state has the shortest expected dwell time, typically only 3 to 7 days, though the microstructural damage it causes is prolonged through a longer stressed state.

The state-machine parameters above are estimated from daily data on a single trading pair (Binance BTC-USDT perpetual) and have three limitations. First, a single-pair model cannot capture state transitions transmitted from other exchanges or other assets; for example, the 2022 FTX collapse first hit FTT and SOL before being transmitted to BTC perpetual futures. Second, daily data may mask state transitions at the intraday microstructural level: a trading day classified as "calm" by a daily-frequency model may contain a localized crisis at the 15-minute level (such as a flash crash followed by rapid recovery). Third, the effective number of observations for some state-condition combinations (such as the low-liquidity range within the stressed state) is limited, so the confidence intervals of the transition-probability estimates may be wide. Multi-exchange joint HMM estimation or extended estimation on hourly data can mitigate these limitations to some extent, but this is beyond the scope of this chapter.

23.4.3 Transition asymmetry

In the three-state volatility model, a prominent dynamic feature is the pronounced asymmetry of state transitions: the market enters a high-volatility regime far faster than it exits one. Empirical data show that falling from the calm state, through the stressed state, fully into the crisis state often takes only 1 to 3 days, and in extreme cases the jump can be completed within hours; whereas cooling down from the crisis state, through the stressed state, and finally returning to a truly calm state typically requires a repair period of 5 to 20 days or even longer. This large difference in time scale is what gives volatility clusters their sawtooth shape—sharp rise, slow decay—on time-series charts.

By contrasting the typical time paths of entering and exiting the high-volatility state, Figure 23-7 illustrates this asymmetry intuitively.

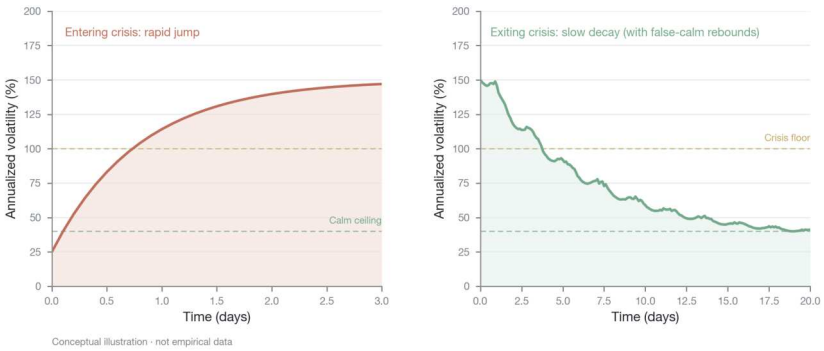


Figure 23-7. Comparison of the temporal asymmetry of volatility state transitions (Drawn by the author based on this chapter's analytical framework; conceptual illustration, not empirical data)

This macro-statistical asymmetry is rooted in the asymmetry of the behavioral patterns of micro-level market participants and of mechanism design. Entering a high-volatility state is usually initiated by a single triggering event. When an information shock exceeding expectations strikes a fragile market that has already accumulated high leverage and thin liquidity, it immediately triggers the automated liquidation engine. The high-leverage nature of perpetual futures causes liquidation orders to hit the order book as market orders regardless of cost, instantly draining buy-side or sell-side liquidity. This algorithmically forced selling is not constrained by human hesitation or wishful thinking, and its transmission speed is limited only by the processing capacity of the exchange's matching engine. At the same time, once market makers' risk-control models detect the jump in volatility and the loss of control over inventory risk, they automatically cancel orders or sharply widen spreads at millisecond speed. The positive feedback of automated liquidation and the instantaneous withdrawal of machine market makers superimpose on each other, so that the market completes its jump to the crisis state in an extremely short time.

By contrast, exiting a high-volatility state is a protracted process involving the psychological reconstruction of multiple parties and the redeployment of capital. First, although the liquidation cascade in the crisis state is fierce, it must wait for excessively leveraged positions to be fully and *naturally digested*, and this deleveraging process takes time to find a new price equilibrium. Second, after experiencing extreme volatility, market makers' VaR models significantly raise their expectations of future volatility, causing them to exercise great caution when resupplying liquidity. Redeploying capital must pass through reassessment and approval by risk-control departments, and market makers typically return tentatively and gradually with wider spreads and smaller sizes, rather than restoring liquidity depth to pre-crisis levels all at once. Finally, the emotional recovery of market participants likewise exhibits a lag effect: the memory of panic brought by extreme volatility causes traders to lower their leverage and reduce their trading frequency for a considerable period. These slow micro-level repair mechanisms jointly determine that the regression from a high-volatility regime to a low-volatility regime must be a gradual, repeatedly interrupted decay process.

23.4.4 The regulation by slow variables

Although the Markov regime-switching model provides an elegant framework for describing the discrete jumps of volatility, treating the transition-probability matrix as a fixed constant would ignore deeper market dynamics. In fact, the state machine does not operate in isolation in a vacuum; its internal transition probabilities are time-varying, and what determines these dynamic changes in probability is precisely the "slow variables" described in the first layer of explanation. The three major slow variables—the macroeconomic environment, the level of systemic leverage, and liquidity depth—constitute the underlying physical environment in which the state machine operates, and by regulating the sensitivity of the microscopic feedback loops, they directly determine the specific values of the individual elements of the transition matrix.

The level of systemic leverage is the most critical slow variable determining the transition probability from the calm state to the stressed state. During a prolonged calm period, the low-volatility environment entices traders to continually raise their leverage multiples in pursuit of capital efficiency, causing the system's overall effective leverage ratio to climb gradually. Once a high-leverage environment has formed, the market's tolerance for adverse price movement drops sharply. At this point, a minor information shock that would originally have been insufficient to cause a ripple can reach the liquidation lines of a large number of highly leveraged positions. Therefore, as the systemic leverage ratio rises, the transition probability from the calm state to the stressed state increases sharply and nonlinearly. High leverage lowers the market's tolerance for price disturbances to an extremely low level, so that a routine-scale information shock can trigger a liquidation cascade.

Liquidity depth and macroeconomic uncertainty mainly determine the probability of deterioration from the stressed state to the crisis state. When market-maker capital deployment is insufficient or the market is in a specific low-liquidity holiday period, the absorptive capacity of the order book is extremely fragile. In such a thin-liquidity environment, once the market enters the stressed state and an initial liquidation flow appears, the lack of sufficient counter-liquidity to absorb it causes prices to shift disproportionately and violently, rapidly triggering a larger-scale liquidation cascade. Empirical research shows that during periods when liquidity depth is in the bottom historical quartile, the probability of the market sliding from the stressed state into the crisis state is more than three times that of liquidity-ample periods (this multiple is based on a conditional HMM estimate on the aforementioned Binance BTC-USDT perpetual daily data: after grouping order-book depth by historical quartile, the state-transition matrix of each group is fitted separately, and the inter-group difference in the stressed-to-crisis transition probability is compared; because the number of joint observations with the stressed state and liquidity in Q1 is limited, the 95% confidence interval of this multiple estimate is wide, at approximately 2.1 to 4.8 times, and it should be regarded as an order-of-magnitude indication rather than a precise parameter). At the same time, if the macroeconomic environment is in a high-interest-rate or high-uncertainty cycle, the overall risk appetite for risk assets is suppressed, and all transition probabilities toward high-volatility states are systematically amplified [29]. It is precisely this deep interaction between slow variables and state-transition probabilities that explains why the intensity and frequency of volatility clustering in perpetual futures markets

vary significantly across different macro and micro market cycles.

23.5 The institutional amplifiers of perpetual futures

The preceding three sections established the three-layer explanation in turn: slow variables set the fragility baseline, feedback loops amplify shocks, and the state machine drives regime jumps. However, these mechanisms may exist in any leveraged derivatives market; the reason perpetual futures markets exhibit clustering intensity far exceeding that of traditional markets lies in the systematic amplification that their distinctive institutional design imposes on these three layers. The institutional features are not themselves independent sources of volatility; each of their core components enhances the clustering effect by lowering the feedback-activation threshold or accelerating the feedback transmission speed. This directly explains why the clustering intensity of perpetual futures observed in Section 23.1 is so much higher than in traditional financial markets. The three institutional features—the liquidation engine, funding rates, and around-the-clock trading—together construct an almost frictionless transmission environment, and their synergistic amplification logic is drawn together in Section 23.5.4.

23.5.1 Funding rates

The funding rate is a periodic payment mechanism designed for perpetual futures to maintain the anchoring of the contract price to the spot price (as described in Chapter 10). By design, the funding rate is first of all a convergence mechanism: by charging the deviating side and paying the reverting side, it incentivizes arbitrageurs to enter and correct the deviation, effectively suppressing excessive deviation between the contract price and the spot price in most market states. However, when the market is in a strong one-sided trend and volatility rises sharply, arbitrageurs cannot maintain opposing positions because they face extremely high liquidation risk, the rate's convergence function fails, and the mechanism flips from a stabilizer into an amplifier. The critical condition for this flip corresponds roughly to the funding rate remaining at an extreme value in the same direction for several consecutive settlement periods (for example, exceeding 0.1% for three consecutive 8-hour periods). In the context of volatility clustering, once the flip occurs, the funding rate amplifies the trend within a trend and amplifies volatility at turning points, becoming an endogenous positive feedback engine. In the calm state, the funding-rate pulse is minuscule and its effect on volatility is negligible; in the stressed or crisis state, however, an extreme rate significantly prolongs the duration of high volatility. The crypto spot-futures basis (carry) has been shown to embed amplification dynamics related to crash risk [30].

In a strong one-sided market, the positive feedback loop of the funding rate is especially evident. When long forces dominate the market, the contract price generates a premium, and the funding rate turns positive. In theory, a positive rate should incentivize shorts to enter and arbitrage, thereby depressing the price; but in the high-volatility state of a bull market, shorting faces extremely high liquidation risk, and shorts struggle to maintain positions. The result is that longs continue to accumulate, prices rise further, the funding rate is pushed to a higher level, and this instead attracts

more momentum-chasing speculators. When the market finally reverses, the destructive power of the feedback loop is released instantly. The funding rate swings sharply from an extreme positive value to negative, and the longs that had enjoyed the trend dividend begin to be penalized and are forced to close positions. The closing of long positions drives prices down, further deepening the negative shift of the funding rate, prompting more longs to be forced to close, and ultimately triggering a liquidation cascade.

Figure 23-8 shows the synchronization between funding-rate pulses and volatility clustering: during the market's transition from calm to crisis, the arbitrage-closing behavior triggered by extreme rates translates directly into additional directional volatility, giving volatility stronger persistence.

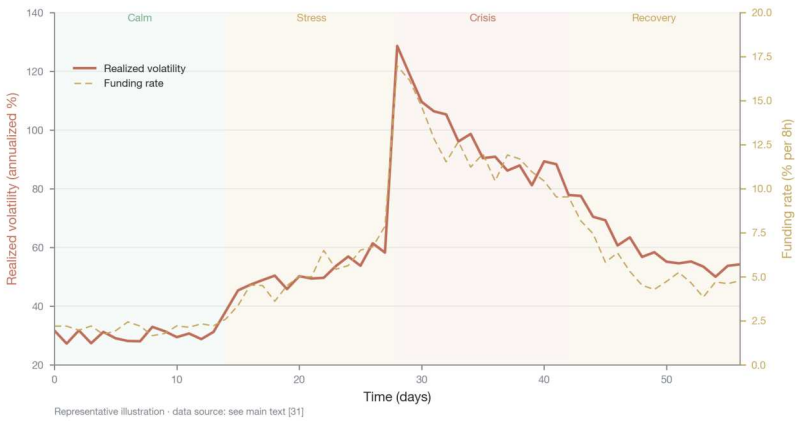


Figure 23-8. The linkage between funding-rate pulses and realized volatility (Representative illustration, synthetic series, not empirical data; the funding-rate axis in the figure is a synthetically amplified series that only illustrates its synchronization with volatility, not the true magnitude of the rate—the true perpetual rate is approximately 0.1%–0.3% per 8 hours, see main text; the volatility × leverage × liquidity framework mechanism is based on a crypto-market derivatives analysis report [31])

This pulse-superposition effect of the funding rate manifests in quantitative models as a higher persistence parameter in the GARCH model. Naimy et al. (2021) showed that, in capturing the volatility clustering characteristics of cryptocurrencies, a GARCH model incorporating asymmetric effects better reflects this mechanism-driven volatility jump, with a persistence parameter significantly higher than in traditional fiat-currency markets [9].

23.5.2 Automated liquidation

The core difference in margin mechanisms between traditional futures markets and perpetual futures lies in the execution speed of liquidation. In traditional delivery futures, when an account's margin is insufficient, the exchange issues a margin call, and the trader typically has a T+1 time window to add funds or close the position independently. This design provides a buffer for the market: liquidation is dispersed, and different traders respond at different points in time. Perpetual futures, by contrast, use

an automated liquidation engine: when the mark price reaches the bankruptcy price, the engine immediately takes over the position and liquidates it at market price, with no manual intervention at any point, and the mark price is monitored at millisecond frequency. From trigger detection to the generation of the liquidation order, this process is indeed completed on a millisecond scale; however, the actual execution of the liquidation order—that is, filling level by level on the order book until the position is fully closed—may last several seconds or longer in extreme markets, because the rapid evaporation of order-book depth forces the liquidation order to seek counterparties at ever more distant price levels. When large numbers of liquidations are triggered simultaneously, the matching engine may become congested or even degrade (for example, in May 2021 the latency of Binance's matching engine spiked to the second level), forming a "liquidation cliff" effect: after the engine recovers, the backlog of liquidation orders is released all at once, causing a price impact far greater than under continuous processing.

This speed difference fundamentally alters the transmission dynamics of volatility. In perpetual futures, liquidation is synchronous and concentrated. The engine simultaneously processes all triggered leveraged positions, and the concentrated liquidation flow directly hits the order book. Because market makers typically widen spreads and withdraw liquidity during high volatility, the order book's absorptive capacity is at its weakest. The liquidation flow creates a deeper price dip in the thin order book, and this new price low immediately triggers the next batch of originally safe positions, forming the classic liquidation cascade.

Using a scatter plot of 24-hour cumulative liquidation volume against contemporaneous realized volatility, Figure 23-9 reflects the positive feedback relationship between the two: the scatter exhibits pronounced nonlinearity, and once liquidation volume breaks through a critical level, the volatility response steepens sharply.

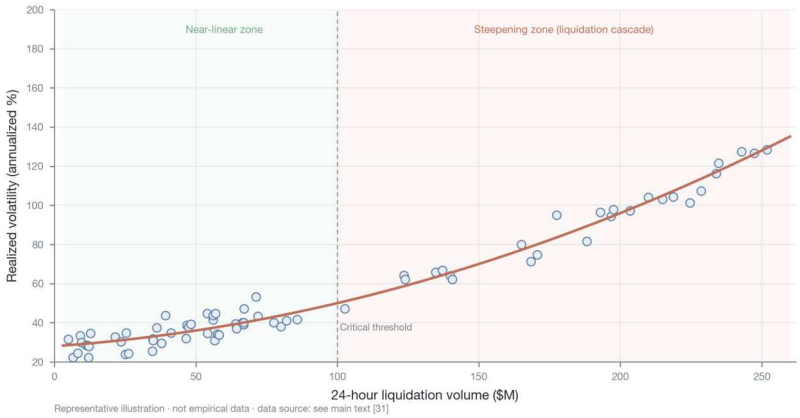


Figure 23-9. The positive feedback effect between 24-hour liquidation volume and realized volatility (Representative scatter illustration, synthetic series, not empirical data; mechanism based on a crypto-market derivatives analysis report [31])

In perpetual futures, leverage routinely reaches 100×, and superimposed on an instan-

taneous liquidation mechanism it gives the positive feedback an extremely high gain. The liquidation volume triggered by a 3% price decline in perpetual futures may be several times that in traditional futures markets. This additional volatility manufactured by the liquidation engine accumulates along the time series and directly causes deep volatility clustering. The liquidation cascade not only creates extreme volatility in the current period but also lays the foundation for high volatility in subsequent periods by thoroughly destroying the liquidity structure. The operational risk of the liquidation engine itself also constitutes an independent amplification channel: when engine overload causes liquidation delays, positions not liquidated in time continue to lose money and accumulate bankruptcy losses, ultimately triggering the exhaustion of the insurance fund and auto-deleveraging; the latter forcibly closes the positions of profitable counterparties, so that even traders who correctly predicted the direction are forced out of the market, further contracting liquidity. The regulatory evolution surrounding liquidation mechanisms is accelerating: for example, the European Union's MiCA framework imposes orderly-operation and risk-management requirements on crypto-asset service providers; some jurisdictions (such as Hong Kong) set caps on retail clients' virtual-asset leverage exposure; and several common-law jurisdictions have begun to examine whether automated liquidation under a manipulated mark price constitutes unfair trading conduct. If these regulatory interventions are fully implemented, they will fundamentally reshape the transmission dynamics of liquidation cascades.

23.5.3 Around-the-clock trading

Traditional financial markets have fixed closing times and weekend market closures. This design objectively acts as a systemic shock absorber. During the market close, market makers have time to reassess risk models and deploy capital, traders can calmly digest information, and algorithmic logic can be manually reviewed. Overnight information shocks are typically reflected as a one-time gap at the next day's open, rather than triggering continuous feedback transmission during the session.

Perpetual futures markets operate around-the-clock, continuous trading, which means the feedback loops are always in an activatable state. Information shocks can be transmitted at any point in time, especially during late-night or weekend hours when liquidity is naturally thin. In these low-liquidity windows, an order-flow shock of the same magnitude produces far greater price movement than during normal hours. Once volatility triggers the liquidation engine or an extreme value of the funding rate, the feedback loop operates continuously in an environment lacking sufficient liquidity buffer.

This continuous-trading mechanism facilitates the cross-time-zone relay transmission of volatility. Volatility originating in the Asian trading session may continue to be amplified in the European session because of insufficient liquidity, and then trigger a larger-scale leverage clearing in the American session. Volatility clustering can therefore cross the closing boundaries of traditional markets, forming high-volatility clusters lasting days or even weeks. The profound impact of around-the-clock trading on market microstructure has attracted wide attention, with related analyses noting that low-liquidity periods such as weekends and late nights have become high-incidence windows for volatility eruptions [32]; traditional exchanges such as CME have also begun to

launch around-the-clock crypto derivatives trading. While amplifying volatility, 24/7 trading also has a mitigating side: continuous trading eliminates the overnight gaps of traditional markets and allows information to be incorporated into prices more smoothly along the timeline; major exchanges have also deployed circuit-breaker-like mechanisms such as price-protection bands and extreme-market trading limits to partly compensate for the systemic fragility of having no market close, though the effectiveness of these mechanisms in extreme markets remains to be verified. Low-liquidity windows also invite strategic exploitation. In *liquidation hunting*, a trader places large directional orders during the thinnest-liquidity periods to trigger a cascade and profit from the resulting dislocation, which makes the fragility of these windows not merely a passive structural feature but an active, adversarial amplification mechanism.

23.5.4 Synergistic effects

The directional amplification of funding rates, the speed amplification of the automated liquidation engine, and the temporal amplification of around-the-clock trading—these three institutional features work together to make the volatility clustering intensity of perpetual futures systematically higher than in traditional markets. Figure 23-10 lays out how the three synergistically reinforce volatility clustering.

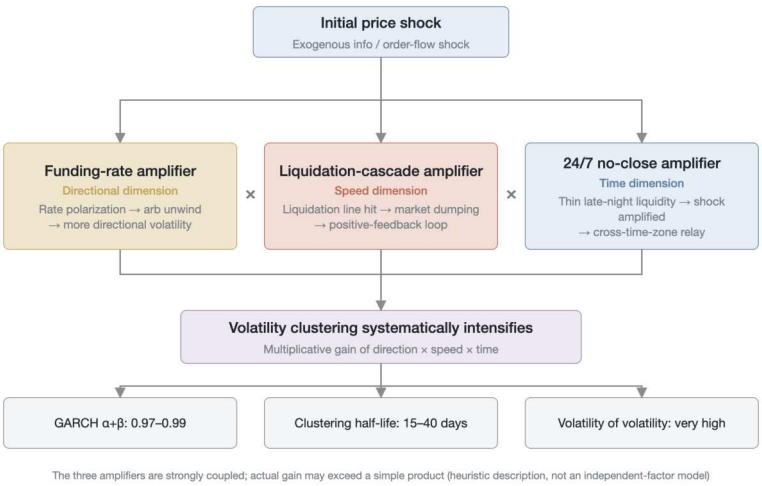


Figure 23-10. The conceptual transmission path of perpetual futures volatility amplifiers (Drawn by the author based on this chapter's analytical framework; conceptual mechanism illustration, not empirical data)

The transmission path in Figure 23-10 reveals the hierarchical relationship among the three amplifiers: the funding rate amplifies the trend and accumulates one-sided positions along the directional dimension, the automated liquidation engine instantaneously converts the fragility of positions into a market-price impact along the speed dimension, and around-the-clock trading eliminates the window for the system to cool naturally along the time dimension. The superposition of the three is not a simple linear summation but exhibits a multiplicative relationship: the larger the one-sided positions accumulated by the funding rate, the more violent the impact released when the liqui-

dation engine is triggered, and around-the-clock trading ensures that this impact can be detonated during the most fragile periods, such as the thinnest-liquidity late-night Asian hours. This "direction × speed × time" multiplicative effect is a heuristic description rather than a strict independent-factor model; strong coupling exists among the three amplifiers (24/7 trading allows funding rates to accumulate continuously, and extreme rates in turn trigger more liquidations), so the actual amplification factor may be higher than the simple product of the three. In addition, the analysis above focuses on the institutional design of centralized exchanges; on-chain order-book decentralized exchanges (DEXs), such as Hyperliquid and dYdX v4, face different but equally significant amplification mechanisms: the block-time constraint limits the liquidation-processing capacity per unit of time, the centralized single-point dependence of the sequencer may become a bottleneck under high load, and the update latency of the price oracle may cause a systematic deviation between the liquidation trigger and the true market price.

Returning to the cross-market comparison data in Table 23-1, the persistence parameter of perpetual futures is systematically higher than in traditional futures and spot markets, its shock-decay speed is the slowest, and its volatility half-life is the longest. The GARCH persistence parameter of traditional fiat-currency pairs is typically markedly lower than that of crypto perpetual futures, with a shorter volatility half-life and a shorter duration of extreme clustering episodes (the author's synthesis of multi-source estimates: fiat persistence of about 0.90–0.95, half-life of about 5–15 days, and clustering episodes of about 1–4 days); this directional difference is consistent with the empirical conclusions of [9].

The extreme volatility-of-volatility observed in perpetual futures markets does not stem from the intrinsic properties of the underlying asset but is the inevitable result of institutional design. If the automated liquidation engine were removed or trading circuit breakers introduced, the intensity of volatility clustering would decline structurally. This conclusion has important policy and design implications: in assessing and managing systemic risk in crypto markets, risk-control models cannot rest on the statistical extrapolation of historical volatility alone but must treat the nonlinear transmission logic of these institutional amplifiers as a core consideration.

23.6 The accumulation of fragility during low-volatility periods

In cryptocurrency perpetual futures markets, a prolonged low-volatility environment is often misinterpreted by market participants as a signal of reduced risk. From the perspective of system dynamics, however, this calm is not true safety but a process in which fragility accumulates covertly within the system. When the market remains in a low-volatility state for an extended period, the behavioral responses of traders, the market-making strategies of liquidity providers, and the parameter updates of risk-management models jointly constitute a self-reinforcing fragility-production mechanism. This mechanism explains why the most intense volatility clusters in cryptocurrency markets often erupt right after the longest calm periods. This section examines the

endogenous evolution of leverage and liquidity during low-volatility periods, traces the microscopic chain through which stability produces instability, and shows how the pattern appears in empirical data.

23.6.1 The endogenous accumulation of leverage

In the financial instability hypothesis, Minsky pointed out that the stable periods of a capitalist economy endogenously breed instability, because prolonged stability induces market participants to take on greater risk [33]. Minsky's theoretical framework divides the financial cycle into three stages: the hedge-financing stage (participants can repay debt through cash flow), the speculative-financing stage (participants can only repay interest but must borrow anew to repay principal), and the Ponzi-financing stage (participants cannot even repay interest and rely entirely on rising asset prices). Perpetual futures trading in cryptocurrency markets is an accelerated version of this evolution, and the low-volatility period is precisely the critical transition from hedge financing to speculative financing. However, the gradual deterioration from hedge financing to Ponzi financing in Minsky's framework depends on the relationship between debt-servicing cash flow and income, whereas perpetual futures trading does not generate "income" in the traditional sense, and participants' solvency depends entirely on the direction of asset-price movement. In this sense, highly leveraged directional perpetual futures positions have the characteristics of speculative or Ponzi financing from the moment they are opened, and Minsky's gradual-deterioration narrative is better understood in the crypto context as the accumulation process of total systemic leverage rather than a stage-by-stage evolution in the nature of financing. Geanakoplos's (2010) [34] leverage-cycle theory—whose core thesis is that changes in collateral value endogenously alter credit availability and produce procyclical fluctuations in leverage—may provide a theoretical anchor more directly applicable to crypto derivatives markets. This theory is especially pronounced in cryptocurrency perpetual futures markets, and its evolution is far faster than in traditional macroeconomic cycles. In the participant structure of cryptocurrency markets, the share of retail investors is far higher than in traditional financial markets, and retail investors' risk perception is often based on recent market performance rather than on long-run statistical regularities. When asset-price volatility remains low for a prolonged period, the behavioral patterns of market participants undergo a systematic change, and this change directly leads to a covert climb in the system's overall leverage level.

The starting point of the behavioral response is the interaction between risk perception and risk-management models. In a low-volatility environment, a VaR model estimated from historical data outputs extremely low risk assessments [35]. For institutional investors, this means that under the same risk budget, they can and must allocate a larger notional position to maintain the target rate of return. For individual traders, prolonged sideways movement or slow appreciation significantly lowers their vigilance toward tail risk, prompting them to adopt higher leverage multiples in perpetual futures. Over time, the ratio of the entire market's open-interest value to spot market capitalization continuously climbs. This phenomenon is especially pronounced in cryptocurrency markets, because the leverage caps of perpetual futures are often far higher than in traditional futures markets (BTC perpetual futures on some exchanges can

reach 100× leverage or higher), and retail investors, with relatively weak risk awareness, are more likely to adopt extreme leverage driven by the false sense of safety of low volatility.

This leverage-accumulation process is further amplified under the funding-rate mechanism unique to perpetual futures. When the market is in a calm period with a slight long bias, the funding rate tends to remain stable and positive. This stable income stream attracts a large number of arbitrageurs into the market to establish basis-arbitrage or funding-rate-arbitrage positions. Although these arbitrage positions are market-neutral by design, they greatly increase the total open interest of the perpetual futures market. At the same time, these seemingly low-risk positions are equally subject to maintenance-margin requirements when the market state changes abruptly. The distinction matters. Funding-rate arbitrage positions are delta-neutral by design—a spot long against a perpetual short—and their risk comes not from directional losses but from the margin calls caused by basis jumps and from the liquidity risk that the spot leg cannot be closed in time in extreme markets. Therefore, arbitrage open interest (OI) and speculative OI contribute to systemic fragility in different ways. In assessing systemic risk, the share of speculative OI is a more informative indicator than total OI. Nevertheless, when the calm period is broken and volatility rises sharply, both speculative and arbitrage positions of enormous size face margin pressure, instantly converting into potential liquidation momentum and providing ample fuel for subsequent volatility clustering.

At the microscopic level, leverage accumulation during low-volatility periods is also reflected in the dynamic adjustment of exchanges' margin systems. Exchange margin adjustment is in fact the superposition of multiple layers of mechanism: volatility-linked dynamic margin (procyclical—that is, lowering requirements when volatility is low, encouraging leverage), position-size tiered margin (countercyclical—that is, large accounts face higher margin rates, suppressing concentration), and, on some exchanges, pre-announced adjustment (buffering—that is, granting a 24-to-48-hour preparation period). On regulated exchanges (such as CME), margin adjustment is constrained by notice-period requirements, and its procyclical effect is buffered; on offshore exchanges, zero-notice margin adjustment can itself become a trigger of volatility. The net effect of the three remains procyclically dominated, because volatility-linked dynamic margin has the widest coverage and the largest adjustment magnitude. During low-volatility periods, dynamic margin requirements are gradually lowered, which further incentivizes traders to adopt higher leverage. Overall, this procyclically dominated margin-adjustment system is itself a fragility-amplification mechanism: when volatility suddenly rises, exchanges are often forced to rapidly raise margin requirements, and this sudden change in requirements triggers the forced liquidation of large numbers of positions, thereby intensifying the rise in volatility.

In extreme cases, exchanges may even suspend the liquidation function when market volatility rises to an extreme, in order to prevent a system collapse. Although this "circuit breaker" mechanism protects the exchange's system stability in the short term, it simultaneously freezes the market's price-discovery mechanism, so that trapped positions cannot be closed at any price, further intensifying the panic of market participants. This phenomenon was fully demonstrated in the "Black Thursday" event of March 2020, when the liquidation systems of several exchanges were temporarily

paralyzed by excessive traffic, leaving large numbers of positions unable to be closed in time and ultimately causing greater losses. On-chain liquidation systems exposed a different but equally serious fragility in the same event: MakerDAO's on-chain liquidation auctions were nearly paralyzed by severe congestion on the Ethereum network, as liquidators could not submit bid transactions because of the surge in gas fees, causing large amounts of collateral to be filled at zero bids. The fragility of on-chain liquidation stems from the throughput bottleneck of the underlying network rather than a design flaw in the liquidation engine itself, and it constitutes an amplification mechanism physically entirely different from the overload of a centralized exchange (CEX) liquidation engine.

23.6.2 Liquidity concentration

Accompanying the covert rise in leverage is a profound distortion of the liquidity structure. During low-volatility periods, market makers face a markedly different competitive environment and risk-return trade-off. Because prices fluctuate within a narrow range, orders far from the current market price have an extremely low probability of being executed, which greatly reduces the capital efficiency of providing liquidity across a wide price range. To capture limited trading volume and maximize capital utilization amid fierce competition, market makers rationally adjust their quoting strategy.

The core manifestation of this adjustment is the extreme concentration of liquidity. Market makers deploy the vast majority of their capital in an extremely narrow range hugging the current best bid and ask, compressing the bid-ask spread to the extreme and making the order book's apparent depth near the price appear exceptionally ample [36]. This surface abundance, however, masks the structural emptiness deeper in the order book. Because capital is drawn tightly toward the touchline, depth far from the current price declines sharply. In automated market maker (AMM) mechanisms, the widespread adoption of concentrated-liquidity design likewise intensifies this phenomenon, as liquidity providers tend to concentrate their funds near the current price to earn the highest fee income.

This distortion of the liquidity distribution greatly increases the fragility of the system. During calm periods, the extremely narrow spread and ample touchline depth create for traders an illusion of excellent market liquidity. However, once an information shock beyond ordinary expectations or a large sell order appears, the liquidity concentrated near the touchline is instantly broken through. Because far-end depth is empty, price jumps sharply as it searches for the next level of liquidity support. This discontinuous jump, triggered by a liquidity fault, produces a price impact far greater than under a normal liquidity distribution, pushing the system directly from a low-volatility state into an extreme high-volatility state within a very short time.

From the perspective of data microstructure, this phenomenon can be quantified through the shape parameters of the order book. During low-volatility periods, the order book typically exhibits a "spiked" shape—that is, an extremely high depth density near the current price, but a sharp decline in depth as the price distance increases. By comparison, the order-book shape under normal market conditions is more uniform and gradual. This spiking not only reduces static shock-absorption

capacity but also weakens the order book's recovery resilience—that is, the speed at which the order book returns to its pre-shock state after a price impact. Bouchaud et al. (2009) [37] showed that liquidity gaps in the order book amplify the price impact of a single order, and that the spread and depth recovery after a shock follow a slow power-law decay; this "slow recovery" feature corresponds directly to the "slow termination" feature of the liquidity loop described in Section 23.3. When the market shifts from calm to volatile, this spiked-shape order book cannot effectively absorb large order flow, forcing prices to jump to more distant price levels to find sufficient liquidity. In this process, each price jump triggers more liquidations, which in turn causes more sell orders to flood in, forming a self-reinforcing vicious cycle.

23.6.3 Empirical relationships

The "calm trap" is not merely a theoretical deduction but an empirical regularity deeply imprinted in the historical data of cryptocurrency markets. The evolution of slow variables (such as the leverage level and the liquidity structure) takes time; the longer a calm period lasts, the deeper these slow variables slide toward the danger range and the greater the fragility the system accumulates. A significant positive correlation therefore exists between the length of a calm period and the intensity of the volatility cluster that subsequently erupts.

Observation of Bitcoin's historical market movements clearly reveals this pattern. For example, before the violent crash of the Bitcoin price in mid-November 2018, the market experienced months of extremely low volatility, during which the Bollinger Bands narrowed to a historical extreme and open interest continuously climbed. Likewise, before the extreme "Black Thursday" market of March 2020, the market had also gone through a relatively calm, optimistic period. The May 2021 crash followed the same pattern: in the preceding weeks, Bitcoin oscillated repeatedly within the $45,000 to $65,000 range, volatility remained at a low level, yet open interest kept increasing. These historical events show that the eruption of extreme volatility clustering is often not an isolated random event but the inevitable result of continuously accumulating fragility during a long calm period.

Figure 23-11 shows the scatter relationship, based on Bitcoin historical data, between the length of calm periods and the intensity of subsequent volatility clustering: low-volatility calm periods lasting more than 50 days are often accompanied by subsequent extreme volatility clusters whose annualized volatility peaks exceed 100%, confirming that time is a key dimension of fragility accumulation. To quantify this relationship, defining a calm period as a continuous period in which the 30-day realized volatility is below an annualized 35%, and defining clustering intensity as the peak of the 30-day realized volatility that follows, the author identified about 15 independent calm-period–eruption-period pairs in BTC historical data from 2018 to 2024. The Spearman rank correlation coefficient is approximately 0.72 ($N=15$, $p<0.01$), though under such a small sample its confidence interval is wide (the 95% confidence interval is approximately 0.35 to 0.90), and removing the single most extreme event (March 2020) reduces the rank correlation to about 0.60. Although the precise value of the point estimate is sensitive to sample composition, the direction of the positive correlation remains robust

across various calm-period definition thresholds (25% to 45%) and volatility measures, indicating statistically meaningful support for the explanatory power of calm-period length over subsequent clustering intensity.

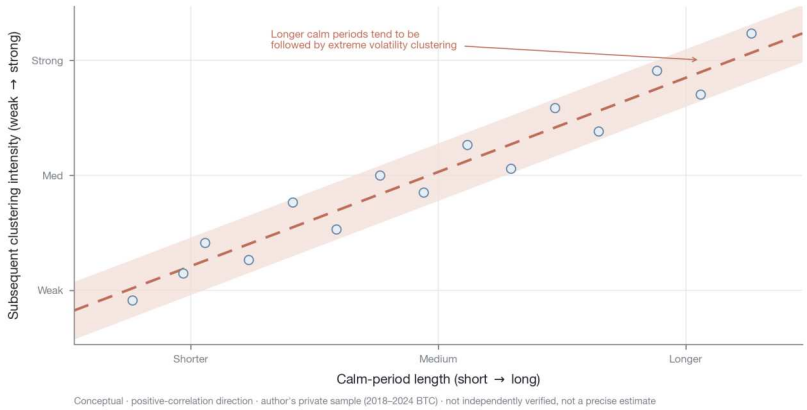


Figure 23-11. The positive correlation between calm-period length and subsequent volatility clustering intensity in Bitcoin historical data (Illustrative: direction of positive correlation; the author's proprietary sample of BTC from 2018 to 2024, not independently verified, not a precise estimate)

The mechanism behind this empirical relationship lies in the time-accumulation property of slow variables. The accumulation of leverage, the concentration of liquidity, and the parameter drift of risk models do not happen instantaneously but evolve slowly, day by day and week by week, during the calm period. A 30-day calm period may accumulate only enough fragility to support 50% volatility clustering, whereas a 70-day calm period may accumulate enough to support 150% volatility clustering. This nonlinear relationship explains why certain relatively small triggering events can provoke extreme market reactions after a long calm period.

23.6.4 Early-warning signals

Once the mechanism of fragility accumulation during low-volatility periods is understood, the core of risk management shifts from "predicting when a shock will occur" to "assessing how fragile the current system is." The VIX paradox in traditional financial markets points out that when the volatility index is at its lowest, it is often the moment when hedging demand is weakest and the system is least protected, and at this point the market's endogenous fragility is in fact highest [38]. In cryptocurrency perpetual futures markets, although a volatility-derivatives market of equal depth is lacking, we can identify the "calm trap" through a set of observable microstructural indicators.

In fragility assessment, leverage indicators are at the most upstream point of the causal chain. The ratio of the notional value of open interest to the market capitalization of the underlying asset is the most intuitive indicator of systemic leverage. In a low-volatility environment, if this ratio shows a persistent upward trend accompanied by the funding rate long deviating from a neutral level, it strongly suggests that a large number of highly leveraged positions extremely sensitive to one-directional price movement are

accumulating in the market. Figure 23-12 simulates the entire process in which open interest gradually climbs during the calm period, converts into liquidation momentum that pushes up volatility after the turning point, and then falls rapidly during the liquidation cascade—a process typically completed within hours to days.

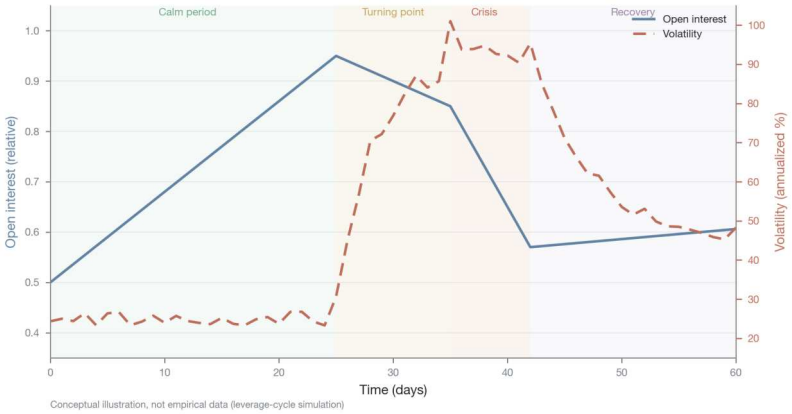


Figure 23-12. The temporal evolution dynamics of open interest and volatility in perpetual futures (Simulated by the author based on leverage-cycle theory; illustrative simulation, not empirical data)

Leverage accumulation determines the potential liquidation momentum, but whether this momentum can be absorbed by the market depends on the distribution structure of liquidity. Simply observing the touchline spread is often deceptive, because the extremely narrow spread of a low-volatility period is precisely the result of excessive liquidity concentration. A more reliable early-warning signal is the concentration indicator of order-book depth—that is, the ratio of quoted volume near the touchline (for example, within a 0.5% range) to quoted volume in a wider range (for example, within a 2% or 5% range). If this ratio rises significantly during a low-volatility period, it indicates that liquidity is becoming extremely concentrated, far-end depth is being hollowed out, and the market's capacity to absorb unexpected shocks is declining sharply. When this ratio exceeds 150% of the historical median, the system has already entered a high-fragility state. However, the depth-concentration ratio based on order-book snapshots has inherent limitations: visible quotes contain a large number of non-committed quotes (such as the ghost liquidity described in Chapter 20) that are rapidly withdrawn when the market moves, so the indicator may overestimate true depth. Alternative indicators based on actual transaction data—such as the time-weighted average price impact or the VPIN (volume-synchronized probability of informed trading) indicator—are less susceptible to illusory liquidity and can serve as cross-validation tools. More fundamentally, the early-warning indicators above have so far only been validated in-sample, and their out-of-sample predictive power (particularly sensitivity and specificity) still needs to be established through testing over a longer time series before operable early-warning thresholds can be set.

Above the objective fragility of the system constituted by leverage and liquidity, behavioral and sentiment indicators reflect market participants' subjective cognitive bias

toward this fragility. The extremization of the retail long-short ratio, the persistent exuberance of social-media sentiment indicators, and the narrowing of the spread between market makers' quoted implied volatility and historical realized volatility all indicate that market participants severely underprice tail risk. During low-volatility periods, if the spread between implied and realized volatility (commonly called the "volatility premium") remains persistently negative with a continuously expanding absolute value, this indicates that the market is collectively underestimating risk. This cognitive bias in turn reinforces leverage accumulation, because participants who underestimate risk are more inclined to raise their leverage multiples. Microstructural indicators such as the directionality of order flow, the arrival frequency of large orders, and market makers' inventory levels can further verify the above trends: when the arrival frequency of large buy orders persistently exceeds that of large sell orders during a low-volatility period, this suggests that market participants are actively building leveraged long positions.

Figure 23-13 summarizes the complete transmission mechanism of fragility accumulation during low-volatility periods. These early-warning signals cannot predict the timing and nature of the triggering event, but they can assess the system's current level of fragility, thereby shifting the focus of risk management from "when it will happen" to "how severe it will be if it happens."

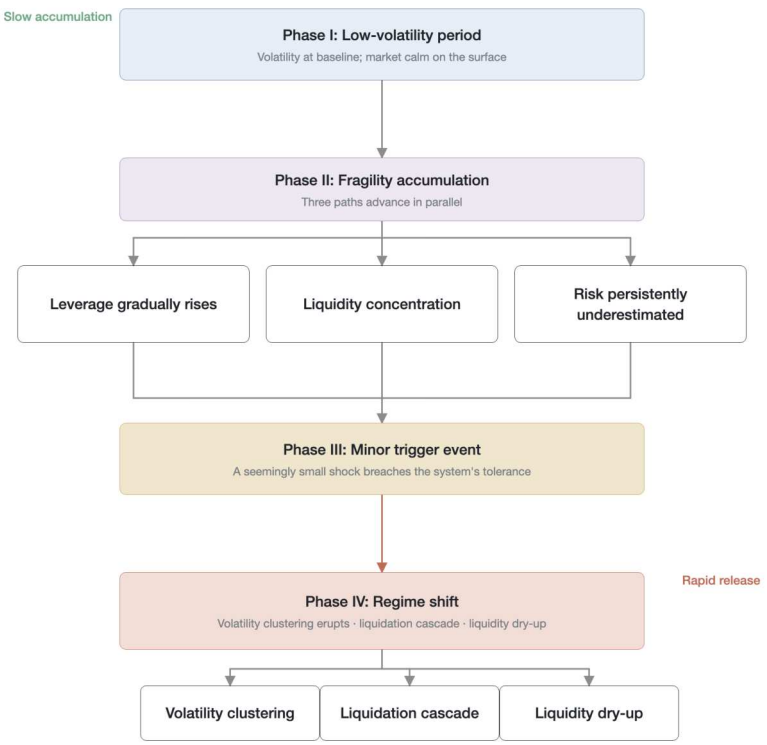


Figure 23-13. The mechanism of fragility accumulation and abrupt state change in a low-volatility environment (Drawn by the author based on this chapter's analytical framework; conceptual mechanism illustration, not empirical data)

The transmission path in Figure 23-13 clearly displays the complete logical closed loop of the "calm trap." Starting from the low-volatility environment at the top, three paths—leverage accumulation, liquidity concentration, and risk underestimation—advance in parallel, jointly pushing the system toward a high-fragility critical state. When an external triggering event breaks through the activation threshold set by the slow variables, the system jumps sharply from the upper part of the figure (the slow-accumulation stage) to the lower part (the rapid-release stage), and the positive feedback loop of liquidation cascade and liquidity exhaustion causes volatility to surge to extreme levels within a short time. The practical significance of this path diagram is that risk managers should focus their attention on the three accumulation paths in the middle of the figure, because they are the observable, measurable early-warning dimensions.

23.7 Case study: the life cycle of the May 2021 crash

To validate the three-layer explanatory framework of "slow variables → feedback loops → state machine" proposed above, we need a real market case with a complete life cycle. The Bitcoin market crash of May 2021 provides a classic natural experiment. This event was not a single random shock but a demonstration of the complete dynamic process of how volatility gradually accumulates fragility from a low level, how it is sharply amplified by positive feedback loops after a triggering event, and how the market state rapidly jumps from calm to crisis. By dissecting this historical case, we can clearly see how microstructural mechanisms shape the macro phenomenon of volatility clustering.

23.7.1 The state of the slow variables

The cryptocurrency market in the first quarter of 2021 was in a typical "calm" state, but this surface stability masked a sharp rise in underlying systemic fragility. The evolution of the slow variables set the tone for the subsequent violent volatility.

The level of systemic leverage, as the primary slow variable, underwent unprecedented expansion during this period. The Bitcoin price climbed from $29,000 at the start of 2021 to $64,000 by mid-April, a gain of more than 120%. In this strong uptrend, the endogenous character of leverage was fully displayed: as prices rose, the effective leverage ratio of long positions passively declined, prompting traders to continually add new leverage to maintain their target risk exposure. According to whole-market BTC perpetual futures data aggregated by Coinglass (including coin-margined and USDT-margined contracts), the notional value of open interest surged from about $10 billion at the start of 2021 to about $27 billion by mid-April, and climbed further to a peak of about $28 billion before the May crash [39]. At the same time, this leverage expansion was accompanied by extremely skewed market sentiment, with the funding rate long remaining at a level of 0.1% to 0.3% per 8-hour settlement period (a simple annualization of about 110% to 330%, though because arbitrageurs typically build positions when rates are low, the actual cost borne during the holding period is lower than the simple annualized value), indicating that longs not only held absolute dominance but were also willing to pay a high cost to maintain leverage.

Liquidity depth, another slow variable, was likewise gradually changing for the worse. Although the number of surface quotes on the order book appeared ample when prices were high, market makers' capital deployment exhibited a highly concentrated character. A large amount of liquidity clustered within an extremely narrow range near the current price, while the deep order book far from the current price was abnormally empty. This liquidity structure meant that once the price broke through the narrow high-liquidity range, the market would face a liquidity-fault zone with almost no buffer [20].

At the same time, changes in the participant structure further intensified the system's fragility. Blockchain data analysis shows that during this period the vast majority of Bitcoin addresses were in an unrealized-profit state, and large numbers of newly entered retail investors had established highly leveraged long positions at high prices [40]. These inexperienced investors had extremely low tolerance for price fluctuations, and their stop-loss orders were densely distributed not far below the current price. From the perspective of behavioral finance, the disposition effect (Shefrin & Statman 1985 [41]) meant that the large number of unrealized-profit holders constituted an enormous pool of potential sellers—profitable holders tend to close positions within the profit range; once a negative catalyst appeared, this seller pool could be rapidly activated, triggering large-scale profit-taking selling.

By this point, the slow variables had completed a dangerous combined configuration: elevated systemic leverage, weak deep liquidity, and highly concentrated stop-loss orders. This state of "high leverage + thin depth + high sentiment" constituted an extremely fragile market environment. In such an environment, the system did not need to withstand a large-scale external shock; a routine-scale piece of negative news was enough to break through the feedback-activation threshold that the slow variables had lowered to an extremely low level.

23.7.2 The trigger and initial shock

In mid-May 2021, the long-accumulated fragility finally met its triggering event. This trigger was not a single catastrophic piece of news but the superposition of a series of negative events that together broke through the critical point at which the system maintained equilibrium.

On May 12, Tesla CEO Elon Musk announced on social media that, out of concern about the environmental impact of Bitcoin mining, Tesla would stop accepting Bitcoin as a means of payment [42]. This statement directly struck one of the core narratives underpinning the first-quarter bull market—institutional adoption. Immediately afterward, on May 13, Chinese regulators reiterated their ban on financial institutions and payment companies providing cryptocurrency-related services, further intensifying market panic. Such regulatory shocks amplify and spread risk through the contagion channel between crypto assets and traditional securities markets [43]. Unlike a one-time information shock such as the Musk tweet, the Chinese regulatory tightening was essentially a structural institutional change; it not only triggered panic in the short term but also, over the medium to long term, set off a series of structural adjustments—including mining migration, the shift of exchange users to offshore platforms, and the closure of fiat on-ramps—and these supply-side changes lasting for months prolonged the tail of

the volatility clustering.

These pieces of news constituted the initial information shock. The Bitcoin price fell at one point in mid-May from about $57,000 to near $52,000, a single-day decline of about 9%. In a traditional spot market, a 9% decline, though significant, would typically be regarded as a normal correction. In the highly leveraged environment of perpetual futures, however, this initial shock underwent a qualitative change.

The initial price decline first caused a sharp reversal in the funding rate. The positive funding rate that had long remained high rapidly turned negative, marking a sudden reversal of market sentiment, and some highly leveraged longs began to be forced to close positions. At the same time, the 9% decline was enough to reach the maintenance-margin lines of a large number of highly leveraged long positions established above $55,000. Initial liquidations began to occur, and the market sell orders automatically generated by the liquidation engine began to hit the order book.

At this stage, the market state machine underwent its first discrete jump: from the "calm" state to the "stressed" state. Volatility indicators responded immediately to this jump, with Bitcoin's realized volatility jumping rapidly from the calm-period range of an annualized 30% to 45% to a level of an annualized 75% to 85%. Market makers began to sense the risk, spreads widened from a normal 3 basis points, and some ghost liquidity (as described in Chapter 20, referring to non-committed market-making quotes that are displayed under normal market conditions but rapidly withdrawn under stress) began to retreat. At this point, however, the system was still in a state of strained support, and a full-blown cascade had not yet erupted.

23.7.3 The cascade and crisis

After the initial shock pushed prices down and triggered the first wave of liquidations, the second-layer mechanism—the feedback loops—was fully activated. From May 13 to 19, the market witnessed one of the most violent liquidation cascades in cryptocurrency history, precisely the result of the mutual reinforcement of the volatility-leverage feedback loop and the volatility-liquidity feedback loop.

As prices fell further, the volatility-leverage feedback loop began to operate continuously. Falling prices eroded the margin of more long positions, the effective leverage ratio rose passively, and the liquidation trigger distance shortened. When the price broke below the key psychological threshold of $50,000, it triggered densely distributed stop-loss orders and liquidation orders. These forced sell orders hit the market indiscriminately, driving prices down further and, in turn, triggering deeper liquidations [12].

At the same time, the volatility-liquidity feedback loop intensified this process. Facing sharply rising volatility and one-directional liquidation selling pressure, market makers faced extremely high adverse-selection risk and inventory risk. To protect their own capital, market makers sharply widened the bid-ask spread and even withdrew buy orders entirely. At the most extreme moment, market depth evaporated and the order book became extremely thin [20]. The exhaustion of liquidity meant that the same quantity of liquidation sell orders caused a far larger price impact than usual.

These two feedback loops formed a self-reinforcing downward spiral: liquidation trig-

gered → price decline → liquidity withdrawal → price impact amplified → more positions liquidated. This process peaked on May 19. That day, the Bitcoin price plunged by about 32% within less than 12 hours (an intraday-peak measure; about −30% on a daily-close basis), touching a low below $30,000; Ethereum's decline was deeper, with an intraday maximum drop of about 46% [44].

The data clearly record the severity of this crisis. According to statistics from several data-aggregation platforms (Coinglass, The Block), the notional value of positions forcibly liquidated across the entire crypto market on May 19 exceeded $8 billion (about $8.6 billion by multi-source aggregation; this figure is the notional size of liquidated positions rather than the actual loss amount, the latter being difficult to estimate precisely because of slippage and bankruptcy); among them, Deribit alone saw 296 Ethereum positions and 609 Bitcoin positions undergo bankruptcy liquidation [44]. Whole-market BTC perpetual open interest plunged from about $28 billion to near $10 billion within just a few days, indicating that leverage within the system was violently cleared out [39].

At this stage, the market state machine completed its second jump: from the "stressed" state fully into the "crisis" state. Volatility was no longer continuous but exhibited jump-like behavior, soaring directly from around an annualized 80% to an extreme level of an annualized 150% to 200%. Cross-exchange spreads widened sharply, and the arbitrage mechanism temporarily failed in the face of an extremely congested network and exhausted capital. The cross-instrument contagion effect of this crisis was equally significant: Ethereum perpetual futures plunged 46% on the same day, a decline far exceeding BTC's 32%. This difference may stem from the ETH market's higher starting leverage ratio and thinner deep liquidity, as well as traders under a unified margin account preferentially liquidating the more liquid ETH positions to free up BTC margin. At the peak of liquidation, the insurance funds of several exchanges suffered large drawdowns, and some platforms triggered the auto-deleveraging mechanism (the reverse cascade described in Section 23.5.2, which forcibly closes profitable counterparties' positions), further compressing the supply of liquidity.

Figure 23-14 superimposes four time series from this life cycle—the price trajectory, 24-hour liquidation volume, open-interest size, and realized volatility—so that the reader can trace, on a single timeline, the successive activation and transmission of each layer of mechanism.

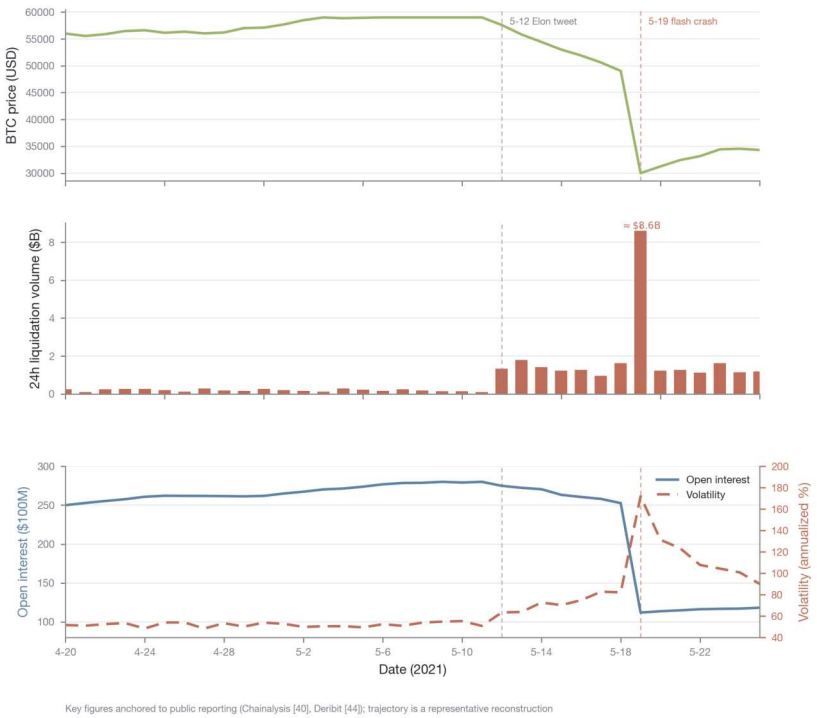


Figure 23-14. The volatility clustering life cycle of the May 2021 BTC crash (Empirical figure; price and liquidation-volume data sources: Chainalysis [40] and Deribit Insights [44]; open-interest series source: Coinglass [39]; realized volatility is the author's calculation based on daily returns; ETH's intraday maximum drop of about 46% is an intraday-peak measure, about −40% on a daily-close basis)

In Figure 23-14, the three stages corresponding to Sections 23.7.1–23.7.3—incubation, stress, and crisis—can be clearly identified: open interest first climbs to a peak during the calm period, then declines off a cliff during the crisis period, while liquidation volume and realized volatility surge synchronously to extreme values on May 19. The violent movements of the four curves are highly synchronized on the timeline, precisely mapping the positive-feedback transmission path of liquidation cascade → liquidity exhaustion → volatility surge.

23.7.4 Theoretical validation

The May 2021 crash case provides a consistent illustration of the three-layer explanatory framework proposed in this chapter, demonstrating the framework's structural advantages in explaining extreme market behavior. The ex-post analysis of a single case carries inherent methodological limitations, however: selective interpretation after the fact may exaggerate the framework's explanatory power, and the framework's applicability to other historical events, such as the 2022 LUNA/UST collapse, still requires independent validation. The discussion that follows should be read under this constraint.

First, the analysis at the slow-variable level successfully explains why the system was

in an extremely fragile state before the shock. Traditional statistical models (such as GARCH) can only make lagged adjustments after volatility has already risen and cannot warn of the accumulation of fragility during the calm period. By observing the expansion of open interest and the persistently elevated funding rate, the slow-variable framework accurately captured the potential of "how severe the reaction would be if a shock occurred." This proves that the crisis was not caused unilaterally by Musk's tweet or the regulatory news alone but was the joint result of an extremely fragile environment and an ordinary information shock.

Second, the analysis at the feedback-loop level explains why volatility exhibits clustering and persistence. In the flash crash of May 19, the plunge in price far exceeded what any fundamental information could explain. The closed loop formed by the liquidation engine's mechanical selling and market makers' liquidity withdrawal clearly demonstrates how the shock was self-amplified by the system's internal mechanisms. It is precisely these microstructural feedback loops that constitute the physical basis of the "long memory" in the volatility time series, allowing the high-volatility state to sustain itself for several days after being triggered.

Finally, the state-machine model accurately describes the nonlinear character of market dynamics. The case shows that the transition from "calm" to "stressed" was relatively gradual, but the jump from "stressed" to "crisis" erupted within a few hours. In the crisis state, the dominant source of volatility switched completely from fundamental information to mechanistic liquidation and the absence of liquidity. This asymmetry of state transitions—very fast entry into crisis, slow exit from it—is the fundamental reason volatility clusters in perpetual futures markets display a sawtooth distribution [45].

However, this case also reveals the inherent limitations of the theoretical framework. Although the three-layer framework can accurately assess the system's fragility and explain the amplification mechanism of the feedback loops, it cannot predict the specific timing, nature, and exact magnitude of the triggering event. This limitation does not diminish the framework's value; rather, it points to the correct path for risk management: rather than trying to predict unpredictable triggering events, it is better to focus on monitoring and managing the observable slow variables and feedback states.

Abstracting the causal chain distilled from this case one step further, one can construct a systemic-crisis formation-path model applicable to more general situations. Figure 23-15 decomposes this path into four progressive stages: preconditions, the trigger layer, the transmission layer, and the amplification layer.

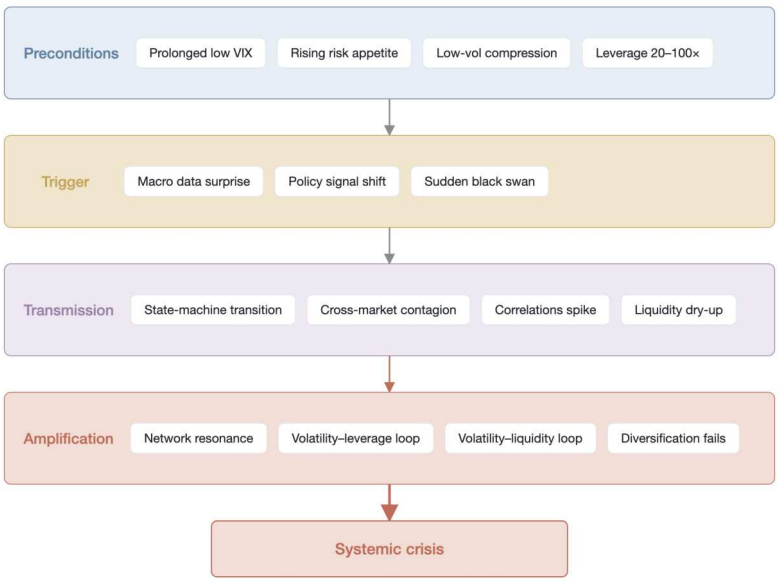


Figure 23-15. The formation path of a systemic financial crisis (Drawn by the author based on this chapter's analytical framework; conceptual mechanism illustration, not empirical data)

Figure 23-15 expresses this chapter's "slow variables → feedback loops → state machine" framework as a structured path and precisely maps the nonlinear dynamics observed in the May 2021 cryptocurrency market crash: a crisis is not triggered by a single external shock but is the inevitable result of the interaction between the long-term accumulation of internal systemic fragility and specific triggering conditions.

Along this path: in the precondition stage, the evolution of slow variables raises the system's fragility baseline, with the 20× to 100× leverage of perpetual futures acting here as a significant amplifier that raises the system's sensitivity to minor price changes exponentially; in the trigger layer, an exogenous shock shatters the illusion of low volatility and reaches the maintenance-margin lines of a large number of long positions, triggering initial liquidations; the transmission layer corresponds to the state machine's state switching, as the market jumps rapidly from calm to crisis and cross-market contagion and liquidity exhaustion appear in succession; and the amplification layer is the macro-network resonance of the feedback loops, in which the positive feedback of volatility-leverage and volatility-liquidity intertwines and self-reinforces, ultimately evolving local volatility into a full-blown systemic crisis. The policy implication of this formation path is that although a systemic crisis is abrupt, the accumulation of its underlying fragility and the evolution of its transmission mechanisms are traceable, structured processes: through continuous monitoring of the slow variables and feedback states, regulators and market participants may be able to identify the system's critical state before a crisis fully erupts, and thereby adopt more forward-looking risk-management measures.

23.8 Chapter summary

In empirical data, volatility clustering manifests as the dependence of the conditional variance on past shocks, a phenomenon that has been widely documented in time-series analysis. However, purely statistical models cannot explain the microscopic source of this memory effect. The three-layer explanatory framework proposed in this chapter fills this theoretical gap, reducing the phenomenon of volatility clustering to an identifiable process of market microstructure evolution. The first layer is the slow variables, including the macroeconomic environment, the level of systemic leverage, and liquidity depth; these variables evolve over longer time scales, and their core role is not to directly generate volatility but to set the baseline fragility for the entire system, determining how high or low the threshold for activating the feedback loops is. The second layer is the feedback loops, which constitute the microscopic dynamic basis for the self-amplification of shocks: significant positive feedback mechanisms exist between volatility and liquidity, leverage, and market behavior, and when an initial shock crosses the activation threshold set by the slow variables, market makers' withdrawal thins liquidity while the erosion of margin triggers passive leverage increases and liquidations—the rational reactions of these micro-agents converge at the macro level into the self-perpetuation of volatility. The third layer is the state-machine model, which discretizes the market's evolution into three states—calm, stressed, and crisis—with transitions among the different states exhibiting strong asymmetry: the process of entering a crisis is often driven by positive feedback and completed within an extremely short time, whereas exiting a crisis requires waiting for the momentum of the feedback loops to be exhausted and for liquidity to recover slowly. Figure 23-16 shows the overall structure of this framework.

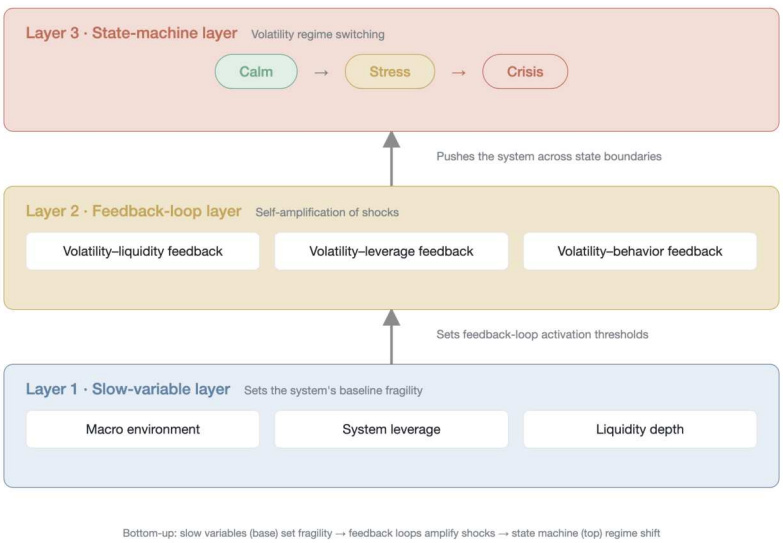


Figure 23-16. The mechanistic explanatory framework for volatility clustering (Drawn by the author based on this chapter's analytical framework; conceptual framework illustration, not empirical data)

Figure 23-16 condenses the core argumentative structure of this chapter into a bottom-

up three-layer architecture diagram. The practical value of this framework is that it provides a layered diagnostic tool for volatility forecasting and risk management: first assess the range in which the slow variables lie to judge systemic fragility; next monitor the activation signals of the feedback loops to identify the precursors of state transitions; and finally select the corresponding forecasting model and hedging strategy according to the current state.

Compared with traditional financial markets, crypto-asset perpetual futures markets exhibit a more violent volatility clustering effect, and this difference stems not from higher uncertainty in the intrinsic value of the underlying asset but from amplification by the market's distinctive institutional design. The liquidation cascade mechanism, under conditions of extremely high leverage and the absence of a margin-call buffer, instantly converts adverse price movements into market-price liquidation orders, producing an extremely high positive-feedback gain. The pulse-superposition effect of funding rates prolongs the duration of high volatility in the stressed or crisis state. The around-the-clock, no-close trading rule eliminates the window for market-maker capital resetting and emotional cooling that the market-close periods of traditional markets provide. To reduce the systemic risk of this market, policymaking and protocol design must intervene in these institutional features—for example, by introducing smoother liquidation-execution algorithms or setting dynamic caps on funding rates. These suggestions are better understood as directions for optimization at the protocol-design level than as top-down regulatory requirements, because trading volume in perpetual futures markets is concentrated mainly on offshore exchanges beyond the reach of most regulators, and globally unified mandatory rules are difficult to achieve in the short term. In addition, if a funding-rate cap is set improperly, it may introduce new distortions: persistent rate suppression would cause a systematic deviation between the contract price and the spot price, or drive trading activity to migrate to platforms without caps, producing regulatory arbitrage.

Low-volatility periods are usually intuitively regarded as safe periods for the market, but this chapter's analysis reveals that they are in essence the stage in which systemic fragility accumulates fastest. During a prolonged calm state, traders gradually raise leverage as confidence strengthens, market makers narrow spreads amid competition and cause liquidity concentration to rise, and the system's overall risk exposure continuously increases—even though all observable volatility indicators are at low levels. This mechanism poses a fundamental challenge to the traditional risk-management paradigm: a VaR model based on historical volatility systematically underestimates tail risk during calm periods. For example, in the calm period before the May 2021 event, the 99% VaR estimated from the past 30 days' realized volatility was about a single-day decline of 8% to 10%, whereas the actual single-day decline on May 19 reached 32%, more than three times the VaR prediction; even switching the risk measure to expected shortfall would struggle to close this gap, because the nonlinear amplification effect of the liquidation cascade exceeds the capture range of any static risk model based on the extrapolation of historical distributions. Therefore, the focus of risk management must shift from merely monitoring the current volatility level to continuously assessing the system's underlying fragility, integrating real-time monitoring of the leverage distribution, liquidity depth, and order-book concentration.

By deconstructing the microstructural mechanisms of volatility clustering, this chapter has answered the core question of why a high-volatility state self-perpetuates. Having clarified the driving mechanisms, the focus of the analysis will turn to how to exploit these regularities. The three-layer framework directly leads to a layered forecasting and decision-making system: first, identify the current market state through observable slow variables and microstructural indicators; next, select the corresponding forecasting model according to the dominant volatility dynamics of each state; and finally, formulate dynamic trading and hedging strategies based on the state-specific path of risk evolution. This logical progression—from mechanistic understanding to quantitative forecasting and then to strategy execution—constitutes the core thread of the next stage of analysis.

References

[1] Nikolova, V., Segovia, J. T., Fernández-Martínez, M., & Sánchez-Granero, M. A. (2020). A novel methodology to calculate the probability of volatility clusters in financial series: An application to cryptocurrency markets. *Mathematics*, *8*(8), Article 1216. https://doi.org/10.3390/math8081216

[2] Silva, L. G. D., & Maciel, L. (2025). Cryptocurrency price returns volatility modeling and forecasting with GARCH models. *RAUSP Management Journal*, *60*(1), 225–239. https://doi.org/10.1108/RAUSP-04-2023-0056

[3] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007. https://doi.org/10.2307/1912773

[4] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

[5] Kim, J. M., Jun, C., & Lee, J. (2021). Forecasting the volatility of the cryptocurrency market by GARCH and stochastic volatility. *Mathematics*, *9*(14), Article 1614. https://doi.org/10.3390/math9141614

[6] Brini, A., & Lenz, J. (2024). A comparison of cryptocurrency volatility-benchmarking new and mature asset classes. *Financial Innovation*, *10*(1), 1–38. https://doi.org/10.1186/s40854-024-00646-y

[7] Cont, R. (2007). Volatility clustering in financial markets: Empirical facts and agent-based models. In *Long memory in economics* (pp. 289–309). Springer. https://doi.org/10.1007/978-3-540-34625-8_10

[8] Scaillet, O., Treccani, A., & Trevisan, C. (2020). High-frequency jump analysis of the Bitcoin market. *Journal of Financial Econometrics*, *18*(2), 209–232. https://doi.org/10.1093/jjfinec/nby013

[9] Naimy, V., Haddad, O., Fernández-Avilés, G., & El Khoury, R. (2021). The predictive capacity of GARCH-type models in measuring the volatility of crypto and world currencies. *PLoS ONE*, *16*(1), Article e0245904. https://doi.org/10.1371/journal.pone.0245904

[10] Wang, J., Ma, F., Bouri, E., & Guo, Y. (2023). Which factors drive Bitcoin volatility: Macroeconomic, technical, or both? *Journal of Forecasting*, *42*(4), 970–988. https://doi.org/10.1002/for.2930

[11] BlackRock. (2025). *Four factors behind bitcoin's recent volatility.* BlackRock. Retrieved June 24, 2026, from https://www.blackrock.com/us/financial-professionals/insights/exploring-crypto-volatility

[12] Easley, D., O'Hara, M., Yang, S., & Zhang, Z. (2024). Microstructure and market dynamics in crypto markets (SSRN Working Paper No. 4814346). https://doi.org/10.2139/ssrn.4814346

[13] Locke, P. R., & Sarkar, A. (2001). Liquidity supply and volatility: Futures market evidence. *Journal of Futures Markets*, *21*(1), 1–17. https://doi.org/10.1002/1096-9934(200101)21:1<1::AID-FUT1>3.0.CO;2-Q

[14] Krishnamurthy, A. (2010). Amplification mechanisms in liquidity crises. *American Economic Journal: Macroeconomics*, *2*(3), 1–30. https://doi.org/10.1257/mac.2.3.1

[15] Dobrev, D., Liu, E., Kim, J., & Rodriguez, M. (2025). Order flow imbalances and amplification of price movements: Evidence from U.S. Treasury markets. *FEDS Notes.* Board of Governors of the Federal Reserve System. https://doi.org/10.17016/2380-7172.3954

[16] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

[17] Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56. https://doi.org/10.1016/S1386-4181(01)00024-6

[18] Jacobs, K., Mai, A. T., & Pederzoli, P. (2025). Market makers and the dynamics of volatility demand (SSRN Working Paper No. 5052438). https://doi.org/10.2139/ssrn.5052438

[19] Wyart, M., Bouchaud, J.-P., Kockelkoren, J., Potters, M., & Vettorazzo, M. (2008). Relation between bid–ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, *8*(1), 41–57. https://doi.org/10.1080/14697680701344515

[20] Baumgartner, T., & Guettler, A. (2022). *Bitcoin flash crash on May 19, 2021: What did really happen on Binance?* [Conference paper]. Paris December Finance Meeting. https://www.paris-december.eu/sites/default/files/papers/2022/Baumgartner_2022_2.pdf

[21] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[22] He, S., Manela, A., Ross, O., & von Wachter, V. (2022). Fundamentals of perpetual futures. *arXiv.* https://doi.org/10.48550/arXiv.2212.06888

[23] Mandaci, P. E., & Cagli, E. C. (2022). Herding intensity and volatility in cryptocurrency markets during the COVID-19. *Finance Research Letters*, *46*(PB), 102382. https://doi.org/10.1016/j.frl.2021.102382

[24] Alexander, C., Deng, J., & Zou, B. (2021). Hedging with bitcoin futures: The effect of liquidation loss aversion and aggressive trading. *arXiv*. https://doi.org/10.48550/arXiv.2101.01261

[25] Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779–1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

[26] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357–384. https://doi.org/10.2307/1912559

[27] Chourdakis, K. (2002). *Continuous time regime switching models and applications in estimating processes with stochastic volatility and jumps* (Working Paper No. 464). Queen Mary University of London, Department of Economics.

[28] Hamilton, J. D., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, *64*(1–2), 307–333. https://doi.org/10.1016/0304-4076(94)90067-1

[29] Trabelsi, N. (2018). Are there any volatility spill-over effects among cryptocurrencies and widely traded asset classes? *Journal of Risk and Financial Management*, *11*(4), Article 66. https://doi.org/10.3390/jrfm11040066

[30] Schmeling, M., Schrimpf, A., & Todorov, K. (2023). *Crypto carry* (BIS Working Paper No. 1087). Bank for International Settlements. https://doi.org/10.2139/ssrn.4268371

[31] Amberdata. (2026). *The volatility framework: How to read crypto's stress signals*. Amberdata. Retrieved June 24, 2026, from https://blog.amberdata.io/the-volatility-framework-how-to-read-cryptos-stress-signals

[32] Bloomberg. (2026, February 11). Bitcoin's 24/7 trading risk spikes while Wall Street sleeps. *Bloomberg*. Retrieved June 24, 2026, from https://www.bloomberg.com/news/articles/2026-02-11/bitcoin-s-24-7-trading-risk-spikes-while-wall-street-sleeps

[33] Minsky, H. P. (1992). *The financial instability hypothesis*. The Jerome Levy Economics Institute Working Paper No. 74. https://www.levyinstitute.org/pubs/wp74.pdf

[34] Geanakoplos, J. (2010). The leverage cycle. *NBER Macroeconomics Annual*, *24*(1), 1–66. https://doi.org/10.1086/648285

[35] Adrian, T., & Shin, H. S. (2014). Procyclical leverage and value-at-risk. *The Review of Financial Studies*, *27*(2), 373–403. https://doi.org/10.1093/rfs/hht068

[36] Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press. https://doi.org/10.1093/oso/9780195301649.001.0001

[37] Bouchaud, J.-P., Farmer, J. D., & Lillo, F. (2009). How markets slowly digest changes in supply and demand. In *Handbook of Financial Markets: Dynamics and Evolution* (pp. 57–160). Elsevier. https://doi.org/10.1016/B978-012374258-2.50006-3

[38] Whaley, R. E. (2009). Understanding the VIX. *The Journal of Portfolio Management*, *35*(3), 98–105. https://doi.org/10.3905/JPM.2009.35.3.098

[39] Coinglass. (2026). *Bitcoin futures open interest* [Data set]. Coinglass. Retrieved June 24, 2026, from https://www.coinglass.com/BitcoinOpenInterest

[40] Chainalysis. (2021, May 19). *The May 2021 cryptocurrency crash explained: Selling by retail and not institutions is driving crypto price plunge.* Chainalysis Blog. Retrieved June 24, 2026, from https://www.chainalysis.com/blog/cryptocurrency-price-crash-may-2021/

[41] Shefrin, H., & Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *The Journal of Finance*, *40*(3), 777–790. https://doi.org/10.1111/j.1540-6261.1985.tb05002.x

[42] Ante, L. (2023). How Elon Musk's Twitter activity moves cryptocurrency markets. *Technological Forecasting and Social Change*, *186*(Part A), 122112. https://doi.org/10.1016/j.techfore.2022.122112

[43] Sajeev, K. C., & Afjal, M. (2022). Contagion effect of cryptocurrency on the securities market. *SN Business & Economics*, *2*(7), Article 57. https://doi.org/10.1007/s43546-022-00219-0

[44] Deribit Insights. (2021, May 22). *The (flash) crash of May 19, 2021.* Deribit. Retrieved June 24, 2026, from https://insights.deribit.com/exchange-updates/the-flash-crash-of-may-19-2021/

[45] Fan, H., Xu, Y., Bouri, E., & Zeng, P. (2023). Clustering effect in higher-order moments across various timescales in the cryptocurrency market (SSRN Working Paper No. 4503527). https://doi.org/10.2139/ssrn.4503527

Volatility Forecasting and Decision-Making

In mid-March 2024, as the price of Bitcoin repeatedly set new record highs, open interest in the derivatives market briefly approached a peak of nearly $38 billion, after which a price pullback of just 7% triggered more than $800 million in cascading liquidations [1]. This episode highlights a core structural feature of the crypto-asset derivatives market: the eruption of systemic risk often stems not from a fundamental reversal in price direction but from a nonlinear expansion of volatility over an extremely short interval. For market participants, understanding the sources and clustering mechanisms of volatility is only the first step; the more challenging question is how to translate this theoretical understanding into executable forecasting models and risk-control decisions. That question is the central analytical objective of this chapter.

A market maker need not forecast the direction of Bitcoin's next-day price move, but must estimate its next-day volatility accurately. When expected volatility rises, the rational response is to widen the quoted spread, reduce order book depth, and tighten inventory limits; when volatility is expected to stay low, the market maker can quote narrower spreads and provide deeper liquidity [2]. What the market maker forecasts, in short, is not direction but the dynamics of volatility. The same holds for a portfolio manager, for whom the precise level of returns matters far less than the width of the return distribution, because that width drives the calibration of value at risk (VaR) and the sizing of positions [3]. And for exchanges and protocol designers, guarding against systemic risk turns on predicting which level of volatility will push liquidation volume past safety thresholds, so that margin requirements can be raised preemptively [4].

Volatility forecasting cannot reveal the specific events that will occur, but it can quantify the magnitude of the impact that potential events would have on market prices. Chapter 22 established a framework decomposing volatility into four sources, and Chapter 23 analyzed the clustering and regime-switching of volatility; the task of this chapter is to build, on that foundation, a set of operational forecasting models and a decision-making system. Classical volatility-clustering theory holds that large price changes tend to be followed by further large changes [5], an empirical regularity widely verified in traditional financial markets. The crypto-asset perpetual futures market, however, has distinctive properties—24/7 continuous trading, periodic funding settlement, and the absence of a deep options market—that call for targeted structural adjustments to the classical generalized autoregressive conditional heteroskedasticity (GARCH) model

and the heterogeneous autoregressive (HAR) model [6].

This chapter begins with the volatility information set specific to the perpetual futures market and examines the predictive power of the funding rate, liquidation heatmaps, and order book shape in the absence of implied volatility. On that basis, it proposes a source-decomposed volatility forecasting framework that models informational, liquidity, mechanical, and transmission volatility separately before aggregating them, precisely because these sources differ so fundamentally in predictability and time scale. Those forecasts are then embedded in a three-layer decision architecture spanning micro-level market-making quotes, meso-level position management, and macro-level systemic defenses. By the end of the chapter, the reader should be able to judge where different volatility-forecasting models apply to perpetual futures data, use alternative signals to read the market's risk state, and locate the boundary between the modelable and the intrinsically unpredictable components of volatility—the knowledge on which more resilient trading and risk-control systems are built.

24.1 The set of alternative volatility signals for perpetual futures

In traditional financial markets, option-implied volatility provides participants with a unified, forward-looking volatility benchmark. In the cryptocurrency perpetual futures market, however, the vast majority of assets lack a deep options market. This structural gap forces participants to construct an alternative system of volatility signals from dimensions such as the funding rate, liquidation heatmaps, order book microstructure, and on-chain data. This section analyzes the information content and predictive power of these signals in turn, assessing the conditions under which each class applies across different time scales.

24.1.1 The absence of implied volatility

In traditional financial markets, the core input for volatility forecasting is the implied volatility supplied by the options market. The Chicago Board Options Exchange's VIX, a measure of the implied volatility of S&P 500 index options, is widely used as a benchmark for gauging market risk expectations and the level of volatility over the next 30 days [7]. The distinctive value of implied volatility lies in its forward-looking nature: it is not a simple extrapolation of historical price movements but a consensus price reached by market participants through actual option-trading positions. When market makers and institutional investors expect volatility to rise, they demand higher option premiums, which directly pushes up the implied-volatility measure.

Although the cryptocurrency market has developed a derivatives ecosystem of some scale, its structure differs markedly from that of traditional finance. Platforms such as Deribit offer options trading on Bitcoin and Ethereum and have launched a VIX-like index, the Deribit volatility index (DVOL) [8]; nonetheless, the depth, liquidity, and breadth of participation in crypto options markets remain far below those of traditional markets. More important, although platforms such as Deribit have extended options coverage to leading altcoins such as SOL and XRP, and decentralized options platforms

such as Aevo offer a broader product line, most altcoins still lack options markets with sufficient depth and liquidity. As a result of this structural feature, perpetual futures participants cannot rely on a unified, deep implied-volatility indicator when making trading and risk-management decisions.

The absence of implied volatility has forced the perpetual futures ecosystem to develop a distinctive system of alternative signals. These signals are scattered across the funding rate, order book shape, liquidation data, and on-chain fund flows. Although no single signal can match the precision of the VIX, by cross-validating microstructure data across multiple dimensions, participants can still construct a highly informative set of volatility predictors. Figure 24-1 maps these six classes of volatility information source against the time scale over which each forecasts effectively and its relative predictive power, giving an overview of the volatility information available in the perpetual futures ecosystem.

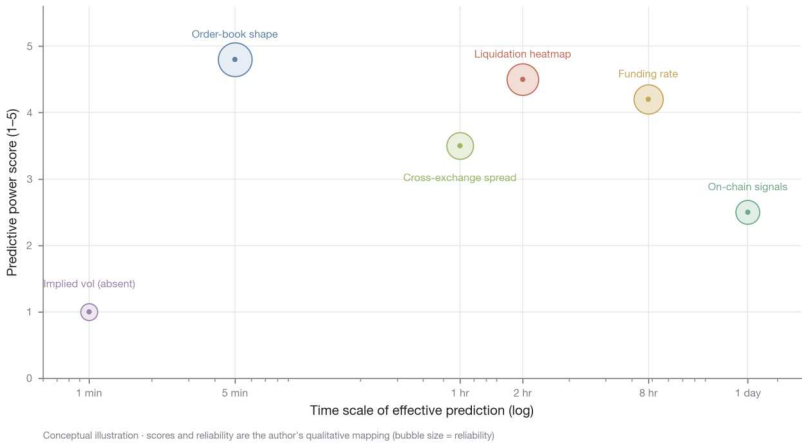


Figure 24-1. A panorama of the volatility information set for perpetual futures (conceptual mapping: the horizontal time-scale axis reflects true orders of magnitude, whereas the vertical predictive-power scores and the bubble reliability are the author's qualitative assignments rather than measured data; data source: compiled by the author)

The signals form a complementary gradient across time scales, running from minute-level order book shape to multi-day on-chain and cross-market signals. It is this layered, multi-scale structure that compensates for the missing implied-volatility benchmark: no single signal replaces the breadth of information in the VIX, but the six sources together yield useful forward-looking judgments within the windows where each is strongest. The subsections below take up the mechanism and predictive logic of each class in turn.

24.1.2 The funding rate as a volatility signal

The funding rate is the core mechanism in perpetual futures that regulates the balance of long and short positions and anchors the contract to the spot price (as discussed in Chapter 17) [9]. Beyond its own arbitrage value, its dynamics carry information for forecasting volatility. When the funding rate turns extreme—when its absolute value deviates sharply from baseline levels—it is often a reliable precursor to rising market

volatility, because an extreme reading directly reflects a severe imbalance between long and short forces in the market.

The intrinsic link between extreme funding rates and rising volatility stems from a severe imbalance in directional demand. A very high positive rate indicates that longs dominate, with large amounts of leveraged capital chasing the uptrend; a deeply negative rate reflects extreme crowding on the short side. Such an imbalance is inherently fragile. When the market price moves in the opposite direction, the side bearing the high funding cost faces enormous pressure to close positions. This closing activity consumes liquidity on one side, which in turn amplifies price volatility. Extreme funding rates therefore not only reflect an excessive concentration of the market's current directional preference but also foreshadow potential future liquidation risk and volatility jumps.

Beyond the absolute level of the rate, the speed at which the funding rate changes—the volatility of the rate—is likewise a highly informative predictor. When the funding rate fluctuates violently over a short interval, especially when it flips rapidly from strongly positive to strongly negative, it signals that the market structure is undergoing an abrupt reorganization. This reorganization is accompanied by the opening and closing of large volumes of positions, which is itself a direct manifestation of a liquidity shock. Empirical observation indicates that the speed of funding-rate changes tends to correlate positively with realized volatility over the subsequent 1 to 24 hours, consistent with the rapid transmission of rate information under the cross-platform tiered structure of the crypto funding rate market [10]. Figure 24-2 illustrates how extreme funding rates and their rate of change predict short-term volatility, plotting subsequent realized volatility against the absolute level of the rate in one panel and against the speed of rate changes in the other.

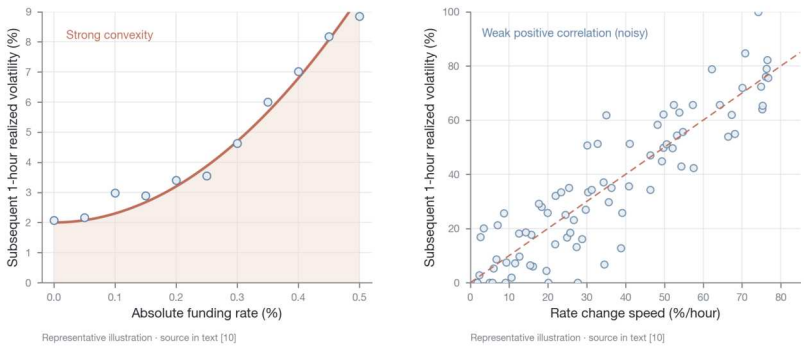


Figure 24-2. The predictive relationship between extreme funding rates and subsequent volatility (representative illustrative relationship: noisy scatter, not an empirical dataset; the tiered structure of the funding rate market follows Zhivkov 2026 [10])

The figure reveals two noteworthy patterns. When the absolute value of the funding rate exceeds two standard deviations above its historical mean, the probability of a marked jump in realized volatility over the subsequent 1 to 24 hours increases substantially; this threshold effect makes extreme funding rates an early-warning signal with a well-defined trigger condition. The speed of rate changes has independent predictive value as well:

even when the absolute level of the rate has not yet reached the extreme zone, rapid changes in it reveal that the market microstructure is undergoing an abrupt adjustment. The combination of these two dimensions forms a two-dimensional rate-warning matrix, in which the "high absolute level plus rapid change" quadrant corresponds to the highest short-term volatility risk.

24.1.3 Liquidation heatmaps

A distinctive source of risk in the perpetual futures market is the cascade effect triggered by the automatic liquidation mechanism (as discussed in Chapter 22). A liquidation heatmap maps the liquidation prices of all current open positions onto specific price ranges, giving a direct view of the scale of liquidations latent at different price levels. This spatial distribution is essentially a map of volatility risk, providing a key microstructure clue for forecasting mechanical volatility [11].

The predictive logic of a liquidation heatmap rests on the spatial distance between the price and dense clusters of liquidations. If the current market price is only a very short distance from a cluster containing a large volume of liquidatable positions, and the expected liquidation volume in that zone far exceeds the absorptive capacity of the current order book, then once the price reaches that zone it will inevitably trigger a liquidation cascade. In such a scenario, the liquidation engine dumps a large number of market orders into the market, instantly draining liquidity on one side and causing a nonlinear, extreme jump in price.

For market makers and risk-control systems, the liquidation heatmap provides a quantitative early warning of tail risk. When the price approaches a high-density liquidation cluster, the model should automatically raise its expectation of future mechanical volatility even if the current realized volatility is low. This conditional forecast, based on spatial distance, enables the risk-management system to take defensive measures—widening quoted spreads or raising margin requirements—before volatility actually erupts.

The data quality of liquidation heatmaps, however, has limitations that cannot be ignored. The liquidation heatmaps in widespread use today (such as the data provided by third-party platforms like Coinglass) do not come from complete liquidation-price distributions officially disclosed by exchanges; rather, they are estimates based on public open-interest data and assumed leverage distributions. Centralized exchanges' liquidation data depend on voluntary disclosure and suffer from coverage bias; decentralized protocols' liquidation data, though verifiable on-chain, currently account for only a small share of the perpetual futures market. More critically, large traders have every incentive to conceal their true liquidation-price levels—for example, by distributing positions across exchanges, dynamically adding margin, or using over-the-counter (OTC) hedges—so that their observable liquidation exposure is far smaller than its actual scale. Observable liquidation heatmaps should therefore be treated as a lower-bound estimate of tail liquidation risk rather than a precise measure, and risk-control systems should systematically reserve a safety margin when using liquidation-heatmap data.

24.1.4 Changes in order book shape

In a millisecond-level high-frequency trading environment, subtle changes in the shape of the order book often lead actual price movements. As the primary providers of liquidity, market makers' perception of market risk is directly reflected in their quoting behavior. When market makers expect short-term volatility to rise, their most immediate response is to reduce order book depth, even before the bid-ask spread widens noticeably. This systematic withdrawal of depth is itself a leading signal that volatility is about to rise [12].

The dynamics of quote skew further reveal market makers' directional expectations. When many market makers simultaneously accumulate thicker resting depth on the bid or the ask, this asymmetry reflects a consensus judgment about the direction of short-term price movement. If such skew is accompanied by a decline in overall depth, it often foreshadows an imminent directional breakout and a sharp amplification of volatility.

In addition, changes in the proportion of ghost liquidity are a key indicator for forecasting the evaporation of volatility. Ghost liquidity refers to resting orders that are observable in the order book but that, because of high-frequency cancellation, cannot in fact be executed. When the share of true depth in the order book falls while ghost liquidity proliferates, the market's capacity to absorb large orders is severely overstated [13]. This divergence between apparent and true liquidity greatly increases the probability that the market will experience a liquidity dry-up and a volatility spike when hit by an external shock. These changes in order book shape typically lead actual increases in volatility by minutes to tens of minutes and are the most informative inputs in ultra-short-term volatility forecasting.

24.1.5 Cross-exchange and cross-asset spillover signals

Liquidity in the cryptocurrency market is highly fragmented, and the process of price discovery for the same asset across different exchanges exhibits small time differences. Under normal market conditions, arbitrageurs maintain tight co-movement of prices across platforms through cross-exchange arbitrage. When the price spread among leading exchanges suddenly widens from the usual few basis points to tens of basis points, however, this often signals that certain large market makers or arbitrage capital have begun to withdraw from the market [14]. An abnormal widening of cross-exchange spreads is an early warning of deteriorating liquidity and imminent systemic volatility. Such widening can also arise, however, from technical factors at a single venue—a matching-engine failure or API latency—which carry no systemic implication. Before treating a widening spread as a warning signal, one must therefore rule out a single-exchange anomaly; the signal acquires systemic-risk value only once spreads are confirmed to have widened simultaneously across several independent platforms.

Cross-asset volatility spillover likewise provides an effective forecasting window. In crypto markets, volatility often transmits along a path spreading from core assets to peripheral assets. If Bitcoin's volatility remains stable while Ethereum's has already begun to rise markedly, this lead-lag relationship greatly increases the probability that Bitcoin's volatility will rise in the future. By constructing a cross-asset volatility spillover

index, a model can capture this micro-level process of risk contagion.

Macro signals from traditional financial markets also affect crypto-market volatility through specific transmission mechanisms. When the VIX index climbs sharply or U.S. Treasury yields move violently, such macro liquidity shocks typically transmit to the cryptocurrency market within hours. For perpetual futures participants, incorporating these cross-market and cross-asset spillover signals into the monitoring framework is a necessary means of forecasting transmission volatility.

24.1.6 On-chain signals

The transparency of the blockchain ledger provides a distinctive on-chain data dimension for volatility forecasting. The fund-transfer behavior of large holders often exerts a disproportionate influence on market liquidity. When a large amount of Bitcoin or Ethereum moves from private cold wallets to centralized exchanges, this large inflow is often interpreted in traditional analytical frameworks as a potential selling signal. Exchange inflows, however, may equally serve many other purposes—collateralized lending, cross-exchange arbitrage, OTC delivery, or replenishing market-making margin—and in recent years their reliability as a selling signal has been substantially challenged. Even so, when cross-validated against other signals, an abnormal large inflow can still serve as a reference factor for downside volatility; conversely, a large inflow of stablecoins to exchanges suggests an accumulation of buying power that may drive upside volatility [15].

Beyond exchange fund flows, a sharp change in the total value locked (TVL) in DeFi protocols is likewise an effective predictor of volatility. When the TVL of a key protocol is withdrawn on a large scale over a short interval, it indicates that capital is undertaking an emergency risk-avoidance move or a reallocation of liquidity. Such a violent restructuring of on-chain liquidity often transmits rapidly, through arbitrage channels, to the perpetual futures markets on centralized exchanges.

On-chain signals are, however, distinctly noisy. Not every large transfer becomes an actual market trade, and on-chain data used in isolation tend to generate a high false-positive rate. Their real value therefore lies in cross-validation with other microstructure signals, such as changes in order book depth or extreme funding rates. Only when signals across several dimensions align does an on-chain volatility forecast become reliable enough to guide trading decisions.

24.2 Volatility forecasting models: adapting classical methods

In traditional financial markets, volatility-forecasting models rely mainly on daily returns computed from daily closing prices. The perpetual futures market, however, has features such as round-the-clock uninterrupted trading, price jumps triggered by high-frequency liquidations, and periodic pulses from funding settlement. These structural differences mean that classical volatility-forecasting models, applied directly to perpetual futures data, often produce systematic forecasting biases. This section evaluates the performance of realized volatility, GARCH models, HAR models, and machine-learning

methods in the perpetual futures market and discusses how to make the necessary structural adaptations.

24.2.1 Realized volatility

Realized volatility estimates daily volatility by summing the squares of intraday high-frequency returns; its theoretical foundation is the theory of quadratic variation. In an idealized continuous-time martingale process, as the sampling frequency tends to infinity, realized volatility converges in probability to the true integrated volatility [16]. In traditional equity markets, a 5-minute or 10-minute sampling frequency typically strikes a balance between microstructure noise and information loss. In the perpetual futures market, however, constructing realized volatility faces three distinctive challenges.

The first is the effect of round-the-clock continuous trading on the definition of a "day." Traditional financial markets have clear opening and closing times, whereas the perpetual futures market lacks natural daily boundaries. Using midnight Coordinated Universal Time (UTC) as the dividing point for the daily cycle can truncate a volatility-clustering process that straddles that moment. Empirical studies show that computing realized volatility over a volume-weighted rolling 24-hour window reflects shifts in market state more smoothly and improves its stability as a predictor [17].

A more serious challenge arises from the price jumps produced by liquidation cascades. As noted earlier, when the market price reaches a dense cluster of liquidation triggers, forced-liquidation orders instantly consume order book liquidity, causing the price to drop precipitously or spike within a very short interval. Left unadjusted, realized volatility computed from high-frequency data becomes dominated by a few extreme 5-minute returns, overstating the market's continuous volatility. To separate the continuous-volatility and jump components, one must introduce bipower variation or truncated realized volatility. By computing the product of the absolute values of adjacent returns, bipower variation asymptotically removes the influence of the jump component, thereby extracting a purer measure of continuous volatility [18].

The periodic pulse injected by funding settlement is a microstructure feature unique to perpetual futures. Around each 8-hour funding settlement window, the concentrated position-building or position-closing of arbitrageurs and directional traders triggers systematic price movements. Whether this mechanism-induced volatility should be removed from realized volatility depends on the forecasting objective. If the goal is to set global margin requirements, this total volatility is a risk that must be covered; but if the goal is to extract a pure informational-volatility signal, then the anomalous returns around funding settlement should be removed through dummy-variable regression or local smoothing techniques, so that the periodic pulse does not contaminate judgments about the true market trend [19].

24.2.2 GARCH models

The GARCH model is the most basic volatility-forecasting tool in traditional finance; its core assumption is that the current conditional variance is jointly determined by past conditional variances and past squared errors. This framework was founded on the autoregressive conditional heteroskedasticity (ARCH) model proposed by Engle

(1982) and generalized by Bollerslev (1986) into its generalized form, GARCH [20][21]. On perpetual futures data, the most basic version of the model clearly captures a pronounced volatility-clustering effect. When the market is hit by an information shock, volatility rises markedly and then exhibits a slowly decaying memory, closely matching the model's characterization of the autoregressive effect [22].

In handling asymmetry, however, the leverage effect common in traditional financial markets exhibits a different directionality in cryptocurrency markets. In equity markets, negative news typically induces a larger rise in volatility than positive news, an asymmetry that can be captured by the exponential GARCH (EGARCH) model or the threshold GJR-GARCH model, which allows for asymmetric effects. In the perpetual futures market, however, because of the large amount of speculative long leverage, the irrational exuberance triggered by positive news can likewise produce a sharp spike in volatility. Empirical tests indicate that volatility asymmetry in cryptocurrency markets is often weak, or in certain market cycles exhibits an inverse leverage effect, in which the volatility increase from a positive return shock may be no smaller than that from a negative shock of equal magnitude. The distinctive price dynamics and volatility-jump characteristics of the Bitcoin derivatives market further corroborate that its volatility structure differs from that of traditional equity markets [23].

The core limitation of the GARCH model in the perpetual futures market lies in its inherent assumption that the volatility process is continuous—that is, that volatility changes smoothly and gradually—whereas liquidation cascades in this market often produce instantaneous jumps in volatility. In extreme conditions, for example, annualized volatility can jump from 20% to 100% within an hour and then fall back rapidly. Conventional continuous models cannot capture such "volatility jumps," leading to a severe underestimation of tail risk when extreme events occur. To remedy this shortcoming, researchers typically relax the assumed return distribution from the normal to a Student's t-distribution or a generalized error distribution to accommodate fatter tails, or directly combine the model with a jump-diffusion process to build a composite volatility model that includes a jump component [24].

24.2.3 HAR models

The HAR model decomposes realized volatility into components at three time scales—daily, weekly, and monthly—and uses each as a predictor, thereby ingeniously capturing the multi-scale volatility structure that arises because market participants trade at different frequencies. High-frequency market makers focus on intraday volatility, intraday arbitrageurs on daily volatility, and medium-term trend traders and long-term holders on weekly and monthly volatility, respectively. This intuitive and computationally simple linear architecture achieves performance comparable to that of complex nonlinear models in forecasting long-memory features [25].

When applied to the perpetual futures market, the model requires domain-specific extensions. The first is to introduce a funding-settlement cycle component. Because funding settlement typically occurs on an 8-hour cycle, the traditional daily, weekly, and monthly partition cannot capture this high-frequency periodic structure. Adding a volatility component based on an 8-hour rolling window can markedly improve the

forecasting accuracy for short-term volatility pulses.

Second, liquidation volume can be integrated into the model as an additional forward-looking predictor. Historical liquidation volume not only reflects past extreme volatility; its spatial distribution (as shown in a liquidation heatmap) also hints at future potential mechanical-volatility risk. Adding lagged relative liquidation intensity as an independent explanatory variable can effectively improve the model's ability to forecast the secondary volatility induced by liquidation cascades.

A cross-market spillover component is another key dimension of extension. Pronounced lead-lag relationships exist within the cryptocurrency market—for example, a jump in Bitcoin's volatility often precedes a rise in volatility across the whole market. Introducing the volatility of a core asset or of a traditional financial market (such as an equity volatility index) as an exogenous variable can capture early signals of transmission volatility, yielding a more comprehensive multi-factor forecasting framework [26].

24.2.4 Machine-learning methods

As high-frequency data have accumulated, machine-learning methods have found increasingly wide application in volatility forecasting. Feature-engineering-based methods, such as XGBoost (extreme gradient boosting) and random forests, can combine multiple market signals nonlinearly. These models can simultaneously process heterogeneous features such as extreme funding rates, distance to a liquidation cluster, and changes in order book depth, and can automatically capture interaction effects among features. For example, when the funding rate deviates extremely from its center and the price approaches a large-scale liquidation trigger point, a tree-based model can identify this specific nonlinear combination and output a very high volatility forecast—something that traditional linear regression models struggle to achieve [27].

Deep-learning methods—such as long short-term memory (LSTM) networks and the Transformer architecture based on the self-attention mechanism—attempt to bypass manual feature engineering and learn the evolution of volatility directly from raw high-frequency return series or order book snapshots. Such end-to-end models have a theoretical advantage in capturing complex temporal dependencies, and they perform especially well on ultra-short-term (for example, 5-minute to 1-hour) microstructure dynamics.

Machine-learning methods, however, face an extremely high risk of overfitting in the perpetual futures market. A model's value lies not in its accuracy in historical backtests but in its robustness when the market microstructure or the macro environment changes. The evolution of the perpetual futures market's rules, the launch of new protocols, and shifts in the composition of participants can all cause a fundamental change in the data distribution. Empirical experience indicates that a complex deep-learning model trained during a particular market cycle may have a predictive half-life of only 3 to 6 months. Once the market enters a new state, such models' forecast errors often expand sharply, sometimes performing worse than the simplest historical-mean model [28].

24.2.5 Model comparison and combination

Different classes of volatility-forecasting model exhibit their own comparative advantages at different time scales. In ultra-short-term forecasting of 5 minutes to 1 hour, machine-learning models based on high-frequency realized volatility and order book microstructure features generally have the highest accuracy. In short- to medium-term forecasting of one day to one week, HAR models that include a jump component exhibit the best robustness, effectively balancing the capture of long memory against responsiveness to sudden shocks. In long-term forecasting beyond one week, performance differences among models gradually narrow, and the mean-reversion tendency of volatility becomes the dominant factor.

Figure 24-3 compares the forecasting accuracy of realized volatility, the GARCH family, HAR models, and machine-learning methods across time windows from ultra-short-term to long-term, and the central lesson is that no single model dominates at every time scale.

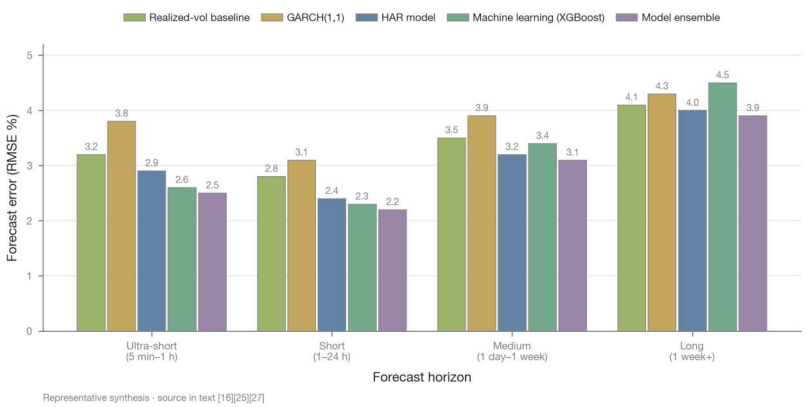


Figure 24-3. A comparison of forecast errors of volatility-forecasting models across time scales (representative composite illustration: the RMSE values shown are representative magnitudes synthesized from the literature, not exact values from any single source; data source: based on empirical studies of volatility forecasting in cryptocurrency and equity markets [16][25][27])

Each class of model has its own zone of comparative advantage: in the ultra-short-term window of 5 minutes to 1 hour, machine-learning models that process order book microstructure features directly (such as XGBoost) exhibit the lowest forecast error among the baseline models; in the short- to medium-term window of one day to one week, HAR models gradually surpass machine learning thanks to their multi-scale decomposition architecture; and once the forecast horizon exceeds one week, the errors of all classes of model converge, with mean reversion becoming dominant and complex models providing little incremental value. This pattern offers clear guidance for a model-combination strategy: emphasize machine learning in the short term, rely on HAR in the medium term, and respect mean reversion in the long term. Table 24-1 further summarizes, at the mechanistic level, the core features and applicable scenarios of each class of model in the perpetual futures market.

Model class	Core mechanism	Perpetual-futures adaptation	Strengths	Weaknesses and limitations
Realized volatility	Sum of squared high-frequency returns	Bipower variation to remove jumps; handling of settlement pulses	Solid theoretical foundation; precise measurement	Only an ex post measure, not a direct forecasting model
GARCH	Autoregression of the conditional variance	Introduce a Student's t-distribution; combine with a jump-diffusion process	Captures volatility clustering; parameters are interpretable	Assumes a continuous process; struggles with liquidation jumps
HAR	Linear combination of multi-scale volatility	Add a funding-settlement cycle component and a liquidation-volume factor	Simple, robust structure; matches the stratification of market participants	Underfits extreme nonlinear interaction effects
Machine learning and deep learning	Nonlinear feature mapping and end-to-end learning	Fuse multidimensional heterogeneous signals (funding rate, order book, etc.)	Excels at ultra-short-term forecasting; captures complex nonlinear patterns	Black-box nature; very high risk of overfitting

Table 24-1. Adaptation and evaluation of volatility-forecasting models in perpetual futures (data source: analysis of model characteristics [22][25][28])

Faced with the limitations of any single model, forecast combination offers a more robust solution. Taking a weighted average of the forecasts of several heterogeneous models can effectively diversify the risk that any individual model fails under a particular market state. The weights can be adjusted dynamically in inverse proportion to historical forecast errors, or set using an exponentially weighted moving average to give greater weight to models that have performed better recently. More important, quantitative models should be embedded within a three-layer risk-control framework that incorporates human judgment and hard stop-loss mechanisms, so that the system retains basic survivability when a model encounters a black swan event or a structural break [29].

24.3 A source-decomposed volatility forecasting framework

Forecasting the total volatility of the cryptocurrency market often runs into a structural dilemma: when we mix volatility from all sources together and try to fit it with a single unified model, volatility components of different natures interfere with one another, degrading forecast accuracy. The solution is not to seek an ever more complex unified model but to adopt a divide-and-conquer strategy. As Chapter 22 discussed, volatility in the perpetual futures market can be decomposed into four distinct sources: informational, liquidity, mechanical, and transmission volatility. Each source has entirely different drivers, time scales, and predictability characteristics. This section shows how to build a separate forecasting model for each of these four sources and then aggregate

them into a forecast of total volatility; this framework constitutes the chapter's core methodological innovation.

The core idea of the source-decomposed framework derives from a basic principle of microstructure theory and signal processing: when multiple independent information sources drive an observed variable simultaneously, the optimal estimation strategy is to extract and process each source separately rather than to invert directly from the mixed signal. This principle is especially applicable in crypto markets, because the four volatility sources differ completely not only in physical mechanism but also in statistical properties. Informational volatility exhibits sparse, pulse-like behavior; liquidity volatility is strongly periodic; mechanical volatility is determined by protocol rules; and transmission volatility manifests as time-lagged cross-asset correlation. Attempting to capture all four properties with a single parameterized model amounts to asking one model to fit four data-generating processes with entirely different statistical characteristics, which inevitably lowers the efficiency of parameter estimation.

Figure 24-4 presents the complete architecture of the source-decomposed volatility forecasting framework in bottom-up logic: from the extraction of alternative signals at the base layer, to the independent modeling of the four volatility sources in the middle layer, to the variance aggregation and total-volatility forecast output at the top layer.

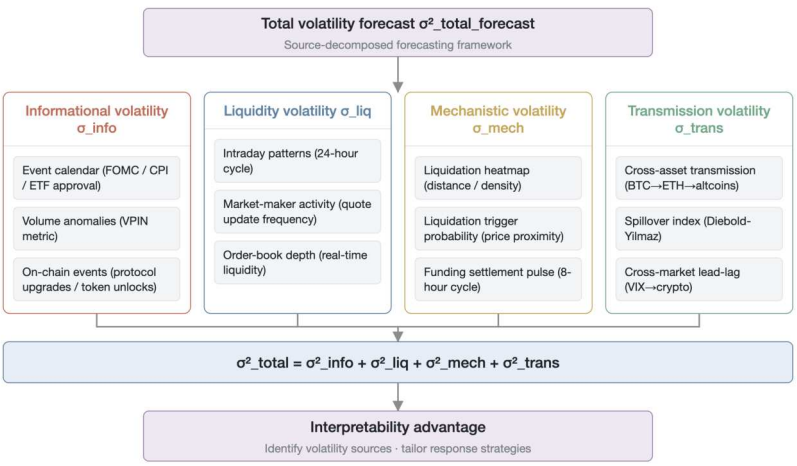


Figure 24-4. The conceptual architecture of the source-decomposed volatility forecasting framework (structural illustration, not measured data; data source: compiled by the author)

The core design principle of this architecture is embodied in three layers of information processing. At the base layer, different market signals (funding rate, liquidation heatmap, order book shape, on-chain data, and so on) are assigned to the volatility-source channel that best matches their physical mechanism, rather than being mixed as input to a single unified model. At the middle layer, each channel adopts the forecasting method best suited to its statistical properties: event-calendar regression for informational volatility, an HAR model for liquidity volatility, a threshold model for mechanical volatility, and a VAR spillover index for transmission volatility. At the top layer, the four independent forecasts are aggregated into a composite forecast of

total volatility through variance summation and calibration of interaction terms. This layered processing ensures that the distinctive time scale and statistical properties of each volatility source are exploited to the fullest, rather than being averaged away in the parameter competition of a unified model.

24.3.1 Forecasting informational volatility

Informational volatility (σ_info) arises from the process by which new information reaches the market and is absorbed into prices. From a forecasting standpoint, informational volatility can be divided into a highly predictable component and an intrinsically unpredictable component. The predictable part consists mainly of the known event calendar, including macroeconomic data releases (such as Federal Open Market Committee (FOMC) decisions, Consumer Price Index (CPI) reports, and nonfarm payrolls data) and scheduled events specific to crypto markets (such as protocol upgrade dates, token unlock schedules, and exchange-traded fund (ETF) approval deadlines). Within the windows of these known events, the rate of information arrival to the market rises systematically.

Ben Omrane et al. (2025) [30] show that the volatility of Bitcoin and Ethereum reacts significantly and asymmetrically to macroeconomic data releases. Analyzing 5-minute-frequency data from 2016 to 2023, the study finds that volatility reacts significantly to only a few news categories, among which U.S. monetary policy news has a longer-lasting effect than the others. U.S. monetary policy news in particular drives volatility higher before, during, and after the announcement, and Ethereum is more sensitive to U.S. macroeconomic announcements than Bitcoin. This rise in volatility ahead of an announcement reflects participants adjusting their positions before the uncertainty is resolved, a phenomenon known as the pre-announcement volatility run-up effect. A forecasting model can therefore use event-calendar regression to automatically raise the baseline estimate of informational volatility during known event windows.

Concretely, one builds a piecewise event-calendar regression model. For each known macroeconomic event, a time window is defined—typically from 2 hours before to 4 hours after the release—and a dummy variable is set within this window. The model can be written as:

$\sigma_{\text{info},t} = \alpha + \sum_i \beta_i \cdot D_{i,t} + \gamma \cdot \text{VPIN}_t + \varepsilon_t$

where D_{i,t} is the dummy variable for event i, and VPIN_t is the volume-synchronized probability of informed trading (VPIN), used to approximate the information arrival rate. The coefficients β_i and γ must be estimated empirically within-sample; their stability depends on the event type and market state, and they should not be applied directly across periods. VPIN was proposed by Easley, López de Prado, and O'Hara (2012) [31]; its essence is to divide trading volume into fixed-size volume buckets and estimate the degree of buy-sell order imbalance within each bucket to infer the active share of informed traders. Compared with the traditional probability of informed trading (PIN) model, it requires no maximum likelihood estimation and is more computationally efficient, and its volume-synchronized nature is naturally suited to the vast differences in activity across the 24/7 crypto market, avoiding the signal distortion of fixed time windows. This hybrid approach captures both the structured

impact of known events and, through VPIN, changes in the information arrival rate outside event periods. One caveat is that in crypto markets VPIN is exposed to the contamination of volume by wash trading: at some exchanges wash trading may account for a substantial share, so a VPIN built on unfiltered volume develops a systematic bias. In practice, one should use wash-trading-filtered volume, or data from exchanges with a low share of wash trading, as input.

For informational volatility outside event periods, although there is no explicit calendar guidance, its baseline level is determined by the market's actual information arrival rate. In traditional microstructure theory, VPIN and volume anomalies are widely used to approximate the information arrival rate. When volume deviates markedly from its historical mean, or when buy and sell order flow becomes severely imbalanced, this usually signals an inflow of undisclosed information and thus foreshadows a near-term rise in informational volatility. The unpredictable component includes sudden "black swan" events such as regulatory raids, exchange bankruptcies, and protocol hacks. The timing of such events cannot be foreseen, so they cannot be incorporated into a point-forecast model and can only be addressed through a portfolio-level tail-risk budget.

In practice, on-chain events in crypto markets—such as protocol upgrades, major governance votes, and token unlocks—are also highly predictable. The dates of these events are typically announced weeks or months in advance, allowing participants to adjust their risk exposure ahead of time. Ethereum's major upgrades, such as the Shanghai and Dencun upgrades, for example, had their dates fixed months in advance, giving forecasting models a clear timing for the information shock.

24.3.2 Forecasting liquidity volatility

Liquidity volatility (σ_liq) is caused by changes in market depth and by market-maker behavior. Unlike informational volatility, liquidity volatility contains a large, highly predictable component. Most notable is the intraday pattern: the level of liquidity in crypto markets fluctuates with high regularity over a 24-hour cycle. Hansen et al. (2024) [32] find that crypto markets exhibit a clear intraday cycle in volatility and liquidity, a periodicity that is closely related to the activity of traders in different geographic time zones and to the execution times of algorithmic trading. For example, at the handover between the Asian and the European-American trading sessions, liquidity often reaches a trough, so that even a small order can trigger a large price move. The study further finds that crypto-market volatility and liquidity metrics exhibit pronounced intraday variation, which means that liquidity is highly predictable.

The mechanisms that produce this intraday liquidity pattern are several. First, traders in different time zones have clearly defined active hours: Asian traders are most active during Asian working hours, while European and American traders dominate during their working hours. During the handover between time zones (such as the gap between the Asian close and the European-American open), the number of participants falls to a trough, market makers' incentive to quote declines accordingly, and liquidity tightens noticeably. Second, the execution times of algorithmic trading are also highly regular: many institutional traders execute large orders at specific times (such as on the hour or half-hour), causing a sudden rise in liquidity demand at those points. Third, funding

settlement times (though belonging to the category of mechanical volatility) also affect liquidity, because traders adjust their positions around settlement.

Beyond static time-of-day effects, market makers' real-time activity provides a dynamic forecasting signal. If the frequency with which market makers update their quotes suddenly drops, or the apparent depth of the order book begins to shrink, this usually signals that market makers perceive some latent risk and are reducing their risk exposure. Such defensive behavior is itself a strong leading signal that liquidity volatility is about to rise. In practice, one can track market makers' activity in real time by monitoring the quote refresh rate of the order book. Taking the BTC perpetual futures on a leading exchange as an example, when the quote refresh rate falls to 10% to 20% of its normal level, this usually presages an impending liquidity crunch.

In terms of forecasting method, the classic HAR model performs well in handling liquidity volatility. Brauneis and Sahiner (2026) [33] demonstrate that, although the HAR model effectively captures the clustering and long-memory features of volatility, forecast accuracy can be further improved by introducing machine-learning methods such as LightGBM (light gradient boosting machine) and XGBoost. The basic form of the HAR model is:

$$RV_{t+1} = \alpha + \beta_d \cdot RV_t + \beta_w \cdot \overline{RV}_{t-6:t} + \beta_m \cdot \overline{RV}_{t-29:t} + \varepsilon_t$$

where RV denotes realized volatility, and the subscripts denote the daily, weekly, and monthly scales, respectively. Here a crypto-market convention is used: the weekly term $\overline{RV}_{t-6:t}$ is the average over the most recent seven calendar days and the monthly term $\overline{RV}_{t-29:t}$ is the average over the most recent 30 calendar days—both including the current period t and thus consistent with the daily term RV_t—rather than the five and 21 trading days used in traditional finance. Even so, crypto-market participants still behave according to calendar cycles (weekend trading volume falls markedly, and intraday liquidity fluctuates regularly under the influence of geographic time zones), so the daily/weekly/monthly decomposition remains empirically effective; researchers need only note the difference in time-scale definitions when making cross-market comparisons. This multi-scale decomposition is naturally suited to forecasting liquidity volatility, because liquidity changes contain intraday, day-to-day, and week-to-week periodicity simultaneously. Within the source-decomposed framework, real-time order book depth, bid-ask spread skew, and market-maker activity can be fed as additional factors into an HAR or machine-learning model to fit the liquidity-volatility component specifically; when the state machine defined in Chapter 23 enters the "stressed" state, the model should automatically raise the baseline expectation for liquidity volatility.

24.3.3 Forecasting mechanical volatility

Mechanical volatility (σ_mech) is a feature unique to the perpetual futures market, created endogenously by protocol rules and the leverage-liquidation mechanism. Of the four volatility sources, mechanical volatility contains the most predictable deterministic component: the funding-settlement pulse. Funding settlement times are known: at traditional centralized exchanges such as Binance, settlement occurs every 8 hours (for example, at 00:00, 08:00, and 16:00 UTC), but a new generation of protocols has markedly increased the frequency—Hyperliquid and dYdX v4 both settle hourly, and

Bybit has moved some contracts to settlement every 4 hours. Differences in settlement frequency directly shape the character of the mechanical-volatility pulse: a higher frequency means a smaller amplitude per pulse (because rate deviations are corrected over shorter intervals) but a greater frequency of pulses. Regardless of the specific frequency, the rate level is already observable in the order book and on derivatives data platforms before settlement. When the rate reaches an extreme level, arbitrageurs and directional traders concentrate their closing and opening of positions around settlement, either to avoid paying the high funding cost or to capture funding income. This foreseeable concentration of order flow inevitably produces a volatility pulse within the settlement window.

The settlement pulse can be forecast with a simple yet effective method. First, compute the deviation of the current rate from its historical mean; when the rate is at an extreme level (for example, an annualized rate above 100% or below −50%), the amplitude of the settlement pulse increases substantially. Second, monitor the speed of rate changes; the faster the rate moves, the stronger the market's directional bias and the greater the closing pressure at settlement. From the absolute level and rate of change of the current rate, the forecasting model can reasonably estimate the direction and magnitude of the settlement pulse. A simplified form of the model is:

$$\sigma_{\text{mech,settlement}} = \alpha + \beta_1 \cdot |f_t| + \beta_2 \cdot |\Delta f_t| + \beta_3 \cdot OI_t$$

where f_t is the current rate, Δf_t is its change, and OI_t is total open interest. The coefficients of this equation must be calibrated empirically within-sample and vary with the exchange's funding-settlement frequency and market state; they should serve as directional guidance rather than a ready-made fixed rule.

Another key source of mechanical volatility is the liquidation cascade. Although the triggering of a liquidation cascade is abrupt, its underlying risk distribution is highly transparent. By aggregating open-interest data across the entire market, a liquidation heatmap can precisely compute the scale of liquidations distributed at different price levels. The distance from the current market price to the nearest "liquidation cluster" is a powerful conditioning factor for forecasting mechanical volatility. For example, if the current Bitcoin price is $45,000, with $10 million of clustered long liquidations at $44,500 and $8 million of clustered short liquidations at $45,500, then these two price zones become high-risk areas.

A model for forecasting liquidation cascades should output a conditional volatility whose magnitude depends on the distance between the price and a liquidation cluster. A common approach uses a Gaussian kernel function to measure this distance:

$$\sigma_{\text{mech,liquidation}} = \sigma_{\text{base}} \cdot \exp\left(-\frac{(P_t - L^*)^2}{2h_d^2}\right)$$

where h_d is the bandwidth of the Gaussian kernel, characterizing the price scale over which a liquidation cluster exerts influence; its dimension is the same as that of P_t and L^* (both a quote currency such as USD), not that of volatility. The larger the bandwidth, the more widely the cluster's influence spreads along the price axis. This form ensures that as the price approaches a liquidation cluster, the estimated mechanical volatility rises sharply. In real markets, liquidation clusters usually exist at several price levels simultaneously, and clustered long liquidations (concentrated below the

current price) and clustered short liquidations (concentrated above it) are distributed asymmetrically. A more complete form of the model should therefore perform a kernel-density summation over all known liquidation clusters:

$$\sigma_{\text{mech}} = \sigma_{\text{base}} + \sum_{k} w_k \cdot \exp\left(-\frac{(P_t - L_k)^2}{2h_{d,k}^2}\right)$$

where L_k is the center price of the k-th liquidation cluster, $h_{d,k}$ is the bandwidth corresponding to the k-th cluster (again a price scale), and the weight w_k is determined by the estimated liquidation scale of that cluster. Note that, whereas the single-cluster equation treats σ_{base} as a multiplicative factor, the multi-cluster equation makes it an additive base: the kernel-summation term is added directly on top of the baseline volatility, so the two equations do not reduce to the same form even when only a single cluster is present. The liquidation-cluster data used here come from third-party estimates rather than precise exchange disclosures (as discussed in Section 24.1.3), so the model output is itself an estimate.

Forecasting liquidation cascades, however, has an inherent nonlinear boundary: whether a cascade will start can be predicted from the probability that the price approaches a liquidation cluster, but how violent the cascade will be once it starts is difficult to estimate precisely. This is because the intensity of a cascade depends on the order book liquidity at the moment of triggering, and that liquidity evaporates rapidly once the cascade begins. This "liquidity evaporation" was especially evident in the March 2020 crypto crash, when Bitcoin fell by more than 50% within a few hours and triggered a liquidation cascade of several billion dollars. A model for mechanical volatility should therefore output an asymmetric probability distribution: as the price approaches a liquidation cluster, the model must not only raise the mean forecast of volatility but also substantially increase the probability weight on extreme tail movements. This is usually achieved by using a Student's t-distribution or a generalized Pareto distribution.

24.3.4 Forecasting transmission volatility

Transmission volatility (σ_trans) reflects the spread of a shock from one asset or market to another. This transmission is not instantaneous but involves a physical delay of minutes to hours, and that delay provides an exploitable time window for volatility forecasting. At the macro level, volatility in traditional financial markets often leads that in crypto markets. For example, when the VIX—the implied volatility of S&P 500 options—rises sharply or U.S. Treasury yields jump, crypto-market volatility typically responds within minutes to hours, and this transmission lag has been shrinking steadily as institutional capital enters crypto through the ETF channel. Moreover, the direction of transmission is not always one-way: in certain events, such as the 2022 FTX collapse, crypto-market volatility can also propagate back to traditional financial assets. This increasingly tight and bidirectional cross-market linkage means that forecasting transmission volatility must incorporate the two-way spillover paths between traditional finance and crypto markets.

Within crypto markets, clear lead-lag relationships also exist among assets. As the market's liquidity anchor and the representative "risk asset," Bitcoin's volatility increases

often lead Ethereum's and then propagate to less liquid altcoins. This transmission lag is typically minutes to hours, depending on each asset's liquidity depth and the attention traders pay to it. Shahrour et al. (2025) [34] quantify this cross-market spillover using a time-varying parameter vector autoregression (TVP-VAR) technique and find that, within their particular sample period and asset set, Bitcoin and the NEAR protocol are the main transmitters of volatility, influencing 68.17% and 71.63% of other assets' volatility, respectively, whereas traditional financial instruments such as Treasury bills act more as net receivers of volatility. The finding that NEAR's spillover influence exceeds Bitcoin's may reflect the particular asset selection and sample-period bias of that study; because NEAR's market capitalization and liquidity are far smaller than BTC's, whether such a high spillover coefficient is robust across different samples remains to be tested.

To exploit this transmission mechanism systematically within a forecasting model, one can introduce a real-time version of the Diebold-Yilmaz spillover index. Through variance decomposition, this index dynamically measures "who is transmitting volatility to whom" and the strength of that transmission. The Diebold-Yilmaz index is computed from the variance decomposition of a vector autoregression (VAR, to be distinguished from value at risk, VaR) model:

$$S = \frac{\sum_{i \neq j} \theta_{ij}}{\sum_{i,j} \theta_{ij}} \times 100$$

where θ_{ij} is the contribution of a shock to asset i to the variance of asset j, and S is the Diebold-Yilmaz spillover index. When the monitoring system detects a sharp rise in the overall spillover index, or a marked strengthening of the directional spillover from a particular asset (such as Bitcoin) to others, the forecasting model should automatically raise the estimated transmission volatility of the target asset. This approach converts changes in the cross-market topological structure into direct inputs for volatility forecasting.

24.3.5 Aggregating the four-source forecast

Once the four volatility sources have been forecast independently, the forecast of total volatility can be obtained by summing the variances of the individual forecasts (plus empirically calibrated interaction terms). This source-decomposed aggregation is significantly superior, both in theory and in practice, to a unified model that forecasts total volatility directly. The aggregation formula is:

$$\hat{\sigma}^2_{\text{total}} = \hat{\sigma}^2_{\text{info}} + \hat{\sigma}^2_{\text{liq}} + \hat{\sigma}^2_{\text{mech}} + \hat{\sigma}^2_{\text{trans}} + 2\rho \cdot \hat{\sigma}_{\text{info}} \cdot \hat{\sigma}_{\text{liq}} + \cdots$$

where ρ is an empirically calibrated interaction coefficient. In practice, these interaction terms are usually small, because the drivers of the four volatility sources are relatively independent.

The reasonableness of the four-source independence assumption depends on the state the market is in. In a calm state, the drivers of the four volatility sources do indeed differ fundamentally: informational volatility is driven by the news arrival rate, liquidity volatility by market-maker behavior and the intraday cycle, mechanical volatility by funding settlement and liquidation trigger conditions, and transmission volatility by cross-market fund flows. Under normal market conditions, these drivers are weakly

correlated, so the interaction terms contribute little to the total forecast.

In a crisis state, however, the independence assumption may be significantly violated. When a major information shock (such as a sudden regulatory action) simultaneously triggers a sharp price drop, the withdrawal of liquidity providers (liquidity volatility), the triggering of liquidation clusters (mechanical volatility), and cross-asset panic transmission (transmission volatility) may all activate synchronously within minutes. In such an extreme scenario, the interaction terms among the four volatility sources may account for a significant share of total variance, and simple variance summation will underestimate the true total volatility. Both the "Black Thursday" of March 2020 and the 2022 FTX collapse exhibited this feature of synchronous multi-source activation. Feasible ways to handle this limitation include introducing state-dependent interaction coefficients that automatically amplify the interaction-term weights when the state machine identifies a "stressed" or "crisis" state, or adopting a conditional-correlation structure (such as a dynamic conditional correlation (DCC) model) to model the time-varying correlations among the four volatility sources. This limitation is a known structural constraint of the framework, and readers should pay particular attention to the reliability of the model's output in a crisis state.

Moreover, the decomposition into four volatility sources is not, in the statistical sense, a unique decomposition. Different decomposition methods (such as principal component analysis, independent component analysis, or attribution based on different microstructure feature sets) may yield different definitions of the "four sources" and different independence properties. This framework's decomposition rests on the physical-mechanism prior classification proposed in Chapter 22—information arrival, liquidity supply, protocol rules, and cross-market transmission—which gives it a relatively stable economic meaning across different samples and model specifications rather than being a purely statistical artifact. In day-to-day operation, however, precise attribution still faces the challenge of multicollinearity: a single large liquidation (mechanical) may simultaneously alter order book depth (liquidity) and transmit through cross-asset positions to other markets (transmission), so the boundaries of attribution are inevitably blurred. In practice, source-decomposed forecasts should be treated as an approximate tool that provides directional guidance rather than a quantitative decomposition accurate to the decimal point.

Table 24-2 systematically compares the differentiated positioning of the four volatility sources within the forecasting framework along three dimensions—predictability profile, core predictors, and optimal forecasting method: the four have almost no overlap in the source of their predictability, their time scale, or their optimal mathematical tool.

Volatility source	Predictability profile	Core predictors	Optimal forecasting method
Informational volatility (σ_info)	High during event windows, low otherwise	Macro event calendar, VPIN, on-chain events	Event-calendar regression + order-flow imbalance model
Liquidity volatility (σ_liq)	Strongly periodic, highly predictable	Order book depth, market-maker quote frequency, time of day	HAR model with depth features / machine learning
Mechanical volatility (σ_mech)	Conditionally triggered, highly nonlinear	Funding rate level, distance on the liquidation heatmap	Threshold model + extreme value theory (EVT)
Transmission volatility (σ_trans)	Time-lagged, traceable	VIX index, cross-asset lead-lag relationships	Diebold-Yilmaz spillover index + VAR

Table 24-2. A comparison of the forecasting characteristics and methods of the four volatility sources (data source: compiled by the author)

The reason divide-and-conquer outperforms a unified model is, fundamentally, that it avoids feature confusion during model learning. A unified GARCH or machine-learning model tries to make the same set of parameters learn the 8-hour funding pulse, the 24-hour intraday liquidity pattern, and the multi-day macro transmission all at once. Such multi-scale, multi-driver mixed input often leaves the model wavering between states, ultimately producing a mediocre mean forecast. Source-decomposed forecasting, by contrast, allows the most suitable mathematical tool and time scale to be chosen for each physical mechanism.

Another core advantage of source-decomposed forecasting is its interpretability. When a unified model forecasts that tomorrow's volatility will reach 80%, a market maker knows only that risk has increased, not how to respond. A source-decomposed model, by contrast, can state explicitly whether the rise in volatility is because "the price is only 2% away from a large-scale liquidation cluster" (mechanical volatility dominant), because "there is an FOMC rate decision tomorrow evening" (informational volatility dominant), or because "the VIX index has spiked suddenly and cross-market spillover has strengthened" (transmission volatility dominant). Different attributions lead directly to different trading and risk-control decisions: if liquidation risk dominates, a trader may choose to reduce leverage or adjust the direction of the position; if a macro event dominates, the trader may choose to increase position size but lower leverage; if cross-market transmission dominates, the trader may choose to hedge the exposure to the source asset. This is precisely the core value of turning volatility theory into an engineering application.

Figure 24-5 makes the case for the source-decomposed framework along two dimensions: the left panel compares the forecast errors of the source-decomposed aggregation method against a traditional unified GARCH model across four time windows, while the right panel reports the predictability of each of the four volatility sources—together giving intuitive support for the decomposition strategy.

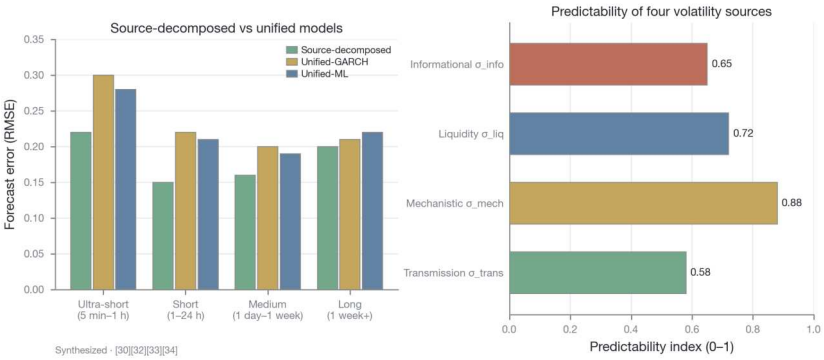


Figure 24-5. A performance comparison of source-decomposed and unified-model forecasting, and the predictability of each source (illustrative synthesis extrapolated from the literature, not independent empirical evidence: the RMSE and predictability values shown are relative magnitudes, not measured data; data source: the author's synthesis based on the literature [30][32][33][34])

Building on the above literature's empirical findings on crypto-market volatility forecasting, this book extrapolates the performance advantage of the source-decomposed aggregation method over a unified model. In ultra-short-term (5 minutes to 1 hour) and short-term (1 to 24 hours) forecasting, the source-decomposed method is expected to achieve a significant improvement in error, because at these time scales the periodic features of liquidity volatility and mechanical volatility are especially pronounced and the source-decomposed method can capture these regularities precisely. In medium- and long-term forecasting, the advantage is expected to weaken but to still preserve some improvement in error. These performance estimates are conceptual judgments synthesized from the literature, not independent empirical conclusions.

Moreover, based on the physical mechanisms of each volatility source and the qualitative analysis in the literature, the predictability of the four volatility sources differs systematically, ranging from highest to lowest as follows: mechanical (funding settlement times are fixed and the liquidation heatmap is transparent), liquidity (the intraday pattern is highly regular; Hansen et al. (2024) [32] document a pronounced intraday periodicity), informational (the event calendar is known, but information arrival outside events remains random), and transmission (the lag and strength of cross-market transmission vary considerably). This systematic difference in predictability across sources further supports source-decomposed modeling over unified modeling.

The performance advantage of the source-decomposed forecasting framework described above currently rests on theoretical extrapolation and a synthesis of the literature and awaits independent empirical testing. An ideal validation scheme should contain the following elements: at the sample-asset level, coverage of perpetual futures data for at least BTC, ETH, and two or more altcoins; in terms of time span, coverage of a complete bull-bear cycle (at least three years is recommended) including at least two known liquidation-cascade events; at the benchmark-model level, a standard GARCH(1,1), an HAR-RV, and a unified machine-learning model as control groups; and in terms of evaluation metrics, simultaneous reporting of root mean square error (RMSE), mean absolute error (MAE), and a quantile loss function, evaluated separately across different

time scales (intraday, daily, weekly). This validation work constitutes the key follow-up research direction for taking the framework from theory to empirical evidence.

The four outputs of the source-decomposed forecasting framework—the independent estimates of informational, liquidity, mechanical, and transmission volatility—constitute the core inputs to the subsequent three-layer decision system. At the micro level, market makers dynamically adjust their quoting strategy based on real-time identification of the volatility source; at the meso level, directional traders and portfolio managers use source-decomposed forecasts for position sizing and tail-risk management; at the macro level, exchanges and protocol designers embed the forecast signals in dynamic margin calibration and liquidity early-warning systems. The following three sections develop the specific applications of these three levels in turn.

24.4 Applying volatility in market-making decisions

Market makers are the most direct consumers of volatility forecasts, and every gain in forecast quality translates into a better quoting decision. This section connects the market-making profit equation of Chapter 19 with the volatility analysis of Chapters 22 through 24 to show how such knowledge enters real-time decision-making.

24.4.1 Volatility and the optimal spread

Starting from the profit equation of Section 19.1.4, the market maker's profit structure comprises four core parts: market-making profit equals spread revenue minus adverse-selection loss, minus inventory-volatility loss, and finally minus institutional cost. This equation provides a basic framework for understanding how volatility affects market-making decisions. The different components of volatility have asymmetric effects on the terms of the profit equation. Adverse-selection loss is positively related to informational volatility: when informational volatility rises, informed traders are more active in the market and the adverse-selection risk faced by market makers rises accordingly. Inventory-volatility loss is positively related to total volatility, and the relationship between the two is markedly convex [35]. When total volatility doubles, the risk cost of a market maker holding inventory rises by more than a factor of two. In addition, institutional cost is positively related mainly to mechanical volatility: extreme events such as liquidation cascades significantly increase a market maker's potential loss in the tail.

The setting of the optimal spread must dynamically reflect this nonlinear risk structure. Theoretical models show that a market maker's optimal spread should widen nonlinearly as volatility rises. Wyart et al. (2008) established, in order-driven markets, that the bid-ask spread is approximately proportional to volatility [35], and under an inventory-risk-based optimal market-making framework the optimal spread contains a nonlinear term directly related to the variance of volatility [36]. This book adopts an illustrative elasticity range slightly above 1 (roughly 1.2 to 1.5) to characterize the convex sensitivity of the spread to volatility; this range is an illustrative modeling assumption rather than a figure from any single empirical source, and given the higher leverage and more extreme liquidity swings of perpetual futures, the actual elasticity

coefficient may deviate from this range and awaits independent validation in the crypto perpetual futures market. This implies that when realized volatility rises by 10%, a market maker seeking to remain risk-neutral should widen the optimal quoted spread by 12% to 15%. Figure 24-6 shows the convex relationship between volatility and the optimal market-making spread derived from this theory.

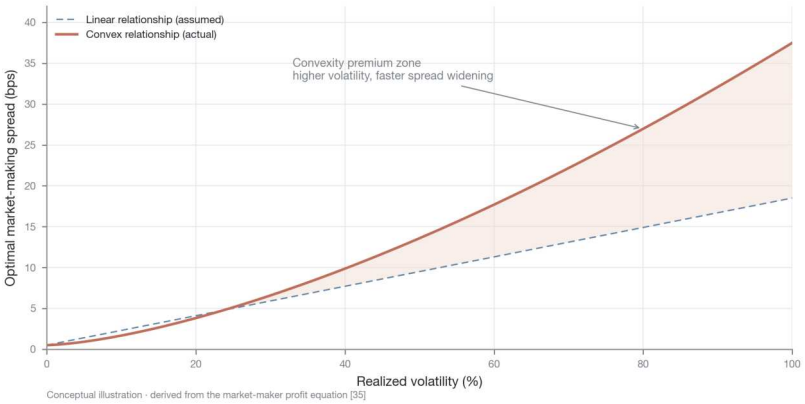


Figure 24-6. The convex relationship between volatility and the optimal market-making spread (theoretical-derivation illustration, not measured data: the elasticity of roughly 1.2 to 1.5 is an illustrative modeling assumption and awaits independent validation for crypto perpetual futures; data source: derived theoretically from the market-maker profit equation [35])

The convex curve in Figure 24-6 carries a concrete practical implication: spread adjustment cannot follow a simple linear mapping but must build in acceleration, so that once the volatility forecast breaks through a threshold, the pace at which the spread widens itself accelerates. A market-making system that handles this convexity poorly will suffer severe adverse-selection losses in the critical interval where volatility jumps from moderate to extreme.

24.4.2 State identification and strategy switching

The state-machine model proposed in Chapter 23 has direct engineering value in market-making strategy. Rather than using one set of static parameters to cope with all market environments, a market maker switches dynamically among three distinct market-making modes based on the volatility state. In the calm state, where volatility is at a historical low with no clear upward trend, the market-making system runs in standard mode, characterized by extremely narrow spreads, deep order book depth, and an active willingness to fill orders, with the goal of maximizing spread revenue. When volatility breaks through the first threshold and enters the stressed state, the system automatically switches to defensive mode. Here the spread typically widens by 50% to 100%, the resting depth at each price level is halved, and inventory limits are tightened sharply. If volatility continues to surge and triggers the crisis state, the market-making system enters minimal-market-making or market-making-suspension mode, quoting extremely wide spreads or withdrawing orders entirely, placing the protection of capital first.

Figure 24-7 integrates the switching logic, trigger conditions, and effects on quoted

spread, order book depth, and inventory limits of these three market-making modes into a single decision flowchart. State identification in real time draws on signals across several dimensions: the system tracks not only the level and trend of current realized volatility but also the leading signals discussed in Section 24.1—the proximity of the liquidation heatmap and micro-level changes in order book shape—which together let the state machine estimate the most probable current market state on the fly.

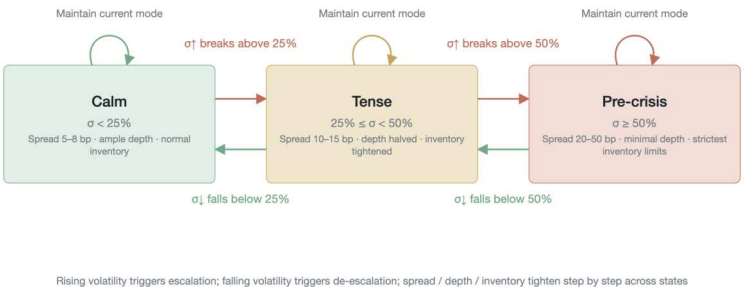


Figure 24-7. Volatility-triggered switching of a three-mode market-making strategy (structural illustration, not measured data: the three bands $\sigma < 25\%$, 25% to 50%, and $\sigma \geq 50\%$ are illustrative regime thresholds introduced by this figure and are not stated in the main text; data source: extrapolated from an application of the state-machine model [37])

The figure makes clear that switching among the three modes is not a symmetric design. The threshold for switching from calm to stressed is deliberately set at a relatively low volatility level, making the system highly sensitive to rising risk; the threshold for returning from stressed to calm is set markedly lower, forming a hysteresis window. The core reason for this asymmetry is that a market maker would rather bear the potential profit loss of defending too early than be exposed to enormous adverse-selection risk when volatility erupts; the system must confirm that volatility has substantively fallen back and that liquidation risk has completely receded before it gradually relaxes inventory limits and narrows spreads, so as to prevent a second loss from resuming normal market-making too early during a volatility aftershock.

24.4.3 Distinguishing liquidation volatility from informational volatility

When facing a sharp price drop, a market maker confronts the core difficulty described in Section 19.4.3: it must judge, within an extremely short interval, whether the current plunge is driven by informed trading or triggered by a liquidation cascade. The optimal responses to these two cases are exactly opposite. If the drop is driven by informed trading, this means the asset's fundamentals have changed, the price will not revert in the short term, and a market maker that fills orders will face a certain adverse-selection loss, so the optimal strategy is to withdraw immediately. If the drop is driven by a liquidation cascade—a form of mechanical volatility—the price often rebounds once the liquidation selling pressure is exhausted, so although filling orders entails short-term inventory risk, the expected return is very high.

Decomposing the volatility signal provides a key tool for distinguishing these two sce-

narios. An informed-trading-driven drop is usually accompanied by a rise in the volume-synchronized probability of informed trading, an increase in permanent price impact, and a refusal of the price to mean-revert after the shock, which manifests as a marked rise in the informational-volatility component. A liquidation-driven drop, by contrast, manifests as a spike in liquidation volume, a sharp fall in open interest, and an amplification of transitory price impact, corresponding to a surge in the mechanical-volatility component [38]. Figure 24-8 uses a decision tree to show how a market-making system, based on real-time signals across multiple dimensions—including the rate of change of liquidation volume, the change in open interest, the VPIN value, and the persistence of price impact—makes a millisecond-level choice between two entirely different response strategies.

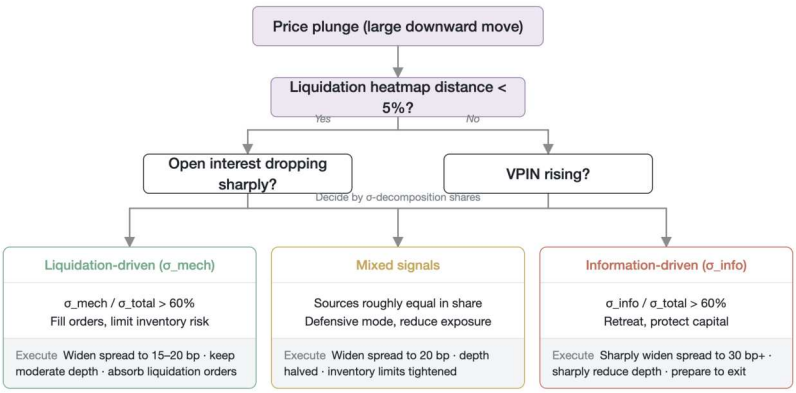


Figure 24-8. Real-time signals for distinguishing liquidation volatility from informational volatility (structural illustration, not measured data: the ratio of mechanical volatility to total volatility uses 60% as a simplified illustrative threshold; data source: extrapolated from the volatility-decomposition framework [38])

The core branching logic of the decision tree is that liquidation-driven and information-driven drops are highly distinguishable at the microstructure level: a liquidation event is accompanied by a sharp contraction of open interest, whereas in an information event open interest usually remains stable or even increases. The market-making system accordingly computes the ratio of mechanical volatility to total volatility in real time; if this ratio exceeds a preset threshold (60% as a simplified example), it judges the move to be driven mainly by liquidations and leans toward staying in the market and absorbing liquidation orders. Conversely, if the share of informational volatility exceeds the threshold, it judges the move to be an information-driven structural decline and triggers the withdrawal mechanism. In practice, real systems use a probabilistic soft threshold—a Bayesian posterior probability, say—rather than a hard cutoff, because the four-source decomposition must run in real time, entails a computational delay of hundreds of milliseconds to seconds, and loses considerable accuracy in extreme conditions.

24.4.4 Risk management for protocol market-making vaults

Automated market-making mechanisms at the protocol level face challenges different from those of traditional market makers. Take Hyperliquid's Hyperliquidity Provider (HLP) vault as an example: as the protocol's native liquidity provider, it bears responsibility for supplying baseline liquidity to the entire platform. Unlike traditional high-frequency market makers, which can exit the market at any time, HLP is designed to provide liquidity continuously in normal operation [39]. Although liquidity providers (LPs) can submit redemption requests (subject to a lock-up period and a queue), and although in extreme circumstances governance mechanisms such as validator voting can intervene to adjust operating parameters (the JELLY incident of March 2025 being one such intervention), HLP is designed to keep running under the vast majority of market conditions. Being unable to withdraw, however, does not mean abandoning risk management; HLP achieves a volatility-driven degradation mode by dynamically adjusting its strategy parameters.

When the risk indicators output by the volatility-forecasting model exceed a warning threshold, the HLP vault automatically triggers a degradation mechanism. This mechanism comprises three levels of response: first, the quoted spread automatically widens several-fold according to the volatility level, to compensate for the sharply rising adverse-selection risk; second, the resting depth on each side of the order book shrinks to a small fraction of its normal level, strictly limiting the maximum inventory that could accumulate in a one-sided market; and finally, the system suspends market-making on high-volatility altcoins, concentrating its limited capital on the most liquid, most controllable core assets [40]. This do-not-withdraw-but-degrade philosophy is precisely the concrete implementation of the crisis behavior of market-making vaults discussed in Section 21.4.2. The volatility-forecasting model constitutes the core engine driving this degradation mechanism, ensuring that the protocol dynamically balances the provision of baseline liquidity against the protection of liquidity providers' capital. A decentralized market-making vault also faces a risk dimension that centralized exchange (CEX) market makers do not: oracle dependence. HLP's mark-price computation and liquidation triggering rely on oracle inputs, and if the oracle is manipulated or delayed, the vault's entire risk-control logic can be bypassed. In the 2023 Mango Markets incident, an attacker manipulated the oracle price and caused protocol-level losses exceeding $100 million, showing that the trust-stack risk faced by decentralized exchange (DEX) market-making vaults (from oracle to mark price to liquidation triggering) differs structurally in nature from the risk faced by CEX market makers.

24.5 Applying volatility in directional trading and portfolio management

Volatility is not only a core input for market makers setting spreads and managing inventory; directional traders and portfolio managers likewise rely on volatility information to decide on position size, risk exposure, and stop-loss settings. In the perpetual futures ecosystem, which lacks a deep options market, participants cannot isolate and trade volatility risk directly through options. This missing market structure requires

traders and portfolio managers to use indirect means to achieve volatility adjustment, volatility timing, and tail-risk management. The core argument of this section is that the application of volatility information is not an optional optimization but a basic competency that perpetual futures participants must master. For individual traders and institutional portfolio managers alike, the degree to which they understand volatility directly determines their probability of survival in a high-risk market.

24.5.1 Volatility-adjusted position sizing

The Kelly formula is a classic method in quantitative trading for determining the optimal position size; its core logic is that the optimal position size is proportional to the strategy's reward-to-risk ratio [41]. In the standard discrete form, the Kelly formula is $f^* = \frac{p \cdot b - q}{b}$, where p is the strategy's win probability, q is the loss probability $(1-p)$, and b is the average win-loss ratio. In the continuous case, when returns follow a normal distribution, the Kelly criterion generalizes to $f^* = \mu/\sigma^2$, where μ is the expected excess return and σ^2 is the return variance. The logic of the generalization from discrete to continuous is that as trading frequency tends to infinity and the return per trade tends to zero, the limiting form of the discrete Kelly formula converges to the continuous expression μ/σ^2, which also intuitively shows that the optimal position is proportional to expected return and inversely proportional to the square of volatility (on the effect of betting frequency on the long-run performance of a Kelly strategy, see [42]). The deeper implication is that, by maximizing the expected growth rate of log wealth, the formula identifies the position size that minimizes the probability of ruin in long-run trading [41]. In the crypto perpetual futures market, however, the return distribution exhibits pronounced fat tails (with kurtosis far above that of the normal distribution), and applying the classical Kelly formula directly often leads to excessive position sizes. In practice, one usually applies a conservative correction via fractional Kelly—investing only 50% to 75% of the Kelly-optimal fraction—in exchange for a markedly lower probability of ruin and a smoother equity curve. The theoretical rationale for fractional Kelly is that when the return distribution departs from the normal assumption (especially in the presence of fat tails), the actual risk of full Kelly is far higher than the model predicts, and moderately reducing the fraction can greatly improve worst-case performance at the cost of only a small sacrifice in long-run growth rate [43].

Volatility adjustment in perpetual futures requires the four-source-weighted forecast rather than simple historical volatility. As Chapter 22 discussed, the four volatility sources carry entirely different implications for trading decisions. Informational volatility may accompany a directional trading opportunity, because the price is moving toward a new fundamental equilibrium; mechanical volatility (such as a liquidation cascade) and liquidity volatility, by contrast, are usually pure microstructure noise. If the forecast shows that the future rise in volatility is driven mainly by mechanical volatility, the trader should sharply reduce position size to avoid being liquidated by disorderly price jumps; but if the rise is caused by an information shock and the trader's strategy itself depends on capturing fundamental trends, the reduction in position size should be relatively small. The two differ significantly in the direction and magnitude of their ef-

fect on position size. This source-decomposed approach to position adjustment is more refined than mechanically reducing positions according to the level of total volatility, and it better fits the actual risk profile of the trading strategy.

The compounding effect of leverage further complicates position-sizing decisions in perpetual futures. In traditional equity markets, reducing position size usually means reducing the notional principal; in the perpetual futures market, however, a trader faces a choice along two dimensions—reducing notional principal and reducing the leverage ratio. In a high-volatility environment, reducing the leverage ratio is more critical than merely shrinking the notional principal. This is because high leverage significantly shortens the liquidation distance, and nonlinear jumps in volatility can easily trigger the liquidation mechanism. As a concrete example, suppose a trader holds a $10 million notional BTC long at 10x leverage, with a liquidation distance of about 10% of the current price. If volatility jumps from a normal 30% to 100% (which is common in a liquidation cascade), the price could fall by more than 12% within a few hours—enough to trigger liquidation. Even 10x, a relatively conservative level of leverage, still carries substantial forced-liquidation risk in an extreme volatility environment, and the risk at higher leverage is naturally more severe. Rules of thumb from practice indicate that there is a convex relationship between volatility and extreme loss: when volatility doubles, in order to maintain the same probability of ruin, position size should be reduced not merely linearly to 50% of the original but usually to between 50% and 70% of it [43]. This range is an empirical estimate from trading practice; the exact value depends on the fat-tailedness of the return distribution and the leverage actually used, and should not be applied as a hard rule. This convexity comes from the nonlinear nature of risk: the loss under extreme volatility is not merely a linear multiple of the loss under normal volatility.

Figure 24-9 shows the nonlinear relationship between the level of volatility and the Kelly-optimal position fraction. This convexly decreasing curve is one of the most important reference benchmarks for perpetual futures traders in managing position size.

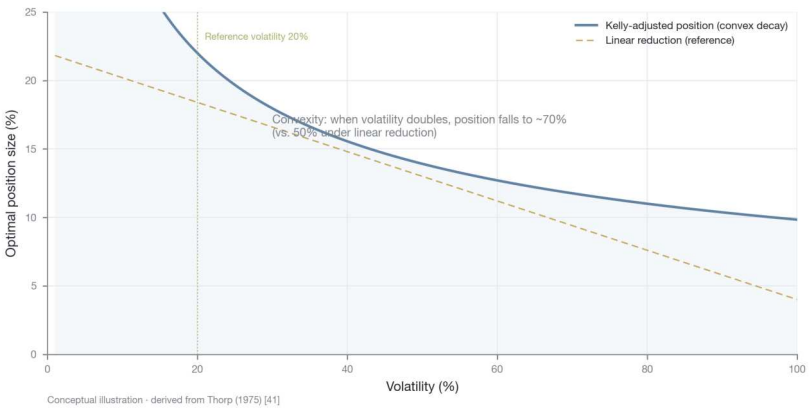


Figure 24-9. The convex function of Kelly position size adjusted for volatility (theoretical-derivation illustration, not measured data: the curve is an illustrative convex decline in which a doubling of volatility reduces position size to roughly 70%, aligned with the fractional-Kelly rule of thumb of 50% to 70% in the main text [43]; the Kelly criterion follows the theoretical derivation of Thorp 1975 [41])

The convexity in the figure has direct practical implications. In the low-volatility interval (such as 20% to 40% annualized), the curve's downward slope is steepest and position size is most sensitive to changes in volatility; when volatility enters the high interval (such as above 80% annualized), the slope flattens markedly, and although position size continues to fall, the marginal reduction per unit change in volatility shrinks. In the extreme volatility environment peculiar to crypto markets (where annualized volatility can exceed 200% during a liquidation cascade), this curve approaches zero, conveying a clear message: from the standpoint of long-run capital growth, maintaining any significant directional position amid extreme volatility is irrational.

24.5.2 Volatility timing

Volatility-timing strategies rest on the premise that market volatility exhibits identifiable regime shifts. As in the state-machine model constructed in Chapter 23, the market switches among three states—calm, stressed, and crisis. Directional traders can dynamically adjust their risk exposure by identifying these state transitions, thereby reducing drawdowns while improving risk-adjusted returns [37]. The key advantage of the state-machine model is that it does not try to predict the absolute level of volatility but instead identifies qualitative changes in market structure. Such identification is often more reliable than a precise volatility forecast, because transitions in market structure usually produce observable signals.

When the market moves from the calm state into the stressed state, the optimal response is to reduce directional risk exposure and raise the cash allocation. The stressed state is usually accompanied by thinning order book depth and extreme funding rates, which not only increases the slippage cost of trading but also raises the probability of mechanical volatility. At this stage, many traders' intuitive reaction is to double down, hoping to capture larger gains amid high volatility. This approach, however, often leads to severe losses, because the stressed state itself signals the fragility of market structure, in which any small shock can trigger a chain reaction. If the market deteriorates further in the stressed state and approaches the crisis state, the trader should reduce risk further or establish explicit hedging positions. The crisis state is characterized by a sharp rise in cross-asset correlation, and traditional diversification often fails at such times because all assets fall in unison. In this situation, holding positions in multiple different crypto assets provides no effective diversification protection.

The transition period from the crisis state back toward the stressed state is often the best window for contrarian traders and mean-reversion strategies to step in. In extreme events such as liquidation cascades, volatility usually overreacts and the price is pushed away from its fundamental equilibrium. As liquidity begins to recover and market risk aversion recedes, this mean-reversion tendency of volatility offers liquidity providers and reversal traders trading opportunities with highly asymmetric returns [44]. In the "Black Thursday" of March 2020 and the June 2022 Luna collapse, traders who established long positions at what proved, with hindsight, to be the optimal entry point earned short-term returns exceeding 200% over the subsequent days to weeks (BTC, for instance, rebounded from a low of about $3,800 on March 12, 2020, and some altcoins staged oversold rebounds after the June 2022 Luna collapse). This figure, however,

suffers from severe survivorship bias: in live trading, participants cannot know whether the current price is the bottom. In the March 2020 event, many traders who tried to buy the dip at the $5,000 to $6,000 level were liquidated by the subsequent decline, and only the few who had reserved an ample margin buffer ultimately profited. In practice, volatility-timing strategies face a core pitfall: state transitions in volatility are not knowable instantaneously and can only ever be confirmed statistically after the fact. In live trading, state identification entails an unavoidable delay, and a strategy's design must be able to tolerate the friction cost of this delay; otherwise, frequent misjudgments and repeated repositioning will erode all of the strategy's excess return.

24.5.3 Indirect volatility trading in perpetual futures

In traditional financial markets, the core instrument for volatility trading is the option; traders can build pure volatility exposure by buying or selling a straddle or by trading volatility-index futures (such as VIX futures) directly [45]. An option's sensitivity to volatility—its vega exposure—enables traders to isolate volatility risk precisely, independent of the underlying asset's directional movement. In the crypto options market, which lacks depth, perpetual futures traders must construct strategies that are long or short volatility through indirect means. The core logic of these indirect strategies is to exploit the microstructure features peculiar to perpetual futures to capture changes in volatility.

The calendar-spread volatility strategy is a method of trading volatility indirectly by using funding-rate forecasts. When future market volatility is expected to rise sharply, the volatility of the funding rate usually rises with it. A trader can build long-short funding-rate positions across different perpetual futures (such as the same asset on different exchanges, or highly correlated assets within the same ecosystem). The return on this strategy does not depend on the absolute price direction of the underlying asset but on the realized volatility of the funding-rate differential. If rate volatility widens as expected, the trader can capture the return from basis widening by dynamically adjusting the weights of the long and short legs. The return of this strategy can be decomposed, in simplified form, as $R_{\text{spread}} = (f_A - f_B) \cdot \Delta t + \Delta\text{basis}_{A,B}$, where $f_A - f_B$ is the difference in funding rates between the two contracts, Δt is the number of funding settlements over the holding period, and $\Delta\text{basis}_{A,B}$ is the change in the basis between the two contracts. The strategy is profitable when the cumulative income from the rate differential exceeds the loss from the change in basis plus transaction costs. The equation above is an illustrative decomposition of the return structure; its parameters depend on the specific contract pair and holding period and must be calibrated empirically before use in live trading. The strength of this strategy lies in its low correlation and high Sharpe ratio, but its weakness is that it requires continuous dynamic management and a deep understanding of the rate structure. In addition, the strategy carries a key margin-fragmentation risk: the trader must lock up margin separately on the two exchanges, and in extreme conditions an unrealized loss on one side may trigger insufficient margin and cause a one-sided forced liquidation, suddenly exposing a supposedly direction-neutral strategy to one-sided directional risk. The mitigation used in practice is to reserve an excess margin buffer on each exchange, but this markedly reduces capital efficiency.

The intraday mean-reversion strategy is, in essence, an indirect strategy that is short liquidity volatility. During periods of high liquidity volatility—especially after a liquidation cascade—the price is often pushed to extreme levels by a liquidity vacuum and then mean-reverts as market makers step back in. By trading against the extreme deviation, the trader is in effect betting on the fading of liquidity volatility and the price's reversion to a short-term mean [46]. Implementing this strategy requires high-frequency order book monitoring and millisecond-level execution speed, so it is usually executed by an algorithmic trading system rather than a human trader. The strategy, however, faces a systematic execution risk: during a liquidation cascade, an exchange's API latency often spikes from its normal level to several seconds, and order requests are rejected or time out in large numbers. The strategy faces a degradation of execution capacity at precisely the moment it most needs to execute, a constraint that significantly limits the practical feasibility of mean-reversion strategies in extreme conditions. In addition, the cross-exchange volatility-spread strategy exploits the realized-volatility differences that may exist for the same asset across different exchanges. When such a difference becomes abnormal because of localized liquidity dry-up or a liquidation event at a specific exchange, a trader can build a cross-exchange hedge whose return comes from the eventual convergence of the volatility difference between the two exchanges.

These indirect volatility-trading strategies face a common core limitation. Unlike options, which can isolate volatility exposure precisely (achieving neutrality to the underlying's price direction—delta-neutral—while retaining only vega and the rate of change of delta, namely gamma exposure), indirect strategies in perpetual futures always contain non-volatility risk components. For example, the calendar-spread rate strategy bears exchange counterparty risk and the cost of margin-use efficiency; the intraday mean-reversion strategy bears the risk of directional trend-continuation in extreme cases. These non-volatility risks make volatility trading in perpetual futures more complex and require more refined dynamic hedging mechanisms.

24.5.4 Portfolio-level tail-risk management

In multi-asset portfolio management, VaR is the standard tool for measuring market risk. Standard VaR models, however, usually assume that asset returns follow a normal distribution or rely on a simple historical-simulation method. In the perpetual futures market, liquidation cascades and leverage effects cause return distributions to exhibit extreme fat tails, so that a VaR based on the normality assumption severely underestimates tail risk [47]. To adapt to this market feature, a portfolio manager must correct the standard VaR, at a minimum using a Student's t-distribution or a Cornish-Fisher expansion (which corrects for skewness and kurtosis) to capture the effect of skewness and kurtosis on the risk measure. The Student's t-distribution has fatter tails than the normal distribution and can more accurately reflect the frequency of extreme events in crypto markets. The effectiveness of these corrections, however, depends heavily on the accuracy of parameter estimation. The degrees-of-freedom parameter of the Student's t-distribution may vary from 3 (extreme fat tails) to above 20 (near-normal) across different market states, and rolling-window estimation and full-sample estimation may give markedly different results. If the degrees-of-freedom parameter is systematically overestimated during calm periods (that is, tail thickness is underestimated), the cor-

rected VaR may, when a crisis arrives, give an even more dangerous false sense of security than the uncorrected version. Parameter estimation should therefore itself use a state-dependent approach, automatically adopting more conservative parameter settings when the state machine identifies a "stressed" or "crisis" state.

The historical-simulation method also requires improvement for use in perpetual futures. The traditional historical-simulation method treats all historical data within the look-back window equally, which is unreasonable in a market with clear regime shifts. A portfolio manager must distinguish normal-period history from crisis-period history. Combined with the state-machine model of Chapter 23, the risk-control system can dynamically adjust the sampling weights of historical data according to the current market state: giving recent data higher weight in the calm state, while in the stressed or crisis state forcibly introducing extreme data from historical crisis periods to calibrate the risk model. This state-dependent VaR computation can avoid being misled by a false low-risk estimate during calm periods.

Conditional value at risk (CVaR) is more important than VaR in perpetual futures risk management. VaR answers only the minimum loss at a given confidence level, whereas CVaR answers what the expected loss actually is in the tail region beyond VaR [48]. Because liquidation cascades in perpetual futures often cause precipitous drops beyond conventional expectations, CVaR is more sensitive to such extreme losses. In managing correlation risk across a multi-asset portfolio, a correlation matrix estimated in normal times often fails in a crisis. In extreme volatility events, correlations among crypto assets rise sharply toward one, and this everything-falls-together scenario must be incorporated into tail-risk management through stress testing. As a concrete example, in the 2022 FTX collapse almost all crypto assets fell by more than 20% within a single day, and this complete rise in correlation caused many portfolios that relied on diversification to suffer severe losses.

24.5.5 Designing stress-test scenarios

Stress testing is a key tool for compensating for the limitations of VaR and CVaR in extreme, nonlinear scenarios. An effective stress test should not merely assume mechanically that asset prices fall by 20% or 30% but should design internally consistent scenarios based on microstructure and volatility-driving mechanisms. Building on the four-source framework of Chapter 22, a stress test of a perpetual futures portfolio can be designed around five core scenarios.

The information-shock scenario simulates sudden events such as a major regulatory ban or a hack of a core protocol. This scenario is characterized by an instantaneous repricing with a clear direction, often accompanied by an isolated plunge in a specific asset. The liquidity-evaporation scenario simulates the situation in which leading market makers withdraw simultaneously for external reasons. In this scenario, even without major negative news, a sharp drop in order book depth causes ordinary order flow to trigger enormous price swings. The liquidation-cascade scenario targets mechanical volatility directly, simulating a price fall into the liquidation cluster of current open positions. This scenario requires the risk-control system to input not only the price shock but also the accompanying liquidity dry-up and extreme imbalance between buying and selling

forces.

The contagion scenario simulates the transmission of a shock from the macro financial market or a specific large entity—for example, a huge single-day net redemption from a Bitcoin spot ETF, or a liquidity crisis in traditional financial markets that pulls capital out of crypto. The full-crisis scenario is the simultaneous activation of the four sources above and represents the market state machine entering its deepest crisis state. Although this scenario has the lowest probability of occurrence, its potential loss to a portfolio is the largest and it must therefore be included in the risk-control framework.

Figure 24-10 arranges these five stress-test scenarios into a matrix by the activation combination of volatility sources and the expected severity, annotating each scenario with its typical price-shock magnitude, degree of liquidity deterioration, and corresponding risk-control response.

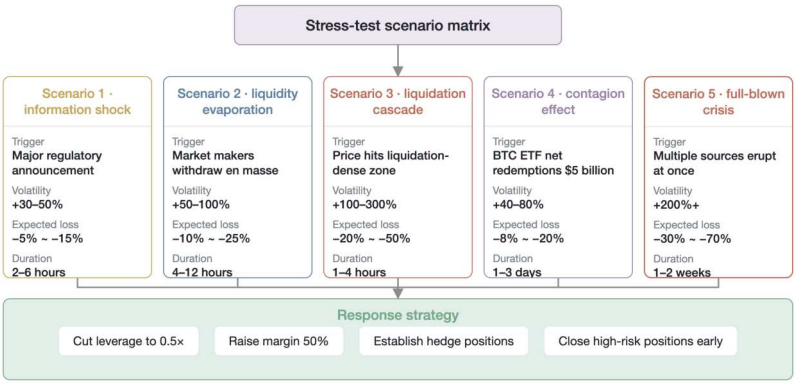


Figure 24-10. A four-source stress-test scenario matrix and response strategies (simulated-stress illustration, not measured data: the volatility, losses, and durations for each scenario are simulated magnitudes drawn from extreme events in crypto history, and the $5 billion net redemption from a BTC ETF is a hypothetical stress rather than a historical observation; data source: simulation parameters based on historical extreme events in crypto markets [49])

A core insight from the matrix is that the expected losses of the five scenarios do not grow linearly with the number of activated sources. Single-source scenarios usually cause manageable losses, because the other, unaffected market mechanisms can still provide a buffer; but when two or more volatility sources activate simultaneously, losses grow markedly superlinearly, owing to the interactive amplification among the volatility sources, and the expected loss of the full-crisis scenario may reach several to several dozen times that of a single-source scenario. In each scenario, the portfolio manager must measure three core metrics: the portfolio's overall expected loss, the position or asset that contributes most to the loss, and whether the remaining margin after the loss is sufficient—thereby identifying the portfolio's key vulnerabilities. The ultimate purpose of stress testing is not to predict when an extreme event will occur but to guide action: once the loss under a plausible scenario is unacceptable or would trigger cross-margin chain liquidations, the portfolio manager must reduce leverage, optimize holdings, or buy tail-hedging instruments before the scenario occurs, rather than reacting passively to depleted liquidity in a crisis. This forward-looking

risk-control framework is a necessary condition for maintaining long-term operation in the perpetual futures market.

In practice, a portfolio manager should establish a systematic stress-testing process. First, the complete five-scenario stress test should be rerun at the end of each trading day, tracking the change in expected loss under each scenario. Second, when the market state machine transitions from one state to another, an additional stress test should be triggered immediately, because market structure is undergoing a qualitative change at that moment. Third, any newly added position or strategy must undergo an independent stress test before being incorporated into the portfolio, to assess its impact on overall portfolio risk under extreme scenarios. This continuous, multidimensional stress-testing framework can significantly improve a portfolio's resilience in a crisis and provides the portfolio manager with an objective, quantifiable basis for risk decisions. Beyond the forward scenario design above, reverse stress testing offers a complementary analytical perspective: instead of assuming a scenario and then computing the loss, it assumes an intolerable loss level (for example, 50% of total capital or the complete exhaustion of margin) and then works backward to infer what combination of market conditions would cause that loss. This approach helps uncover hidden vulnerabilities that forward scenarios miss. For example, reverse stress testing may reveal that certain seemingly diversified multi-asset portfolios in fact have highly concentrated risk exposure under a specific correlation-jump condition.

24.6 Applying volatility in systemic risk management

Volatility forecasting is a tool not only for micro-level traders and portfolio managers; it is equally a core input to the systemic risk management of cryptocurrency exchanges and decentralized derivatives protocols. In traditional risk-control frameworks, exchanges often rely on static parameters to manage risk, but in the perpetual futures market—which trades around the clock and carries extremely high leverage—static parameters cannot keep pace with rapid transitions in the volatility state. This section explores how to embed volatility-forecasting signals in the underlying architecture of exchanges and protocols, from dynamic margin calibration, to sensitivity adjustment of liquidation parameters, to a multi-tier liquidity early-warning system, ultimately building a volatility-driven real-time monitoring loop.

24.6.1 Dynamic margin calibration

In current perpetual futures market practice, most exchanges and protocols use a fixed maintenance margin rate. Although mainstream exchanges have adopted a tiered margin system based on position size (for example, Binance's maintenance margin rate for BTC perpetual futures rises in tiers from 0.4% to 5% with notional position size), the margin rate still does not adjust in real time to the market's volatility state [50]. This static design maximizes capital efficiency during calm periods, but when volatility rises sharply, a fixed margin rate cannot provide enough buffer, leaving the liquidation engine facing a substantial risk of accounts breaching into negative equity.

The core logic of a dynamic margin design is to establish a functional relationship

between the margin rate and forecast volatility. When the system forecasts that future volatility will rise significantly, the exchange proactively raises margin requirements, forcing highly leveraged traders to reduce leverage or add margin before systemic risk accumulates. From the standpoint of statistical calibration, given the volatility forecast for a specific time window, the maintenance margin rate should be set so that, within a specified confidence interval, an extreme price move will not drive position equity below zero [51]. This mechanism ties micro-level margin requirements directly to macro-level volatility expectations.

Such dynamic calibration is not merely a mathematical tool of risk management but a concrete implementation of the "prevention layer" mechanism in protocol governance (as discussed in Section 21.2.2). When the volatility-forecasting model indicates an imminent high-volatility state, dynamically raising margin in effect transmits a risk signal to the market, suppressing speculative position-building by raising the cost of capital. Research shows that treating crypto assets as an independent risk class and calibrating margin according to their distinctive volatility characteristics can significantly reduce the probability of systemic default [52].

The dynamic margin mechanism, however, itself harbors a procyclicality risk that cannot be ignored. Raising margin requirements when volatility rises forces undermargined traders to close positions, and this concentrated closing itself pushes volatility higher, forming a positive feedback loop—precisely the margin-level counterpart of the model-homogeneity problem discussed in Section 24.7.2. Traditional financial regulatory frameworks have proposed systematic responses to this problem: Basel III's countercyclical capital buffer requires banks to accumulate extra capital during booms to meet demand during downturns, and the CPMI-IOSCO central counterparty (CCP) margin standards explicitly require central counterparties to incorporate anti-procyclicality mechanisms into their margin models. Dynamic margin design in crypto markets can draw on this experience—for example, by setting a speed cap on margin increases to avoid violent adjustments over a short interval, or by pre-collecting a "countercyclical margin add-on" during low-volatility periods to reduce the size of margin calls in high-volatility periods. In addition, dynamic margin faces challenges on the compliance front: in traditional finance, margin changes usually require advance notice to traders (CME, for example, gives one trading day's notice), and a crypto exchange that raises margin in real time without prior notice may face legal challenges regarding users' right to be informed, especially under emerging regulatory frameworks such as MiCA.

24.6.2 Adjusting liquidation parameters

The parameter settings of the liquidation engine directly determine the market's survivability in extreme conditions. In crypto derivatives markets, liquidation is usually benchmarked to the mark price rather than the last traded price, to prevent malicious manipulation caused by a short-term absence of liquidity [53]. The mark price is usually computed by combining the spot index price with a funding-rate basis component, and platforms differ in the specific method: Binance uses a moving average of the spot index plus basis, Hyperliquid uses a weighted median price, and dYdX v4 uses a

multi-source median. Regardless of the specific method, the mark price always contains some form of smoothing parameter, and the fixity of this parameter in a high-volatility environment can become a source of systemic risk.

In a high-volatility environment, a longer smoothing window causes the mark price to lag severely behind the spot price. This lag not only provides arbitrageurs with risk-free attack exposure but, more dangerously, triggers an accumulation-release effect (see Section 22.4.3). When the lagging mark price finally catches up with the spot price, liquidation orders that should have triggered dispersed across different price levels are instead triggered simultaneously within an extremely short window, detonating a liquidation cascade [54].

Simply shortening the smoothing window, however, is also risky: it significantly increases the probability of noise liquidations, causing normal small price fluctuations to trigger unnecessary forced liquidations. Here the volatility-forecasting model provides a dynamic benchmark for resolving this trade-off: in the calm state, the system maintains a longer window to filter noise; when the volatility forecast rises sharply, the system automatically shortens the window to improve the mark price's responsiveness. This volatility-sensitive parameter-adjustment mechanism lets the liquidation engine strike a dynamic balance between smoothness and responsiveness. Dynamic parameter adjustment can, however, itself become an attack vector: an attacker who can predict the timing and direction of a parameter switch can profit from the change in liquidation rules at the moment it occurs. Parameter adjustment should therefore use gradual, continuous change rather than discrete jumps, so as to reduce the exploitable time window.

24.6.3 Liquidity early-warning systems

A more advanced form of embedding volatility signals in systemic risk management is to build a multi-tier liquidity early-warning system. Based on the five-stage liquidity model proposed in Chapter 20, an exchange can design an automated response mechanism spanning from a micro-trigger to a liquidation cascade. When a crisis erupts, human decision-makers often cannot keep pace with the execution speed of algorithmic trading, so the warning system's action responses must be highly automated.

Figure 24-11 uses a flowchart to show the three-tier response architecture of this liquidity early-warning system—tier-one micro-trigger detection, tier-two liquidity-ebb response, and tier-three liquidation-cascade emergency intervention—with each tier corresponding to progressively stricter trigger conditions and increasingly forceful risk-control measures.

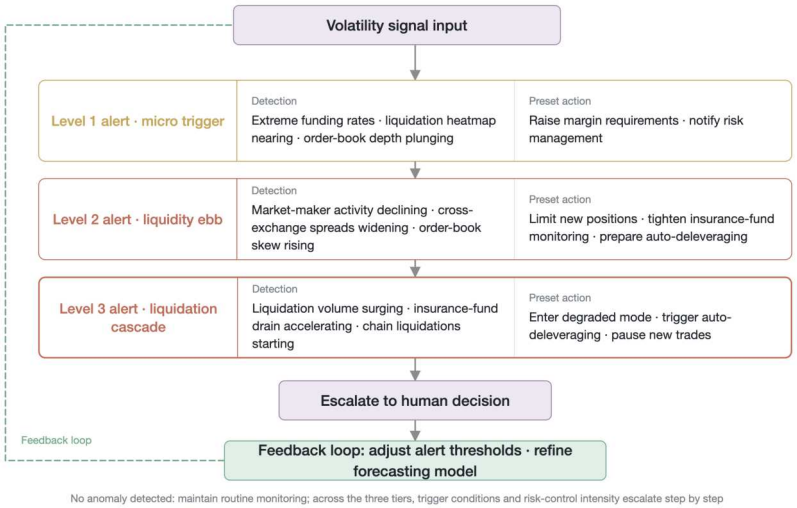


Figure 24-11. A flowchart of a volatility-signal-based liquidity early-warning and automated response system (structural illustration, not measured data; data source: an abstract construction of a protocol risk-management framework [55])

The key design feature of the architecture in the figure is that each warning tier pairs a "trigger condition" with a "preset action," and the mapping between the two is fixed in advance and executes automatically, without human intervention. Tier one is micro-trigger detection: triggered when a signal of surging volatility resonates with the proximity of a liquidation heatmap, its preset actions include raising the initial margin for new positions and sending a liquidity-demand signal to market makers. Research shows that such early turning points usually appear days or hours before violent market swings, providing an ample time window for proactive intervention [55].

If order book depth drops sharply and market makers' quote frequency declines markedly, the system escalates to tier two (liquidity ebb), taking more forceful measures such as restricting new positions in high-risk contracts and raising the maintenance margin rate. When liquidation volume grows exponentially and insurance-fund depletion breaches a safety threshold, the system enters tier three (liquidation cascade), and the protocol must automatically initiate a degradation mode—including suspending trading or activating auto-deleveraging (ADL)—to sever the positive feedback loop between falling prices and forced liquidations [56].

Implementing the warning and degradation mechanisms above faces fundamentally different governance challenges at centralized exchanges and decentralized protocols. A centralized exchange can have its risk-control team make an immediate decision to suspend trading—fast in response but opaque in process; a decentralized protocol's parameter adjustments usually require a governance process such as a multisig or a DAO vote, which simply cannot respond within the few minutes of a crisis. Concentrating the suspension authority in a single admin key, however, introduces a centralized single point of failure and a trust assumption. Both Hyperliquid's JELLY incident of 2025 and the Mango Markets incident of 2023 exposed this governance dilemma: the former was ultimately resolved through an emergency validator vote, while the latter was handled only retroactively through a governance process after the attack had

occurred. Decentralized protocols must find a balance between response speed and decision legitimacy—for example, by using pre-authorized automated risk parameters (which trigger automatically under preset conditions, without a real-time vote) to ease this tension.

In addition, the calibration of warning thresholds is itself an ongoing practical challenge. If thresholds are set too sensitively, they generate too many tier-one warnings, and the system and its users gradually grow immune to the warning signals (the "crying wolf" effect); if set too insensitively, warnings may arrive only after the crisis has already erupted. In practice, the thresholds themselves must also be adjusted dynamically as the market's baseline volatility level and structure change, creating a second-order "calibration of the warning system" problem.

24.6.4 The volatility dashboard

To put the risk-control logic above into practice, exchanges and large protocols need a real-time monitoring framework that integrates all the underlying signals. This system, called the "volatility dashboard," not only aggregates forecast data from different sources but also serves as the hub connecting signals to automated decisions.

Figure 24-12 shows the conceptual design of this volatility dashboard, arranging five core monitoring panels in logical order from micro to macro, with each panel corresponding to an independent risk dimension while achieving cross-panel correlation through an underlying data pipeline.

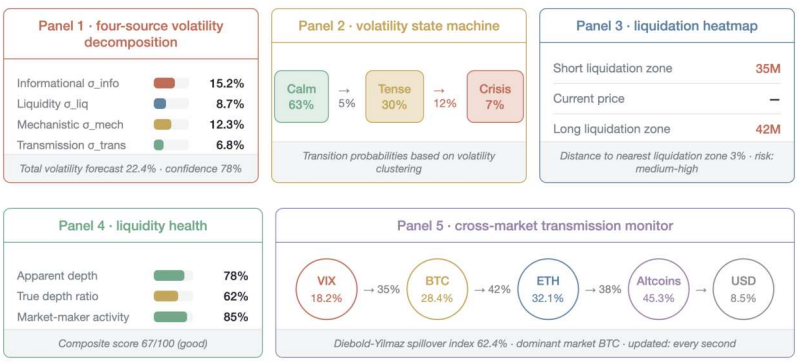


Figure 24-12. The conceptual design of a volatility dashboard and its real-time monitoring panels (conceptual-design illustration: the panel values are representative examples, not a measured snapshot; the variance sum of the four source components is approximately 22.4%, consistent with the conventions of the main text; data source: the author's design based on financial risk-monitoring architecture)

The dashboard's design follows the principle of "single view, multidimensional drill-down": on one interface, risk-control personnel can simultaneously observe the real-time decomposition of four-source volatility, the transition probabilities of the state machine, the spatial distribution of the liquidation heatmap, liquidity health, and cross-market transmission strength. The interconnection among these five panels is the core value of

the system. For example, when panel one shows a sharp rise in estimated mechanical volatility, panel three simultaneously shows a narrowing distance to a liquidation cluster, and panel four confirms that liquidity depth is falling, the resonance of the three automatically triggers an escalated response from the warning system. This monitoring framework closes the full loop—from signal collection, to warning triggering, to execution of preset actions, to feedback adjustment—and is where systemic risk management in the perpetual futures market finally takes concrete form.

24.7 The limits of volatility forecasting

Understanding the limits of forecasting capability is an indispensable part of volatility modeling. Volatility forecasting is an effective tool, but over-reliance on models leads to an accumulation of model risk. This section systematically reviews the inherent limitations of volatility forecasting to help the reader develop a healthy skepticism toward models and avoid treating forecast outputs as certainties. Effective risk control is not about predicting everything but about modeling the predictable part and defending against the unpredictable part.

24.7.1 The decay of predictability

The capability of volatility forecasting does not extend linearly over time but has a clear physical boundary. Within this boundary, market microstructure and historical data provide rich information; beyond it, a forecasting model's effectiveness rapidly decays to no better than a random walk. The speed and degree of this decay are the core metric for evaluating the practical value of any volatility model.

Short-term (one to four hours) realized volatility is highly predictable. This predictive power comes mainly from the clustering effect of volatility—periods of high volatility tend to be followed by periods of high volatility. As earlier chapters discussed, the state-machine model provides powerful medium-term conditional forecasting capability by identifying the market's calm, stressed, or crisis state. In addition, intraday and intra-week periodic patterns, as well as the periodic pulse of funding settlement, all provide highly deterministic inputs for short-term forecasting. When the price approaches a liquidation cluster, the conditional probability of a rise in volatility can be computed precisely from the liquidation heatmap, which is the most valuable forecast of tail risk.

When the forecast horizon exceeds one to two days, however, a half-life effect on information begins to appear. The classic study by Engle and Patton (2001) [57] shows that volatility is highly persistent, with markedly different characterizations of persistence across model specifications—their estimate of the half-life of market volatility is about 73 days. This divergence itself reflects a deeper problem: our understanding of volatility dynamics is far from mature. In cryptocurrency markets, this decay is even more severe. The empirical study of cryptocurrency volatility forecasting by Ftiti et al. (2021) [58] shows that the predictive power of every class of volatility model decays rapidly as the forecast horizon lengthens, with forecast accuracy falling sharply beyond the short-to-medium term. This finding has major implications for risk control in the perpetual futures market: position-sizing decisions that rely on multi-day volatility forecasts lack

sufficient statistical support, and their forecast errors exceed the precision acceptable for risk management.

Figure 24-13 depicts the complete trajectory along which the forecasting power (measured by R^2 or the correlation coefficient) of different types of volatility-forecasting models decays as the forecast horizon lengthens. This decay curve is the most intuitive tool for understanding the practical boundary of volatility forecasting.

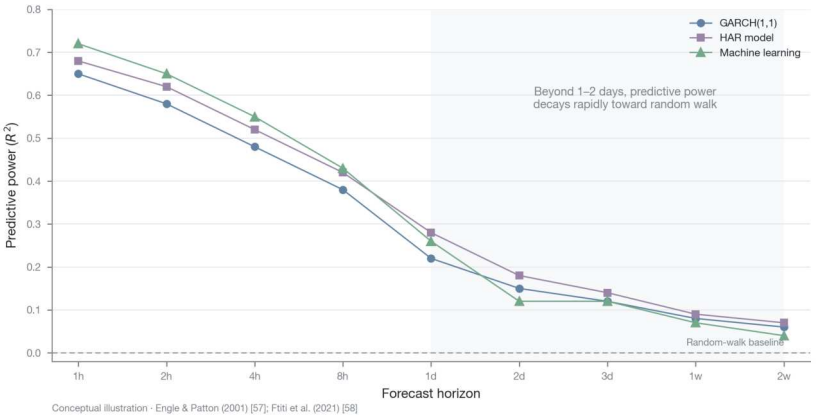


Figure 24-13. The decay curve of volatility-forecasting power over the forecast horizon (conceptual depiction, not measured data: the R^2 magnitudes and the shape of the decay are anchored to the empirical findings of Engle & Patton 2001 [57] and Ftiti et al. 2021 [58], and the claim that predictability approaches a random walk beyond one to two days is a qualitative conclusion; this figure depicts the decay of forecast R^2 over the forecast horizon, a distinct concept from the roughly 73-day half-life of volatility persistence in [57])

The most striking feature in the figure is the nonlinear shape of the decay curve. Within the short-term window (one to four hours), every class of model maintains relatively high forecast accuracy and the curve declines gently; but in the critical region of one to two days, the curve exhibits a sharp "cliff," with forecasting power collapsing dramatically over a very short interval. In cryptocurrency markets, this cliff is steeper than in traditional financial markets, reflecting their faster structural change and shorter information half-life. Beyond the cliff the curve flattens, the differences in forecasting power among the classes of model disappear, and all approach a random walk. The implication for risk-control practice is clear: volatility forecasts beyond two days are statistically unreliable, and position decisions based on such long-term forecasts lack adequate support. Especially difficult to predict is volatility driven by sudden events—true "news" such as a regulatory raid or the bankruptcy of a large exchange—whose timing is intrinsically impossible to infer from any known signal; and market-institutional changes such as the entry and exit of new market makers, the launch of new types of derivatives, or shifts in regulatory policy fundamentally alter the underlying structure of volatility. Even when a liquidity crisis's vulnerability can be identified from the liquidation heatmap, its exact trigger point and the subsequent nonlinear evolution path lie beyond the forecasting capability of statistical models, because extreme events often involve a qualitative rather than a quantitative change in participants' behavior.

24.7.2 Model-homogeneity risk

When most participants in the market use similar volatility-forecasting models, the models themselves become a source of risk. This phenomenon, the accumulation of "model risk," reveals a deep paradox in modern quantitative finance: rational risk management at the individual level may, at the macro level, lead to systemic irrational collapse. The root of this paradox is that participants' behavior is not independent but is tightly coupled through market mechanisms.

Model homogeneity leads to synchronization of collective behavior. When the market experiences a slight anomaly, highly similar forecasting models issue "high-volatility warning" signals to their respective users at the same time. Participants receiving these signals make rational individual responses: market makers widen the bid-ask spread to protect inventory, directional traders reduce positions to control risk, and exchanges' risk-control systems automatically raise margin requirements. These behaviors are robust risk-control measures under normal conditions, but when they occur synchronously, driven by homogeneous signals, they trigger a systemic liquidity dry-up.

Figure 24-14 uses a cyclical flowchart to depict how model homogeneity, through the positive feedback loop of "model warning → synchronized withdrawal → liquidity dry-up → volatility spike → renewed model warning," amplifies a tiny initial disturbance into a systemic liquidity crisis.

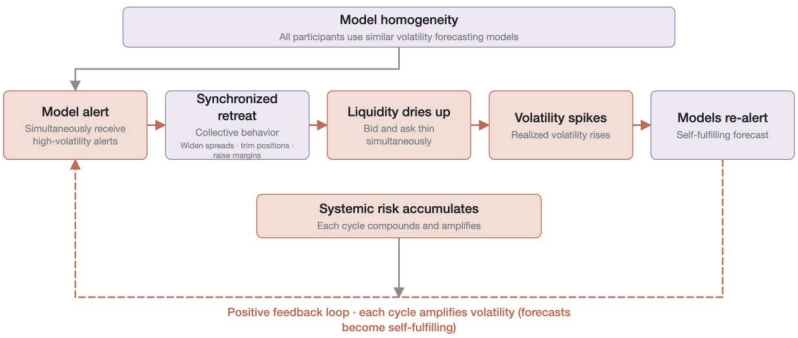


Figure 24-14. The accumulation of systemic risk triggered by the homogeneity of volatility-forecasting models (structural illustration, not measured data: a closed positive-feedback loop; the empirical anchor for the October 2025 cascade of roughly $19 billion in notional liquidations is given in the main text [60]; data source: compiled by the author)

The key node in the positive feedback loop shown in the figure is the step "volatility spike → renewed model warning": the rise in volatility induced by synchronized withdrawal is itself a product of deteriorating market microstructure, not a genuine fundamental information shock, but homogeneous forecasting models cannot distinguish the two and simply feed the higher volatility data into the next round of computation, outputting a higher-level warning signal. This synchronized withdrawal directly causes market liquidity to evaporate suddenly. The absence of liquidity means that even normal trading demand generates enormous price slippage, which further pushes up the actually observed volatility. This higher volatility data is then fed back into the forecasting models, triggering a new round of higher-level warnings. This positive feedback

loop turns the models' forecasts into a self-fulfilling process through the synchronized behavior of participants. In traditional financial markets, in the Flash Crash of May 6, 2010, the homogeneous responses of algorithmic trading systems greatly worsened the market collapse, causing the S&P 500 index to fall by more than 5% within about 15 minutes and the Dow Jones Industrial Average to briefly plunge nearly 1,000 points [59].

In cryptocurrency markets, this homogeneity risk manifests even more extremely. Ali (2025) [60] dissects the October 2025 liquidation cascade and shows that similar risk-control parameter settings and automated liquidation programs triggered roughly $19 billion in liquidations (notional liquidation scale, not wiped-out open interest or actual losses) within about 36 hours. This liquidation cascade was characterized by a self-reinforcing spiral once triggered: the initial liquidation triggered a price drop, the drop set off more liquidations, and the additional liquidations further worsened the drop. Throughout this process, all exchanges and risk-control systems using the same liquidation trigger conditions executed the same operations at the same time, so that liquidity vanished entirely and price gaps became the norm.

The fundamental cause of model homogeneity is the concentration of "best practice" in the market. When a particular volatility-forecasting method is recognized by academia or industry as the "standard," all rational participants have an incentive to adopt it. This produces a "competitive convergence" in which most participants ultimately use the same or highly similar models. This convergence is itself harmless—even efficient—until the market faces a situation outside the models' assumptions. At that point, the homogeneous models fail simultaneously, producing systemic risk.

In crypto markets, model homogeneity has an additional, behavioral source that goes beyond the rational-choice framework: the concentration of information sources. When large numbers of traders follow the same social-media volatility-warning accounts, the same liquidation-data dashboards, or the same on-chain analytics platforms, the synchronization of their behavior comes not only from the similarity of the models themselves but also from the resonance of information inputs and narrative frames. This information-concentration-driven behavioral synchronization also exists in traditional financial markets (such as reliance on the Bloomberg terminal and on the same set of analyst reports), but in crypto markets it is further amplified by the instantaneous transmission effect of social media. Looking ahead, as AI-agent trading spreads, if these agents are based on similar model architectures and training data, model homogeneity may reach an unprecedented magnitude, creating an "algorithmic monoculture" risk in which all agents produce almost identical outputs when facing the same inputs, and the system's diversity defense is completely dismantled.

The paths to mitigating model-homogeneity risk can be developed at three levels. At the individual level, participants can deliberately introduce a heterogeneous mix of models—for example, combining a statistical time-series GARCH model with a machine-learning-based nonparametric model, or incorporating contrarian indicators into the model inputs (such as a contrarian volatility signal: when most models forecast falling volatility, deliberately retaining a certain share of high-volatility estimates as a hedge). At the market level, large market makers and institutions can design asyn-

chronous risk-control triggers, avoiding a situation in which all participants execute the same operation at exactly the same threshold. At the protocol level, decentralized derivatives protocols can consider introducing asynchronous liquidation triggers—for example, setting slightly different liquidation trigger conditions or time delays for positions of different sizes—so as to structurally break the positive feedback loop of synchronized liquidation. Although these measures cannot completely eliminate model-homogeneity risk, they can significantly reduce the probability of a systemic liquidity dry-up caused by synchronized behavior.

24.7.3 Overfitting and out-of-sample failure

A model that shines in backtesting but fails in live trading is a characteristic hazard of volatility forecasting: strong historical fit often gives way to mediocre—or even loss-making—out-of-sample performance. The core cause is overfitting: the model learns not only the true regularities in the data but also memorizes historical noise.

Overfitting is especially severe in cryptocurrency markets. The history of the perpetual futures market is relatively short, the data sample is limited, and samples of extreme events are especially rare. When researchers use high-dimensional machine-learning models with hundreds of features, the algorithm inevitably overfits these rare extreme events. Moreover, multiple-testing bias further worsens the problem. Researchers usually test hundreds or thousands of parameter combinations and select the best-performing one, ignoring the risk of statistical invalidity introduced by this selection process itself. López de Prado (2018) [61] points out that without appropriate correction for multiple-hypothesis testing, backtest results often severely overstate true predictive power.

A deeper cause lies in the continual evolution of market structure. The entry of new participant types (such as institutional investors or AI agents), the launch of new types of derivatives, and changes in regulatory policy all continually alter the data-generating process, causing models trained on past data to fail in the future. As discussed in Section 24.2.4, the "half-life" of an effective volatility-forecasting model in the perpetual futures market may be only 3 to 6 months; even a model that performs perfectly on the past two years of data cannot be guaranteed to remain effective over the next two months.

Practical recommendations for coping with overfitting include rigorous out-of-sample validation and time-series cross-validation. Keeping the model simple is another key principle: following Occam's razor, it is better to use a simple model that is slightly less accurate but robust than to rely on a complex black box that is perfect in backtesting but fragile in live trading. In addition, one should not over-rely on any single model; a combination of models with different underlying assumptions usually provides more robust forecasts. Finally, backtest results should be systematically discounted, assuming that live performance will be 20% to 50% lower than the backtest result—a conservative practice, but one that avoids major losses caused by overconfidence.

24.7.4 Black swan events

Extreme events, or "black swans," are by their nature events beyond the historical distribution and beyond the cognitive range of existing models. Any volatility-forecasting model calibrated entirely on historical data is, by definition, unable to predict the occurrence of such events. Attempting to use a statistical model to predict the timing and specific form of such events is methodologically infeasible, and such attempts produce a false sense of security that reduces the risk-management system's actual defense against extreme events.

Volatility-forecasting models fail in specific ways during extreme events. GARCH-family models assume that volatility evolves continuously and cannot capture the price jumps triggered by liquidation cascades (as discussed in Section 24.2.2). Liao (2013) [62] points out that if jumps are not modeled explicitly, the value of traditional realized-volatility models in risk management is greatly diminished. In the perpetual futures market, price jumps caused by liquidation cascades are a common phenomenon rather than a rare event, which casts serious doubt on the applicability of the standard GARCH model in this market.

Likewise, VaR models based on historical simulation severely underestimate risk during extreme events, because the current tail event may be more extreme than anything recorded in the historical sample. Daníelsson et al. (2013) [63] give a rigorous characterization of the subadditivity condition for VaR under multivariate fat-tailed distributions, showing that the reliability of tail-risk measures depends heavily on the assumptions about the distribution's tails: once the actual tail is fatter than the model assumes, standard methods easily distort the characterization of extreme losses. This means that even a risk model that has performed well over the past decade cannot be guaranteed to correctly estimate risk under extreme market stress.

More deceptive still is the "calm trap." On the eve of many black swan events, the market often maintains a low volatility level, and all forecasting models show risk at a historical low. This false sense of security prompts participants to increase leverage, thereby accumulating structural fragility for the subsequent market collapse. This mechanism has a deep structural correspondence with Minsky's financial instability hypothesis: low volatility $\rightarrow$ low realized volatility $\rightarrow$ low margin consumption $\rightarrow$ traders increase leverage $\rightarrow$ systemic fragility accumulates $\rightarrow$ eventual collapse, which is precisely the crypto-market mapping of Minsky's "stability breeds instability." The quantitative study of systemic risk in cryptocurrency markets by Franco and Laurini (2025) [64] shows that systemic risk contagion often occurs within hours, and conventional daily risk metrics simply cannot provide a warning. In the FTX collapse of November 2022, BTC fell from about \$21,000 to about \$15,500 (a decline of more than 25%) over the three days from November 6 to November 9, and yet in the days before the event erupted the market still maintained a low volatility level, with no forecasting model having given an adequate warning.

Faced with this intrinsic unpredictability, the design of a risk-management system should acknowledge the limits of models' capability. Effective risk control is not about trying to predict all extreme events but about building systemic resilience to them. Constructing this resilience rests on two layers: the first is the statistical-model layer,

used to optimize decisions under normal market states; the second is the rule-constraint layer, a set of hard risk-control rules that do not depend on model output, ensuring that the system can still maintain basic operation when the model fails.

Through rigorous position-size control, participants can ensure that the maximum loss on any single trade does not exceed a fixed proportion of total capital, for example 2%. This proportion must be adjusted for the leverage actually used in a high-leverage environment: at 10x leverage, a 2% capital loss corresponds to a roughly 0.2% price move, still a reasonable stop-loss range; but at 50x leverage, the same capital loss corresponds to only a 0.04% price move, which can be triggered within seconds even under normal market conditions, rendering the stop-loss rule useless. Through firm stop-loss discipline, a trader can close a position immediately once the loss reaches a preset level, avoiding a further widening of the loss. Through an ample liquidity reserve, exchanges and risk-control systems can maintain operational capacity during extreme events rather than being forced to liquidate at extreme prices. None of these measures aims to predict a black swan; they aim to ensure that, when a black swan inevitably appears, losses are bearable rather than unbearable.

Volatility can be studied and partly predicted, but its inherent randomness cannot be completely eliminated. Understanding the sources, dynamics, and predictability boundaries of volatility can improve the quality of risk-management decisions, but it cannot eliminate all uncertainty. A clear-eyed recognition of the limits of models' capability, and a tail-risk budgeting mechanism based on that recognition, are the prerequisites for volatility forecasting to deliver value in practice.

24.8 Chapter summary

In the perpetual futures ecosystem, which lacks a deep options market, participants face a core challenge: how to sense and forecast volatility risk without the benchmark of implied volatility. This chapter has built a complete closed-loop framework running from knowledge of volatility, to forecasting, to decision-making, turning the theoretical insights of the previous two chapters into operational tools. The starting point of this framework is to identify the set of alternative volatility signals specific to perpetual futures, including the degree of extremity of the funding rate, the spatial distribution of open interest in the liquidation heatmap, leading changes in order book depth, and cross-exchange spread anomalies [65]. Although these signals are limited in any single dimension, a reasonable combination of them can effectively capture volatility risk from different sources.

Based on this signal set, the chapter proposed a source-decomposed volatility forecasting method. Traditional unified modeling tries to capture total volatility with a single econometric model, but often performs poorly in the face of the extreme jumps of liquidation cascades and the periodic pulses of funding settlement [66]. The source-decomposed framework, by contrast, models informational, liquidity, mechanical, and transmission volatility separately and then aggregates them, fully exploiting the differences among the sources in time scale and predictability. Its core advantage is interpretability: when the forecast shows a rise in volatility, the risk manager can accu-

rately judge the driving source (informational, liquidity, mechanical, or transmission) and respond in a targeted way.

The ultimate purpose of moving from signal to forecast is to guide decisions. The chapter embedded the forecast outputs in the micro- and macro-level decision processes of different market participants, forming a three-layer decision architecture. At the micro layer, market makers dynamically adjust their quoting strategy based on real-time identification of the volatility state; at the meso layer, portfolio managers use volatility-adjusted position sizing and the CVaR model to manage directional exposure; at the macro layer, exchanges and protocol designers guard against systemic risk through volatility-sensitive dynamic margin mechanisms and liquidity early-warning systems [60]. The core innovation of this closed-loop framework is that it provides not only a tool for measuring risk but also a set of conditional action rules based on the volatility state.

The framework established in this chapter has three key improvements over volatility analysis in traditional finance. First, it explicitly acknowledges the distinctive phenomenon of mechanical volatility in the perpetual futures market—volatility created out of nothing by liquidation cascades and the funding-settlement mechanism, rather than arising from information arrival or liquidity changes. Second, it exploits the special data advantages of perpetual futures, including real-time liquidation heatmaps, funding-settlement schedules, and order book microstructure—data that are hard to obtain in traditional futures markets. Third, it combines volatility forecasting with the specific incentive structures of market participants rather than remaining at the level of abstract risk measurement.

Different types of market participant draw different tools from this framework. Market makers rely on volatility-state identification to achieve dynamic spread adjustment and strategy-mode switching; their core competency is the real-time discrimination between liquidation-driven mechanical volatility and information-driven structural volatility, enabling a correct stay-or-withdraw decision during a sharp price drop. Directional traders and portfolio managers use source-decomposed forecasts for volatility-adjusted position sizing, proactively reducing the leverage multiple rather than merely shrinking the notional position when volatility is high, and identifying a portfolio's vulnerabilities in advance through stress-test scenario design based on the four-source theory [67]. Protocol designers and exchanges embed volatility forecasts in dynamic margin calibration and tiered liquidity early-warning systems, triggering a degradation mode before systemic risk accumulates. The output of the same volatility-forecasting model must be translated into three different decision signals, which requires the risk-management system to be highly flexible and modular in design.

Having built a complete framework from theory to practice, one must be clear about the epistemological boundary of this system: volatility forecasting is essentially a tool for reducing uncertainty, not a method for eliminating it. Intraday liquidity patterns, the periodic pulse of funding settlement, and the inertia effect of volatility clustering belong to the highly predictable category and usually account for 30% to 50% of total volatility. Sudden information shocks, extreme events, and the nonlinear evolution path after a liquidation cascade begins are intrinsically difficult to predict precisely

[67]. A robust risk-management system must be built on the principle of modeling the predictable part and defending against the unpredictable part—even if the model forecasts that volatility will remain at 20%, the risk manager must still reserve an adequate capital buffer for the possibility that volatility suddenly jumps to 100%. In the perpetual futures market, the best risk managers are not those with the most complex forecasting models but those who maintain the clearest recognition of their own models' limitations and reserve an ample tail-risk budget accordingly.

As the final chapter of Part Eight, this chapter has not only completed a full account of volatility from its sources and dynamics to its forecasting and decision-making but has also laid the micro-level foundation for subsequent parts on market-quality assessment and the economics of regulation. Having understood the endogenous interaction between liquidity and volatility, the following analysis turns to a more macro perspective, exploring how these micro-level mechanisms shape the efficiency of price discovery across the whole market, and under what conditions external regulatory intervention must be introduced to correct market failures. The volatility-analysis framework established in this chapter will be used in subsequent parts to evaluate multiple dimensions of market quality, including the accuracy of price discovery, the reasonableness of transaction costs, and the controllability of systemic risk.

References

[1] Coinglass. (2024). *Cryptocurrency derivatives market liquidation data*. https://www.coinglass.com/LiquidationData

[2] Ho, T., & Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, *9*(1), 47–73. https://doi.org/10.1016/0304-405X(81)90020-9

[3] Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367–381.

[4] Mathias, S. E., & Sahoo, S. (2025). Dynamic connectedness and systemic risk in global futures: Evidence from cryptocurrency, financial, and commodity markets. *Eurasian Economic Review*, *16*(1), 311–350. https://doi.org/10.1007/s40822-025-00353-8

[5] Mandelbrot, B. B. (1963). The variation of certain speculative prices. *The Journal of Business*, *36*(4), 394–419. https://doi.org/10.1086/294632

[6] Cont, R. (2007). Volatility clustering in financial markets: Empirical facts and agent-based models. In G. Teyssière & A. P. Kirman (Eds.), *Long Memory in Economics* (pp. 289–309). Springer.

[7] Mayhew, S. (1995). Implied volatility. *Financial Analysts Journal*, *51*(4), 8–20.

[8] Mixon, S. (2009). Option markets and implied volatility: Past versus present. *Journal of Financial Economics*, *94*(2), 171–191. https://doi.org/10.1016/j.jfineco.2008.09.010

[9] Ackerer, D., Hugonnier, J., & Jermann, U. J. (2025). Perpetual futures pricing. *Mathematical Finance*. Advance online publication. https://doi.org/10.1111/mafi.700

[10] Zhivkov, P. (2026). The two-tiered structure of cryptocurrency funding rate markets. *Mathematics*, *14*(2), 346. https://doi.org/10.3390/math14020346

[11] Gornall, W., Rinaldi, M., & Xiao, Y. (2024). *Perpetual futures and basis risk: Evidence from cryptocurrency* (Working Paper). SSRN.

[12] McGroarty, F., ap Gwilym, O., & Thomas, S. (2006). Microstructure effects, bid–ask spreads and volatility in the spot foreign exchange market pre and post-EMU. *Global Finance Journal*, *17*(1), 23–49. https://doi.org/10.1016/j.gfj.2006.06.004

[13] Degryse, H., De Winne, R., Gresse, C., & Payne, R. (2020). *Cross-venue liquidity provision: High frequency trading and ghost liquidity* (ESMA Working Paper No. 4). European Securities and Markets Authority.

[14] Makarov, I., & Schoar, A. (2019). Price discovery in cryptocurrency markets. *AEA Papers and Proceedings*, *109*, 97–99. https://doi.org/10.1257/pandp.20191020

[15] Chi, Y., Chu, Q., & Hao, W. (2024). Return and volatility forecasting using on-chain flows in cryptocurrency markets. *arXiv preprint arXiv:2411.06327*. https://doi.org/10.48550/arXiv.2411.06327

[16] Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*(4), 885–905.

[17] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*(2), 174–196. https://doi.org/10.1093/jjfinec/nbp001

[18] Barndorff-Nielsen, O. E., & Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, *2*(1), 1–37.

[19] Kim, J.-M., Jun, C., & Lee, J. (2021). Forecasting the volatility of the cryptocurrency market by GARCH and stochastic volatility. *Mathematics*, *9*(14), 1614.

[20] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007. https://doi.org/10.2307/1912773

[21] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

[22] Duan, J.-C., Ritchken, P. H., & Sun, Z. (2006). *Jump starting GARCH: Pricing and hedging options with jumps in returns and volatilities* (Working Paper No. 06-19). Federal Reserve Bank of Cleveland.

[23] Chen, K. S., & Yang, J. J. (2024). Price dynamics and volatility jumps in bitcoin options. *Financial Innovation*, *10*(1), 132. https://doi.org/10.1186/s40854-024-00653-z

[24] Zhu, H., Bai, L., He, L., & Liu, Z. (2023). Forecasting realized volatility with machine learning: Panel data perspective. *Journal of Empirical Finance*, *73*, 251–271.

[25] Corsi, F., Audrino, F., & Renò, R. (2012). HAR modeling for realized volatility

forecasting. In L. Bauwens, C. Hafner, & S. Laurent (Eds.), *Handbook of Volatility Models and Their Applications* (pp. 363–382). Wiley.

[26] Kakade, K., Mishra, A. K., Ghate, K., & Gupta, S. (2022). Forecasting commodity market returns volatility: A hybrid ensemble learning GARCH-LSTM based approach. *Intelligent Systems in Accounting, Finance and Management*, *29*(2), 103–117.

[27] Zhuo, Y., & Morimoto, T. (2024). A hybrid model for forecasting realized volatility based on heterogeneous autoregressive model and support vector regression. *Risks*, *12*(1), 12.

[28] Mishra, A. K., Renganathan, J., & Gupta, A. (2024). Volatility forecasting and assessing risk of financial markets using multi-transformer neural network based architecture. *Engineering Applications of Artificial Intelligence*, *133*, 108223. https://doi.org/10.1016/j.engappai.2024.108223

[29] Blanc, S. M., & Setzer, T. (2016). When to choose the simple average in forecast combination. *Journal of Business Research*, *69*(10), 3951–3962. https://doi.org/10.1016/j.jbusres.2016.05.013

[30] Ben Omrane, W., Dabbou, H., Saadi, S., Savaser, T., & Sebai, S. (2025). Exploring volatility reactions in cryptocurrency markets using intraday macroeconomic news analysis. *International Review of Economics & Finance*, *103*, 104509. https://doi.org/10.1016/j.iref.2025.104509

[31] Easley, D., López de Prado, M. M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *The Review of Financial Studies*, *25*(5), 1457–1493. https://doi.org/10.1093/rfs/hhs053

[32] Hansen, P. R., Kim, C., & Kimbrough, W. (2024). Periodicity in cryptocurrency volatility and liquidity. *Journal of Financial Econometrics*, *22*(1), 224–251. https://doi.org/10.1093/jjfinec/nbac034

[33] Brauneis, A., & Sahiner, M. (2026). Crypto volatility forecasting: Mounting a HAR, sentiment, and machine learning horserace. *Asia-Pacific Financial Markets*, *33*(1), 379–411. https://doi.org/10.1007/s10690-024-09510-6

[34] Shahrour, M. H., Lemand, R., & Mourey, M. (2025). Cross-market volatility dynamics in crypto and traditional financial instruments: Quantifying the spillover effect. *The Journal of Risk Finance*, *26*(1), 1–21. https://doi.org/10.1108/jrf-07-2024-0185

[35] Wyart, M., Bouchaud, J.-P., Kockelkoren, J., Potters, M., & Vettorazzo, M. (2008). Relation between bid–ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, *8*(1), 41–57.

[36] Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217–224. https://doi.org/10.1080/14697680701381228

[37] Kalimipalli, M., & Susmel, R. (2004). Regime-switching stochastic volatility and short-term interest rates. *Journal of Empirical Finance*, *11*(3), 309–329. https://doi.org/10.1016/j.jempfin.2003.10.001

[38] Auh, J. K., & Cho, W. (2023). Liquidation cascade and anticipatory trading: Evidence from the structured equity product market. *The Review of Asset Pricing*

Studies, *13*(1), 53–98. https://doi.org/10.1093/rapstu/raac010

[39] Park, A. (2023). The conceptual flaws of decentralized automated market making. *Management Science*, *69*(11), 6731–6751. https://doi.org/10.1287/mnsc.2021.02802

[40] Hyperliquid. (2025). *Protocol vaults: Hyperliquidity Provider (HLP)*. https://hyperliquid.gitbook.io/hyperliquid-docs/hypercore/vaults/protocol-vaults

[41] Thorp, E. O. (1975). Portfolio choice and the Kelly criterion. In *Stochastic Optimization Models in Finance* (pp. 599–619). Academic Press.

[42] Hsieh, C. H., Barmish, B. R., & Gubner, J. A. (2018). At what frequency should the Kelly bettor bet? *2018 Annual American Control Conference (ACC)*, 5485–5490. https://doi.org/10.23919/ACC.2018.8431224

[43] Laureti, P., Medo, M., & Zhang, Y.-C. (2010). Analysis of Kelly-optimal portfolios. *Quantitative Finance*, *10*(7), 689–697. https://doi.org/10.1080/14697680902991619

[44] Ahmed, R. R., Vveinhardt, J., Štreimikienė, D., Ghauri, S. P., & Ashraf, M. (2018). Stock returns, volatility and mean reversion in emerging and developed financial markets. *Technological and Economic Development of Economy*, *24*(3), 1149–1177.

[45] Abken, P. A., & Nandi, S. (1996). Options and volatility. *Economic Review (Federal Reserve Bank of Atlanta)*, *81*(3–6), 21–35.

[46] Renault, E., & Touzi, N. (1996). Option hedging and implied volatilities in a stochastic volatility model. *Mathematical Finance*, *6*(3), 279–302. https://doi.org/10.1111/j.1467-9965.1996.tb00117.x

[47] Huisman, R., Koedijk, K. G., & Pownall, R. A. (1998). VaR-x: Fat tails in financial risk management. *Journal of Risk*, *1*(1), 47–61.

[48] Sarykalin, S., Serraino, G., & Uryasev, S. (2008). Value-at-risk vs. conditional value-at-risk in risk management and optimization. In *State-of-the-Art Decision-Making Tools in the Information-Intensive Age* (pp. 270–294). INFORMS.

[49] Likitratcharoen, D., Chudasring, P., Pinmanee, C., & Wiwattanalamphong, K. (2023). The efficiency of value-at-risk models during extreme market stress in cryptocurrencies. *Sustainability*, *15*(5), 4395.

[50] TD Securities. (2025). *Perpetual futures: The missing link in tokenized equities*. https://www.tdsecurities.com/ca/en/tokenized-equities-missing-link-perps

[51] Gate.io. (2026). *Cryptocurrency derivatives risk weights are independent, higher volatility requires higher margin*. https://www.gate.com/news/detail/18744636

[52] Amirdjanova, A., Lynch, D., & Zheng, A. (2026). *Initial margin for crypto currencies risks in uncleared markets* (FEDS Working Paper No. 2026-009). Federal Reserve Board. https://www.federalreserve.gov/econres/feds/files/2026009pap.pdf

[53] Allen, F., & Carletti, E. (2008). Mark-to-market accounting and liquidity pricing. *Journal of Accounting and Economics*, *45*(2–3), 358–378. https://doi.org/10.1016/j.jacceco.2007.02.005

[54] Soska, K., Dong, J. D., Khodaverdian, A., Zetlin-Jones, A., Routledge, B., & Christin, N. (2021). Towards understanding cryptocurrency derivatives: A case study of BitMEX. *Proceedings of the Web Conference 2021*, 3450–3459.

[55] Song, S., & Li, H. (2025). Can topological transitions in cryptocurrency systems serve as early warning signals for extreme fluctuations in traditional markets? *Physica A: Statistical Mechanics and its Applications*, *657*, 130194. https://doi.org/10.1016/j.physa.2024.130194

[56] Qin, K., Ernstberger, J., Zhou, L., Jovanovic, P., & Gervais, A. (2023). Mitigating decentralized finance liquidations with reversible call options. In *Financial Cryptography and Data Security* (pp. 344–362). Springer.

[57] Engle, R. F., & Patton, A. J. (2001). What good is a volatility model? *Quantitative Finance*, *1*(2), 237–245.

[58] Ftiti, Z., Louhichi, W., & Ben Ameur, H. (2021). Cryptocurrency volatility forecasting: What can we learn from the first wave of the COVID-19 outbreak? *Annals of Operations Research*, *330*(1), 665–690. https://doi.org/10.1007/s10479-021-04116-x

[59] Kirilenko, A. A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The Flash Crash: High-frequency trading in an electronic market. *The Journal of Finance*, *72*(3), 967–998. https://doi.org/10.1111/jofi.12498

[60] Ali, Z. (2025). Anatomy of the Oct 10–11, 2025 crypto liquidation cascade: Macroeconomic triggers, market microstructure, and systemic risk lessons. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5611392 https://doi.org/10.2139/ssrn.5611392

[61] López de Prado, M. (2018). *Advances in financial machine learning*. Wiley.

[62] Liao, Y. (2013). The benefit of modeling jumps in realized volatility for risk prediction: Evidence from Chinese mainland stocks. *Pacific-Basin Finance Journal*, *23*, 25–48. https://doi.org/10.1016/j.pacfin.2013.01.002

[63] Daníelsson, J., Jorgensen, B. N., Samorodnitsky, G., Sarma, M., & de Vries, C. G. (2013). Fat tails, VaR and subadditivity. *Journal of Econometrics*, *172*(2), 283–291. https://doi.org/10.1016/j.jeconom.2012.08.011

[64] Franco, J. P. M., & Laurini, M. P. (2025). Quantifying systemic risk in cryptocurrency markets: A high-frequency approach. *International Review of Economics & Finance*, *102*, 104214.

[65] Naimy, V., Haddad, O., Fernández-Avilés, G., & El Khoury, R. (2021). The predictive capacity of GARCH-type models in measuring the volatility of crypto and world currencies. *PLoS ONE*, *16*(1), e0245904. https://doi.org/10.1371/journal.pone.0245904

[66] Silva, L., & Maciel, L. (2025). Cryptocurrency price returns volatility modeling and forecasting with GARCH models. *RAUSP Management Journal*, *60*(1), 225–239. https://doi.org/10.1108/rausp-04-2023-0056

[67] Joshi, P. (2025). Regime-specific interdependencies in cryptocurrency markets: A high-frequency GMM-VAR approach. *Data Science in Finance and Economics*, *5*(3), 419–439.

Part IX
Engineering: From Diagnosis to Construction

Core question: Once we understand how this market works, how do we evaluate it, regulate it, improve it, and ultimately build a better version?

A Five-Dimensional Evaluation of Perpetual Futures Market Quality

By early 2025, cryptocurrency derivatives trading had been fundamentally reshaped: perpetual futures accounted for 93% of all cryptocurrency derivatives volume [1]. That dominance is the endpoint of a long climb. Since BitMEX launched the first Bitcoin perpetual futures in 2016, an instrument with no expiry date, anchored to the spot price through the funding rate, has grown from a fringe experiment into core infrastructure for pricing crypto assets worldwide. As volumes have grown exponentially and the participant base has become more complex, however, volume alone can no longer answer a more basic question: does the market's overall quality match its scale? High leverage and around-the-clock trading may impose hidden structural costs that headline volume conceals. This tension frames the central question of the chapter: how can we construct a unified, multidimensional framework for evaluating market quality—one that diagnoses the true health of the perpetual futures market and confronts the systematic trade-offs among its quality dimensions?

Evaluating a trading market's quality means asking how well it satisfies the most fundamental demands of its participants. From the perspective of a perpetual futures trader, these expectations can be summarized as standards across five dimensions: low transaction costs (encompassing not only explicit fees but also the implicit taxes formed by the bid-ask spread, market impact, and funding rate), accurate price discovery (market prices that reflect supply-and-demand information rapidly and without bias), strong resilience (the maintenance of liquidity supply amid extreme volatility), fair trading rules (ordinary participants are not placed at a systematic disadvantage for lacking market-maker privileges), and open market accessibility (the ability to participate in global pricing regardless of geographic location, capital scale, or technical sophistication). Together, these five dimensions constitute the core framework of market quality. The microstructure analyses in Chapters 13 through 24—covering price discovery, arbitrage mechanisms, liquidity supply, and the sources of volatility—all ultimately serve a single macro-level question: how does the perpetual futures market perform overall across these five dimensions? The task of this chapter is to integrate the fragmented conclusions of the preceding chapters into a structured diagnostic report on market quality, a framework that measures both the absolute level of each dimension and the intrinsic conflicts among them. The high-leverage mechanisms introduced to lower

transaction costs, for example, often come at the expense of the market's resilience under extreme conditions, while the strict admission screening implemented in pursuit of absolute fairness inevitably undermines the market's accessibility.

This chapter first measures the five dimensions of transaction costs, price discovery, resilience, fairness, and accessibility (Sections 25.2 through 25.6), and then incorporates three market states—normal, stressed, and crisis—to analyze the time-varying characteristics and state dependence of market quality. On this basis, the chapter constructs an evaluation model that unifies the five dimensions, delivers a structured quality assessment of perpetual futures platforms or specific assets, and explains the systematic compromises that different mechanism designs make among multidimensional quality objectives, thereby laying a diagnostic foundation for the subsequent discussion of regulatory intervention and institutional redesign.

25.1 Theoretical foundations of market quality

25.1.1 The economic meaning of market quality

At its most fundamental level, evaluating market quality examines how well a market's institutions allocate resources efficiently and distribute welfare among participants. A high-quality market should maximize social welfare, an objective that, at the microstructure level, manifests as the unity of investor welfare, market stability, and allocative efficiency. Investor welfare depends on the level of transaction costs, the accuracy of pricing, and the fairness of participation rules. Market stability requires that a trading venue be able to withstand external shocks and the risks of internal positive-feedback loops. Allocative efficiency, in turn, depends on how accurately price signals reflect fundamental information, thereby directing capital toward its highest-value uses.

From the standpoint of social welfare, improvements in market quality translate directly into gains in economic efficiency: lower transaction costs improve the efficiency of capital allocation, greater price-discovery efficiency improves the quality of investment decisions, and stronger market resilience reduces systemic risk. These seemingly microscopic changes in metrics manifest at the macro level as improved allocative efficiency, reduced financial risk, and greater overall economic welfare. Market quality is therefore not only a concern for traders but also a strategic issue that regulators and policymakers must address.

The economic study of market quality can be traced back to the perfect-market benchmark proposed by Stigler (1961) [2]. In this theoretical abstraction, a perfect market is characterized by instantaneous trading, zero transaction costs, and homogeneous participants, with information transmitted through the market without friction. Real financial markets cannot attain this ideal state; the existence of information costs necessarily produces price dispersion and friction. The degree of departure from the perfect-market benchmark constitutes the central yardstick for measuring the loss of market quality. Within this framework, every evolution of market institutions—from manual matching to electronic order books, from centralized clearing to on-chain automated market makers (AMMs)—can be viewed as an attempt to narrow the gap between real and perfect markets.

Stigler's framework provides a systematic method of analysis. By taking the perfect market as a benchmark, we can decompose the defects of any real market into specific sources of cost: transaction costs, information asymmetry, participant heterogeneity, and so on. This decomposition allows us to diagnose exactly where a market's quality problems lie. If a market's bid-ask spread is wide, for example, we can ask further: is it because market makers face high adverse-selection risk (a problem of information asymmetry), because liquidity providers face high funding costs (insufficient market depth), or because the exchange's technical infrastructure is outdated (slow trading)? Different root causes call for different solutions.

Evaluating market quality requires a conceptual expansion from a single efficiency metric to a multidimensional evaluation system. Informational efficiency is an important dimension of market quality, measuring the speed and accuracy with which prices absorb new information. Yet informational efficiency is not the whole of market quality. A market that is highly informationally efficient but has extremely high transaction costs, or that is prone to systemic collapse under extreme conditions, cannot be called high-quality. Evaluating market quality requires jointly considering the efficiency of price discovery, the cost of liquidity provision, and the market's performance under stress. The need for this multidimensional perspective is especially pronounced in perpetual futures markets, because their distinctive mechanism designs—such as high leverage, automatic liquidation, and funding rates—do not exist in traditional financial markets.

The core insight of a multidimensional evaluation framework is to acknowledge the many-sided nature of market quality. A market may excel on some dimensions while exhibiting clear deficiencies on others. Some highly specialized perpetual futures platforms, for example, may be very competitive on transaction costs yet see their liquidity dry up under extreme conditions; conversely, some platforms with abundant liquidity may carry high effective transaction costs because of complex fee structures. This heterogeneity across dimensions means that a simple *ranking* approach is inappropriate. What we need is a framework capable of structured diagnosis across multiple dimensions, not a system that compresses all dimensions into a single score.

25.1.2 Traditional theories of market quality

Modern market microstructure theory provides rigorous analytical tools for quantifying market quality. O'Hara (1995) shifted the research focus from macroeconomic equilibrium to pricing mechanisms at the transaction level; centering on the bid-ask spread and information asymmetry, she revealed how market makers provide liquidity under the constraints of adverse-selection risk and inventory costs [3]. From this perspective, the bid-ask spread is not merely a trader's execution cost but also a structural indicator of the market's degree of information asymmetry and the efficiency of its liquidity supply. The more severe the information asymmetry that market makers face, the higher the spread compensation they demand, which lowers overall market quality.

The insight that information asymmetry is a fundamental source of market friction points directly to one key avenue for improving market quality: reducing the information asymmetry among participants.

As electronic trading became widespread, Harris (2003) proposed a more systematic

multidimensional evaluation framework that decomposes market quality into four core dimensions: liquidity, transaction costs, price discovery, and market integrity [4]. This framework clarified the interactions among these dimensions: liquidity determines the implicit component of transaction costs (market impact), while the efficiency of price discovery is constrained by market integrity and transparency. Building on these theoretical foundations, regulators such as the U.S. Securities and Exchange Commission and the European Securities and Markets Authority (ESMA) established a system of regulatory metrics centered on best execution, order-handling rules, and market-data transparency, translating the academic concept of market quality into operational compliance standards.

Harris's four-dimensional framework provides a direct theoretical anchor for the five-dimensional model of this chapter. While retaining its core dimensions, the chapter makes three key adjustments: it splits *market integrity* into two independent dimensions, *fairness* and *accessibility*, to reflect the sharply distinct constraints on institutional fairness and participation barriers in crypto markets; and it adds a *resilience* dimension to capture the market's capacity to absorb shocks, which the traditional framework did not model explicitly.

The academic literature has developed a variety of single-dimension tools for measuring market quality. The illiquidity ratio proposed by Amihud (2002) captures a market's sensitivity to price impact through the ratio of price change to trading volume and has become one of the most widely used liquidity proxies in empirical research [5]. Such single-dimension metrics, however, cannot capture the multilayered quality characteristics specific to perpetual futures markets. A market with a very low Amihud ratio (that is, superficially abundant liquidity), for example, may prove extremely fragile during a liquidation cascade. The incremental contribution of the five-dimensional model lies in providing a structured diagnostic framework that can identify efficiency advantages and institutional deficiencies at the same time.

The important contribution of the Harris framework lies in making explicit the multidimensionality of market quality and the mutual constraints among its dimensions. A market with abundant liquidity can execute large trades with smaller market impact, thereby lowering transaction costs; but excessive liquidity supply may erode market makers' risk-bearing capacity, so that liquidity dries up rapidly in periods of stress. The efficiency of price discovery depends on participants' ability to acquire and process information, but this also implies the presence of informed traders, whose participation raises the transaction costs of uninformed traders. This trade-off among dimensions is a central difficulty in market design.

The traditional market-quality framework, however, reveals clear limitations when confronted with perpetual futures markets. These classical theories and regulatory metrics embed several structural assumptions that do not hold in crypto-asset markets. The traditional framework presumes that the market has a central clearing counterparty, is regulated within a single jurisdiction, and trades products with standardized expiry dates and delivery mechanisms. The fragmented liquidity of perpetual futures markets, their margin systems without a central clearing house, and their diverse liquidation mechanisms all fall outside the explanatory range of the traditional framework.

Traditional models also fail to adequately account for the systemic risk introduced by high-leverage automatic liquidation, and they do not consider the distinctive features of around-the-clock trading, the funding-rate mechanism, and decentralized architectures.

The failure of these assumptions means that we cannot simply transplant the quality metrics of traditional markets onto perpetual futures markets. The traditional standard of *best execution*, for example, is typically based on the spread and execution speed; but in perpetual futures markets, funding-rate payments can far exceed spread costs and become the primary source of trading cost. Likewise, *market integrity* in traditional markets usually refers to the prevention of manipulation and fraud; but in perpetual futures markets, because of exchanges' proprietary market making and information asymmetry, *fairness* becomes a far more complex and critical problem. It is precisely this inapplicability of the traditional framework that motivates the new five-dimensional model proposed in this section.

25.1.3 The distinctive challenges of perpetual futures

The mechanism innovations of perpetual futures and the distinctive structure of crypto-asset markets pose five new questions for market-quality evaluation that the traditional framework cannot answer.

The first challenge is how to measure market resilience. Traditional financial markets rely on circuit breakers and central clearing counterparties as buffers against systemic risk, but in perpetual futures markets these external protections are typically absent. The market's capacity to repair itself under extreme conditions, the speed at which order-book depth recovers, and the ability to withstand chained liquidations must be measured explicitly as an independent dimension. In the absence of circuit breakers, resilience becomes the key indicator of whether a market can continue to function normally under stress.

The second challenge concerns how to handle institutional noise. Mechanism designs specific to perpetual futures—such as the periodic settlement of the funding rate and the tiered forced liquidation performed by the automatic liquidation engine—introduce non-informational volatility into the price-discovery process. Ruan and Streltsov (2022) show that the funding-rate cycle of perpetual futures produces a systematic intraday pattern in spot-market quality: market makers face higher adverse-selection risk around funding settlement and therefore widen the bid-ask spread [6]. This quality loss, induced by the contract mechanism itself, has no corresponding measurement dimension in traditional evaluation frameworks.

The third challenge is how to evaluate market fairness. Traditional markets are bound by unified regulatory standards, whereas in crypto-asset markets the agreements between exchanges and market makers often lack transparency, and some platforms even face conflicts of interest from proprietary market making. The asymmetric distribution of information privileges and execution latency means that fairness can no longer be assumed as a default premise but must be assessed explicitly through concrete indicators. In perpetual futures markets, certain market makers may obtain special data or priority liquidation rights from the exchange, and these hidden privileges directly affect market fairness.

The fourth challenge is how to compare centralized exchanges (CEXs) and decentralized exchanges (DEXs) within a unified framework. Barbon and Ranaldo (2024) find that CEXs have a cost advantage in small trades, whereas DEXs offer a more competitive execution environment for large trades [7]. Capponi and Jia (2025) further note that the infrastructure of DEXs confronts liquidity providers with a distinctive collective-action dilemma, giving them little incentive to withdraw from the liquidity pool [8]. These two architectures differ fundamentally in their trust assumptions, settlement efficiency, and modes of organizing liquidity, which requires the evaluation framework to be sufficiently inclusive. CEXs rely on the credit of a central counterparty, whereas DEXs rely on the transparent execution of smart contracts; the quality characteristics of the two are entirely different.

The fifth challenge is how to evaluate accessibility. The global nature of perpetual futures and the *permissionless* property of decentralized protocols transform the participation barrier from an exogenous regulatory setting, as in traditional markets, into an endogenous quality variable that directly determines the diversity of participants and the depth of liquidity pools (elaborated in Section 25.6).

25.1.4 The architecture of the five-dimensional model

To address these challenges, this section proposes the five-dimensional model of perpetual futures market quality (5D-MQ). The model decomposes the market quality of perpetual futures into five interrelated dimensions: transaction costs, price discovery, resilience, fairness, and accessibility. Together, these five dimensions constitute a complete evaluation framework capable of a structured diagnosis of how different platforms and different assets perform across various market states.

The five-dimensional model rests on a core recognition: the quality characteristics of perpetual futures markets differ fundamentally from those of traditional markets and require an evaluation framework designed specifically for them. This framework must be able to accommodate the mechanisms specific to perpetual futures (funding rates, automatic liquidation, and the like) while also comparing markets with different architectures (CEX versus DEX). More importantly, it must be able to identify the multiple dimensions of market quality and make explicit the trade-offs among them, thereby offering market participants and regulators a clear diagnosis and a direction for improvement.

Figure 25-1 presents the hierarchical architecture of the five-dimensional model: market quality sits at the center, from which the five first-level dimensions—transaction costs, price discovery, resilience, fairness, and accessibility—radiate outward, and each in turn resolves into several operational second-level indicators. The structure thus offers macro-level generality while grounding itself in concrete measurement indicators at the micro-operational level.

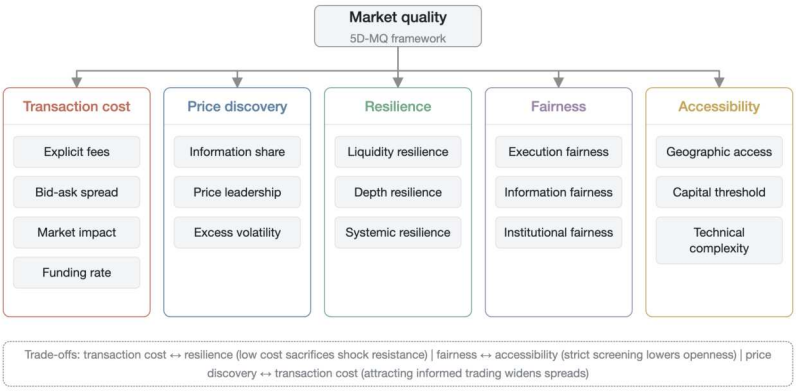


Figure 25-1. Architecture of the five-dimensional model of perpetual futures market quality (5D-MQ) (Data source: author's construction)

The design of the five-dimensional model follows four core principles. The first is mutual exclusivity: each dimension covers an independent facet of market quality. Transaction costs measure the execution friction of entering and exiting the market; price discovery assesses the efficiency with which price signals absorb information; resilience gauges the market's capacity to recover after a shock; fairness examines how equitably the rules treat different participants; and accessibility defines the breadth of the market's openness. This partition avoids overlap and ambiguity among dimensions. Mutual exclusivity matters because it ensures that each dimension has a clear definition and an independent measurement method, avoiding the double-counting or logical confusion that dimensional overlap would cause.

The second is completeness: these five dimensions form a comprehensive mapping of market microstructure and macro-level performance, and omitting any one of them would lead to a biased judgment of market quality. A market with very low transaction costs that is prone to collapse under stress, for example, has a true quality far below what its low-cost metric would imply. Likewise, a market with efficient price discovery that is entirely unfriendly to small traders has a deficiency in the accessibility dimension that offsets its price-discovery advantage.

The third is measurability: each dimension is decomposed into concrete operational indicators. Transaction costs, for example, include not only explicit fees and the bid-ask spread but also market impact and the funding rate, the implicit cost unique to perpetual futures; price discovery is quantified through information share and excess volatility; and resilience is measured through the spread-recovery half-life and the depth-recovery rate. Alexander et al. (2020), analyzing the Ethereum market, demonstrate that perpetual futures dominate price discovery—a leadership metric that is itself an important component of the price-discovery dimension [9].

The fourth is the presence of trade-offs: the five-dimensional model reveals systematic intrinsic conflicts among the dimensions of market quality. Lower transaction costs often come at the expense of market resilience (high leverage, for example, lowers the cost of capital tied up in margin but intensifies liquidation fragility); greater institu-

tional fairness may require stricter identity verification, thereby reducing accessibility; and greater price-discovery efficiency may attract more informed traders, raising market makers' adverse-selection costs and ultimately pushing transaction costs up. The five-dimensional model shows that no market is perfect on all dimensions; the design of market institutions is essentially a matter of weighing trade-offs along the efficient frontier formed by these five dimensions. Recognizing these trade-offs helps make sense of the complexity of market design, because it explains why different markets make different choices and why the *optimal* market design depends on the specific use case and the needs of participants.

The five-dimensional model lays the theoretical foundation for the sections that follow. Sections 25.2 through 25.6 examine, one by one, each dimension's measurement methods, cross-platform comparisons, and evolutionary trends; Section 25.7 analyzes the trade-offs among the dimensions; and Sections 25.8 through 25.10 further explore the state dependence of market quality, cross-platform comparison, and long-term evolution, providing an empirical foundation for the discussion of regulatory economics in Chapter 26 and the thought experiments in Chapter 27.

The five-dimensional model is a diagnostic framework rather than a scoring system (its *non-scoring* stance is detailed in Section 25.1.1). Its diagnostic power lies in its ability to answer concrete questions such as "Why does a given exchange have very low transaction costs yet see its liquidity dry up in periods of stress?" or "Why does a given DEX offer greater fairness but relatively lower accessibility?" Such questions are valuable not only for academic research but also as practical guidance for participants' decisions and regulators' policymaking.

Finally, one intrinsic limitation of the five-dimensional model should be openly acknowledged: the framework embeds an assumption of rational participants—that market participants can correctly assess risk and make decisions in their own interest. In crypto perpetual futures markets, however, the combination of high leverage, around-the-clock trading, and extreme volatility may systematically exploit traders' behavioral biases, including overconfidence, loss aversion, and a propensity to gamble. The five-dimensional model evaluates the operating quality of market institutions, not the behavioral health of participants. This means that a market performing excellently on the five-dimensional metrics may still produce large-scale individual welfare losses because of participants' behavioral irrationality. The absence of this dimension is not a design oversight but reflects a fundamental distinction between assessing institutional quality and assessing participant welfare, the latter of which requires the independent analytical tools of behavioral finance.

25.2 Transaction costs

Transaction costs are the most traditional and most measurable dimension of market quality. In perpetual futures markets, the composition of transaction costs is more complex than in traditional financial markets. Beyond explicit fees and the implicit bid-ask spread, the funding rate constitutes a distinctive holding-period transaction cost. In a one-sided trending market, the cumulative burden of the funding rate can

far exceed the sum of fees and spread, becoming the largest cost a trader faces. This section deconstructs the total transaction cost of perpetual futures across three layers: explicit fees, implicit friction, and the funding-rate burden.

25.2.1 Explicit costs

Fees are the explicit costs a trader must pay to enter and exit the market. Cryptocurrency exchanges generally adopt a two-tier maker-taker fee structure—based on whether an order provides or removes liquidity—combined with a tiered fee schedule keyed to trading volume and asset holdings. The core aim of this design is to incentivize market makers to supply liquidity while charging a premium to traders who actively consume it.

Among the leading centralized exchanges, fee schedules display a high degree of convergence alongside subtle differences. Take Binance's Ethereum perpetual futures: for ordinary users the maker fee is 0.02% and the taker fee is 0.05% [10]. By comparison, OKX charges ordinary users a 0.1% maker fee and a 0.2% taker fee for spot trading, which indicates that explicit transaction costs in the derivatives market are markedly lower than in the spot market [11]. A decentralized exchange such as Hyperliquid establishes a competitive advantage in explicit costs by waiving maker fees entirely and compressing the taker fee to 0.02%. Bybit, a newer centralized derivatives exchange, has a base fee structure close to Binance's but attracts high-frequency traders through a more aggressive rebate program.

There is a marked difference between nominal and effective fee rates. Through tiered fee schedules and market-maker rebate programs, exchanges compress the actual execution rate on the vast majority of trading volume to a small fraction of the nominal rate. Once a trader's monthly volume reaches a certain threshold, the maker fee can fall not merely to zero but into negative territory, becoming a rebate subsidy. When a Binance VIP 3 user reaches $100 million in monthly volume, the maker fee on perpetual futures can fall to 0.012%, roughly three-fifths (about 60%) of the ordinary-user rate [10]. An OKX VIP 9 user can obtain a 0% maker fee, achieving entirely free market making [11]. This structure means that retail traders are, in effect, subsidizing the liquidity-provision costs of high-frequency market makers. Assessing a market's explicit transaction costs through the official fee schedule alone is therefore incomplete; one must compute a weighted-average effective rate that reflects the structural distribution of market participants.

25.2.2 The bid-ask spread and market impact

Explicit fees make up only part of transaction costs; the bid-ask spread and market impact constitute a far larger implicit friction. The bid-ask spread is not a fixed value set by the exchange but compensation that market makers demand for bearing inventory risk, adverse-selection risk, and order-processing costs. The spread directly reflects the market's liquidity depth and degree of information asymmetry.

The quoted spread and the effective spread are two distinct dimensions for measuring this implicit cost. The quoted spread is the difference between the best bid and the best ask on the order book and represents the immediate execution cost of a small market

order. The effective spread accounts for how an order actually fills across multiple price levels and thus more truthfully reflects the real cost borne by large traders [12]. In markets dominated by traditional market makers, the effective spread is typically about half the quoted spread, reflecting limit-order price improvement and participants' adjustments in anticipation of price movements [12]. In the electronic limit order books of crypto perpetual futures, however, the ratio between the effective and quoted spreads varies with the market state, owing to the high proportion of aggressive market orders and market makers' rapid quote withdrawals during volatile periods; under stress, the effective spread may approach or even exceed the quoted spread [7]. In the highly liquid Bitcoin and Ethereum perpetual futures markets, the quoted spread is typically held within an extremely narrow range of 0.5 to 2 basis points [7][13]. For less liquid altcoin contracts, however, the spread can widen to more than 50 basis points, forming a significant implicit barrier to trading.

The market impact generated by large trades is another important component of implicit costs. When the size of a market order exceeds the depth at the best bid and ask, the order eats through successively worse price levels along the order book, so that the actual average execution price departs from the initial quote. This price impact can be decomposed into a permanent information cost that cannot be eliminated and a temporary liquidity cost that can be reduced by improving liquidity. In calm markets, a $1 million Bitcoin market order may generate only about 5 basis points of impact cost [7]; but during periods of extreme volatility or liquidity exhaustion, an order of the same size may trigger more than 100 basis points of impact, giving implicit transaction costs a nonlinear amplification (the specific basis-point figures above are illustrative estimates based on the methodology of [7]). This nonlinearity carries important implications for risk management, because it means that during market stress the real cost a trader faces can be more than 10 times the normal level.

25.2.3 The holding cost of the funding rate

The funding rate is a transaction-cost dimension unique to perpetual futures markets and constitutes the holding cost of a position. In traditional futures markets, the cost of holding a position is chiefly a deterministic interest charge; in perpetual futures, by contrast, longs and shorts must periodically exchange the funding rate in order to keep the contract price anchored to the spot price. This rate is a random variable that fluctuates in real time with market supply and demand, and it can spike to extreme levels under extreme conditions.

In economic terms, the funding rate is the mechanism by which the market prices leverage imbalance. When market sentiment is extremely optimistic and long leverage demand far outstrips short demand, the perpetual futures price develops a positive premium, driving the funding rate positive. Longs must then pay shorts. At the peak of the 2021 bull market, the annualized funding rate on Bitcoin perpetual futures held at a high of 30% to 80% for several weeks at a stretch. This means that even if the underlying asset's price merely churns sideways, long holders suffer significant capital erosion from the heavy funding burden—and the funding rate has been confirmed by empirical research as a risk factor independent of price direction [14].

The way the funding rate is calculated varies across exchanges and continues to evolve. BitMEX first established the model of a fixed base rate plus a premium index, with settlement every 8 hours as the early industry standard. Leading exchanges such as Binance and OKX built on this with more flexible floating-rate models that adjust dynamically to the real-time deviation between the contract price and the spot price. As of early 2025, these platforms had introduced 4-hour or even higher-frequency settlement cycles for some major contracts, while a new generation of platforms such as Hyperliquid was exploring hourly settlement. Higher-frequency settlement reduces the absolute magnitude of any single funding payment but also increases microstructure noise around the settlement points, requiring market makers to rehedge their funding-rate exposure before and after each settlement window.

In bull-market cycles, the funding rate exhibits a marked positive skew, with longs systematically paying this holding tax. Yet this positive skew is not an inherent property of perpetual futures: during the deep bear market of 2022–2023, the funding rate was persistently negative, and shorts instead became the net payers. Over data spanning a full bull-bear cycle, the net positive skew of the funding rate is far smaller than the value observed during a pure bull market, but it remains positive, reflecting the structural feature that long leverage demand dominates over the long run in crypto markets. This is an additional implicit cost of taking a long exposure through perpetual futures rather than spot. For high-frequency traders, the effect of the funding rate is small; but for swing traders or hedging institutions holding for more than a few days, the cumulative funding outlay can far exceed the combined fees and spread of entering and exiting. A trader holding a $1 million Bitcoin long position for 30 days, in an environment where the funding rate is annualized at 30%, would pay about $25,000 in funding costs, equivalent to a 2.5% erosion of capital. This cost level is enough to meaningfully eat into a trading strategy's return. Annualizing the expected funding rate and weighting it by the holding period is therefore a necessary step in accurately assessing the true transaction cost of perpetual futures.

25.2.4 Measuring total transaction cost

A comprehensive assessment of market quality requires integrating explicit fees, implicit friction, and the funding-rate burden into a single total-transaction-cost metric. Total transaction cost equals the sum of fees, half the effective spread, the market-impact cost, and the expected funding-rate holding cost. This composite measure reveals the true cost constraints faced by different trading strategies.

A concrete worked example helps clarify this framework. Suppose a trader opens a $1 million Bitcoin long position on Binance as a taker, holds it for 7 days, and executes both the opening and closing trades with market orders. The total transaction cost should include all the friction of the round trip: the entry taker fee, $0.05\% \times \$1$ million $= \$500$; the entry effective-spread cost, assumed at 5 basis points, or $500; the entry market-impact cost, assumed at 10 basis points in a calm market, or $1,000; the 7-day funding cost, assumed at an annualized 20%, or $20\% \times 7/365 \times \$1$ million $\approx$ $3,800; the exit taker fee, likewise $500; the exit effective-spread cost, $500; and the exit market-impact cost, $1,000. The total cost of a full round trip is therefore about

$7,800, equivalent to a capital cost of 0.78%. This cost level may be acceptable for strategies with short holding periods but constitutes a significant drag on returns for long-term holders. If the trader closes with a limit order rather than a market order, the spread and impact costs on the exit side can be avoided, lowering the total cost to about $6,300 (0.63%)—though at the expense of immediacy in closing the position.

Investors with different trading frequencies and holding periods have sharply different sensitivities to total transaction cost. For high-frequency market makers, fees and the spread are the core consideration, while the funding rate is nearly negligible because holding times are so short. For lower-frequency trend followers, the funding rate is the largest cost item, especially since building a position with the trend often requires paying a steep funding premium. Total transaction cost is also strongly state-dependent. In calm markets, abundant liquidity and low-volatility funding keep total cost at a low level; during market stress, the spread widens sharply, market impact amplifies nonlinearly, and the funding rate produces extreme pulses, so total transaction cost can spike dozens of times over within a short span. More important still, the total-cost formula above treats each cost item as an independent variable to be summed, whereas in real markets these items interact significantly. When the funding rate spikes, market makers widen the bid-ask spread because they bear greater inventory risk and funding-rate exposure; that is, an extreme funding rate indirectly pushes up the other implicit cost items. This interaction means that the true total transaction cost during stress can be higher than the simple sum of the individual costs.

Comparing total transaction cost across assets further confirms the stratification of market quality. Major assets such as Bitcoin and Ethereum enjoy the narrowest spreads, the deepest liquidity, and relatively stable funding rates, and their total transaction cost is the lowest. As market capitalization and liquidity decline, the total transaction cost of altcoin perpetual futures rises along a steep gradient. This cost differential reflects not only how market makers price the risk of different assets differently but also a friction barrier that impedes the efficient allocation of capital across long-tail assets.

25.2.5 Cross-platform comparison and evolution

Applying the total-transaction-cost framework to individual platforms, Figure 25-2 decomposes the total transaction cost of Bitcoin perpetual futures across exchanges under typical market conditions.

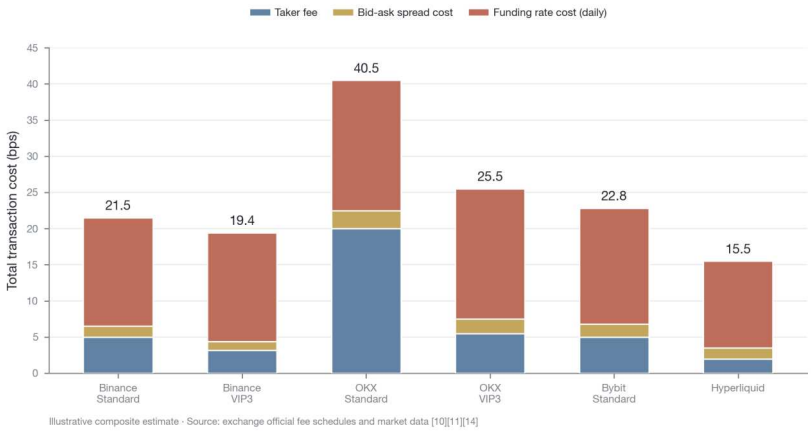


Figure 25-2. Comparison of total transaction cost for Bitcoin perpetual futures across leading exchanges (illustrative composite estimate, not direct measurement; the OKX VIP 3 and Bybit ordinary tiers are representative extensions not discussed in the main text, while the remaining tiers appear in the text; data source: exchanges' official fee schedules and market data [10][11][14])

In cross-platform comparison, ordinary users of centralized exchanges face the highest explicit fee burden, but the exchanges' deep order books effectively lower market-impact costs. The total transaction cost for a Binance ordinary user is about 21.5 basis points, of which the taker fee accounts for 5 basis points, the bid-ask spread cost for 1.5 basis points, and the annualized funding cost for 15 basis points. As the fee tier rises, total transaction cost falls markedly: a Binance VIP 3 user's total cost drops to 19.4 basis points, reflecting the platform's tilt toward high-frequency, large-capital users. Because its spot fees are higher, an OKX ordinary user's total transaction cost reaches 40.5 basis points, roughly twice that of a Binance ordinary user. Through a zero-fee strategy and an innovative liquidity-pool mechanism, the decentralized derivatives platform Hyperliquid compresses total transaction cost to 15.5 basis points, exerting significant competitive pressure on total cost and forcing traditional centralized platforms to reexamine their fee models.

Over the past 10 years, the transaction cost of perpetual futures markets has fallen structurally. From the tens-of-basis-points wide spreads when BitMEX launched the first Bitcoin perpetual futures in 2016, to today's extremely narrow spreads on major assets that approach those of traditional financial markets, this progress is attributable chiefly to three drivers. The first is the full-scale entry of high-frequency market makers and the spread of algorithmic trading, which sharply compressed implicit friction and reduced spreads on major assets from more than 10 basis points to less than 1. The second is fierce competition among exchanges over liquidity and market share, which drove explicit fees steadily downward, from an early 0.1% to today's 0.02%. The third is the maturation of liquidation engines and risk-management mechanisms, which lowered the system's tail-risk premium and markedly reduced funding-rate volatility. Yet the funding rate, as the core mechanism for balancing longs and shorts, still imposes a holding cost that remains a distinctive property setting perpetual futures apart from other financial instruments—and a computational hurdle that traders must clear when assessing the cost of long-term positions. The existence of this cost dimension gives the

total-transaction-cost structure of perpetual futures markets a fundamental difference from that of traditional futures markets and is central to understanding perpetual futures market quality.

25.3 Price discovery

Price-discovery quality measures the efficiency of the market as an information processor—that is, how quickly and accurately prices reflect new information once it arrives. This section treats informational efficiency as a key dimension of market quality, examines how it interacts with the other dimensions, and constructs a multilayered framework for evaluating price-discovery quality. Perpetual futures have become the dominant venue for crypto-asset price discovery, but this leadership is conditional and varies significantly with the type of information shock, the market state, and the asset class.

25.3.1 Measurement methods

When the same asset trades across multiple markets, determining which market leads the price-discovery process is a central question in assessing market quality. The academic literature generally uses two classic econometric methods to quantify the price-discovery contribution of different markets. The first, proposed by Hasbrouck (1995), centers on the concept of the information share. This method assumes an implicit, unobservable common efficient price and defines a given market's information share as its proportional contribution to innovations in that common efficient price [15]. Within this framework, a market that absorbs and reflects new information more quickly has a higher information share, indicating that it dominates price discovery.

The second method derives from Gonzalo and Granger (1995), who proposed a technique for decomposing a price series into permanent and transitory components. This permanent-transitory decomposition is widely used to estimate the price-discovery contributions of multiple cointegrated markets, particularly in analyzing price leadership between futures and spot markets [16]. Unlike the information share, which focuses on the contribution to variance, the Gonzalo-Granger method focuses on measuring the speed at which each market's price adjusts to the common long-term trend. When two markets depart from equilibrium, the market that adjusts more slowly is regarded as the price leader, while the market that adjusts more quickly is the follower.

Applying these traditional methods directly to perpetual futures markets, however, poses particular challenges. The fundamental difference between perpetual futures and traditional futures is their lack of an expiry date, which requires the funding-rate mechanism to anchor them to the spot price [17]. The periodic payment or receipt of the funding rate introduces a distinctive mechanical friction that makes the cointegration relationship between the perpetual futures price and the spot price more complex than in traditional markets. Moreover, the around-the-clock trading and very high microstructure noise of crypto markets require the use of high-frequency data and finer adjustments to model specification when applying these econometric models, so as to accurately separate genuine information shocks from short-term liquidity disturbances.

Notably, the Hasbrouck information share and the Gonzalo-Granger component share can, in practice, produce contradictory rankings of market leadership. This contradiction stems from the two methods' different treatments of microstructure noise: the information share's variance decomposition of price innovations depends on the ordering of the variables, whereas the component share is highly sensitive to the estimation of short-term noise components. Because microstructure noise in perpetual futures markets is far higher than in traditional markets, the choice of method has an especially pronounced effect on the conclusions. The empirical discussion in the remainder of this chapter therefore draws on the results of both methods to strengthen the robustness of its conclusions.

25.3.2 The leadership relationship between perpetual futures and spot

A large body of empirical evidence shows that perpetual futures lead price discovery in the spot market during most periods. Alexander et al. (2020), in their study of the Ethereum market, find that the Ethereum perpetual futures on BitMEX are absolutely dominant across all price-discovery metrics, with an information share far higher than that of the spot exchanges. The study further notes that the introduction of perpetual futures actually improved the pricing efficiency of the spot exchanges [9]. Robertson (2025), analyzing the latest high-frequency data for the Bitcoin market, reaches a similar conclusion, confirming that the futures market holds a systematic lead over the spot market in absorbing information [18]. This leadership stems from the built-in leverage, lower transaction costs, and deeper liquidity that perpetual futures offer—features that make them the preferred venue for informed traders to express their views. The full picture of price discovery, however, should not be confined to a binary comparison between offshore perpetual futures and the spot market. Since the approval of Bitcoin exchange-traded funds (ETFs) in 2024, the open interest of CME Bitcoin futures has grown significantly, and institutional participants increasingly express their pricing views through the CME during U.S. trading hours. In certain periods, the price-discovery contribution of the regulated futures market may exceed that of offshore perpetual futures, forming a tripartite pricing structure of offshore perpetuals, spot, and regulated futures.

Although perpetual futures hold the overall lead, this leadership is markedly conditional. The relative roles of the markets shift depending on the type of information shock. For macroeconomic data releases or market-wide liquidity shocks, perpetual futures typically react fastest, thanks to their high liquidity and leverage. For information specific to a single asset, however—such as a protocol upgrade, a security vulnerability, or a large spot transfer—the spot market sometimes briefly reclaims the lead in price discovery. This division of labor in information processing reflects differences in the composition of participants and their trading motives across markets.

Sharp shifts in the market state likewise reshape the price-discovery hierarchy. During the prolonged 2021 bull market, the leadership of perpetual futures reached its peak: persistently positive funding rates and intense speculative sentiment made the derivatives market the absolute pricing engine. After the series of systemic crises and the deep bear market of 2022, however, the information share of perpetual futures declined some-

what as extensive leverage was liquidated and market-maker capital withdrew. During phases of extreme panic selling, the actual selling pressure of the spot market often becomes the decisive force driving prices, and the leadership of perpetual futures appears more fragile in these stress tests. This state dependence shows that the price leadership of perpetual futures is not a static constant but a variable that evolves dynamically with the market environment.

25.3.3 Cross-asset differences

The price-discovery efficiency of crypto-asset markets varies enormously across assets. Major assets and altcoins exhibit sharply different information-processing capacities. As the assets with the deepest liquidity and the largest market capitalizations, Bitcoin and Ethereum have perpetual futures markets that display high microstructure efficiency. In these two markets, prices integrate new information rapidly, the information share is stable at a high level, and the bid-ask spread is extremely narrow. This efficient price discovery benefits from a mature market-maker ecosystem, broad institutional participation, and highly developed cross-market arbitrage mechanisms.

By contrast, the perpetual futures markets of most altcoins are markedly less efficient. Liquidity in these assets is typically thin, and order-book depth is insufficient to absorb medium-sized trades without generating significant price impact. In such an environment, the price-discovery process is often accompanied by violent volatility and abundant noise. For lack of sufficient arbitrage capital to maintain cross-market price consistency, the prices of altcoin perpetual futures frequently deviate from their spot underlyings for extended periods, resulting in low informational efficiency. This positive-feedback loop between liquidity and price-discovery efficiency allows highly liquid assets to keep improving in efficiency while low-liquidity assets struggle to escape their inefficient state.

The distribution of price-discovery shares across exchanges further reveals the complexity of market structure. The fragmentation of liquidity across platforms disperses pricing power. Leading centralized exchanges, with their vast user bases and deep market-maker resources, typically hold the largest information shares. As decentralized derivatives exchanges have grown, however, this pattern is shifting subtly. Through innovative liquidity-supply mechanisms and permissionless access, decentralized platforms have begun to capture a sizable share of price discovery for particular assets or during particular periods. This multi-centered price-discovery process increases arbitrage opportunities but also introduces cross-platform transmission latency and friction, posing new challenges to the pricing efficiency of the market as a whole.

25.3.4 Excess volatility

In assessing price-discovery quality, relying on the information share and similar metrics alone is often insufficient to capture the full complexity of market microstructure. Excess volatility offers a composite inverse indicator of considerable analytical value. Excess volatility is defined as the difference between a market's actual total volatility and the volatility it ought to exhibit if driven by fundamental information. Operationally, this warranted volatility can be estimated in several ways: one is to use the

Hasbrouck information share to decompose realized volatility into information-driven and noise components; another is to use the degree of departure from a random walk, measured by a variance ratio test over a rolling window, as a proxy [15][18]. High excess volatility means that price movements contain a great deal of non-informational noise, which directly reflects a decline in price-discovery quality. Such noise can mask genuine supply-and-demand signals, raising traders' execution costs and adverse-selection risk.

Several distinctive sources of excess volatility exist in perpetual futures markets. The first is volatility amplification triggered by the liquidation mechanism. When rapid price movements trigger large-scale forced liquidation, the market orders that the system automatically dumps cause a severe chain of price impacts on an already thin order book, producing violent movements that far exceed fundamental value. The second is institutional noise from funding-rate pulses. In a one-sided market, an extreme funding rate forces arbitrageurs to adjust positions frequently, and this non-informational order flow likewise injects additional volatility. In addition, the defensive quote adjustments that market makers make when facing inventory risk increase the random-walk component of prices at the micro level.

Figure 25-3 shows the relationship between excess volatility and the price-discovery efficiency index for different crypto assets across multiple periods.

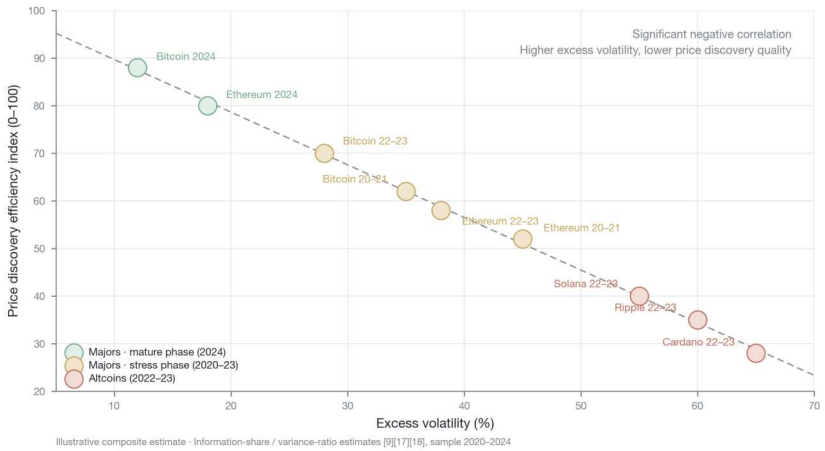


Figure 25-3. The inverse relationship between excess volatility and price-discovery efficiency (illustrative composite estimate, not direct measurement; the two are significantly negatively correlated and are presented qualitatively only, with no precise correlation coefficient reported so as to avoid false precision; based on the information-share and variance-ratio estimates of Alexander et al. [9], He et al. [17], and Robertson [18], with excess volatility defined as the difference between realized volatility and the fundamental volatility weighted by the Hasbrouck information share, sample period 2020–2024)

The scatter plot clearly reveals the significant negative correlation between the two. In more mature market phases, such as 2024, Bitcoin and Ethereum exhibit low excess volatility and high price-discovery efficiency. During the high-stress period of 2022–2023, by contrast, the excess volatility of even major assets rose significantly, accompanied by a decline in efficiency. More strikingly, altcoins such as Solana, Ripple, and Cardano sit systematically in the lower-right region of the chart—the zone of high excess volatility and low price-discovery efficiency. This empirical evidence strongly

supports treating excess volatility as an important dimension for gauging perpetual futures market quality and informational efficiency, providing an intuitive, quantifiable yardstick for comparisons across assets and across cycles.

25.4 Resilience

Resilience is an underrated yet critical dimension of perpetual futures market quality. In day-to-day trading, traders usually attend only to the narrowness of the spread or the depth of the order book, overlooking resilience altogether. Resilience becomes visible only in a crisis, yet the loss from a single failure of resilience can exceed the returns accumulated over years of calm. In a market like crypto—highly leveraged, running around the clock, and lacking a central bank as lender of last resort—understanding and measuring the market's capacity for self-repair is key to assessing systemic risk. This section analyzes resilience across three layers—liquidity resilience, depth resilience, and system resilience—moving from the micro level of spread recovery and order-book reconstruction, to differences in the design of liquidation mechanisms and margin systems, to the resilience function of the funding rate under extreme conditions, and finally to the macro-level systemic risk of cross-exchange contagion.

25.4.1 Definition and measurement framework

In financial microstructure theory, market liquidity is typically decomposed into three interrelated dimensions. Under the classic taxonomy proposed by Kyle, these are tightness, depth, and resilience [19]. Tightness measures the direct cost of trading, usually expressed as the bid-ask spread; depth reflects the market's ability to absorb large orders without causing large price movements; and resilience is defined as the speed at which prices recover from random, uninformative shocks [19]. In perpetual futures markets, resilience is not merely the recovery of price but the self-repair of the entire market structure after extreme stress.

Resilience differs fundamentally from the other dimensions of liquidity. Liquidity is static: it describes how many trading opportunities the market can offer at a given moment. Resilience is dynamic: it concerns the market's ability to replenish liquidity once it has been consumed. A market may exhibit excellent tightness and depth in calm periods yet dry up in an instant when shocked, for lack of resilience. This dynamic property matters especially in crypto markets, because high leverage and programmatic trading tend to amplify local shocks into systemic events.

A comprehensive assessment of perpetual futures market resilience requires a multilayered measurement framework. The first layer is liquidity resilience, concerned mainly with the speed at which the bid-ask spread returns to normal after widening. The second is depth resilience, which measures the time it takes for limit orders to refill the gaps after a large market order has punched through the order book. The third is system resilience, which assesses the ability of an entire exchange to keep functioning normally and to prevent risk from spreading to other markets when confronted with a large-scale liquidation cascade. These three layers are interwoven and together determine a market's survival capacity under extreme conditions.

In analyzing resilience, we must distinguish two conceptually distinct layers: price resilience and system resilience. Price resilience concerns the speed at which an asset's price reverts to fundamentals after a short-term shock, whereas system resilience measures the market's ability to keep its core functions—trade execution, clearing and settlement, and price discovery—uninterrupted when it suffers a severe external shock or internal liquidity exhaustion. For perpetual futures markets, whose underlying assets are themselves extremely volatile, pursuing absolute price stability is neither realistic nor necessary; the core of a resilience assessment should therefore focus on system-level robustness: whether the market can keep operating under stress, not whether the price stays unchanged.

Perpetual futures markets face particular structural challenges in maintaining system resilience. Perpetual futures have no expiry date, which means risk exposure can accumulate indefinitely and the market cannot force a periodic release of accumulated pressure through delivery settlement. These markets generally offer very high leverage multiples, so that small price movements can trigger forced liquidation and, in turn, set off a chain reaction. Empirical observation indicates that when the market's aggregate leverage ratio—defined as the ratio of the notional value of network-wide open interest to total margin deposits—climbs to a high level, systemic fragility rises significantly, and at that point any small negative shock can trigger an uncontainable liquidation cascade [20]. The specific leverage threshold that triggers systemic fragility varies with the measurement approach and sample (illustratively, somewhere around 25%), and there is no single, precise critical value. Actual fragility depends not only on the absolute level of leverage but also on its distribution: high-leverage positions concentrated among a few large holders are far more dangerous than low-leverage positions dispersed across many retail traders, because liquidating the former releases far greater instantaneous selling pressure.

25.4.2 The spread-recovery half-life

The most intuitive expression of liquidity resilience is the spread-recovery half-life. When the market is hit by a large trade or an external news shock, market makers typically withdraw their limit orders quickly to avoid adverse-selection risk, causing the bid-ask spread to widen sharply. The spread-recovery half-life is the time it takes for the spread to fall halfway from its post-shock peak back to its initial normal level [21]. This indicator directly reflects the confidence and speed with which liquidity providers reassess risk and return to the market.

In traditional financial markets, electronic limit order books usually recover quite quickly. Research shows that in mature markets such as the London Stock Exchange, the order book's post-trade recovery half-life is about 20 seconds [21]. In crypto perpetual futures markets, however, the spread-recovery process is often more complex, owing to the widespread use of leverage and the high degree of market fragmentation: under normal conditions the recovery half-life of top exchanges can match or even surpass that of traditional markets, but during periods of stress it may be significantly delayed. Figure 25-4 compares the spread-recovery performance of several major exchanges under both normal and stressed conditions.

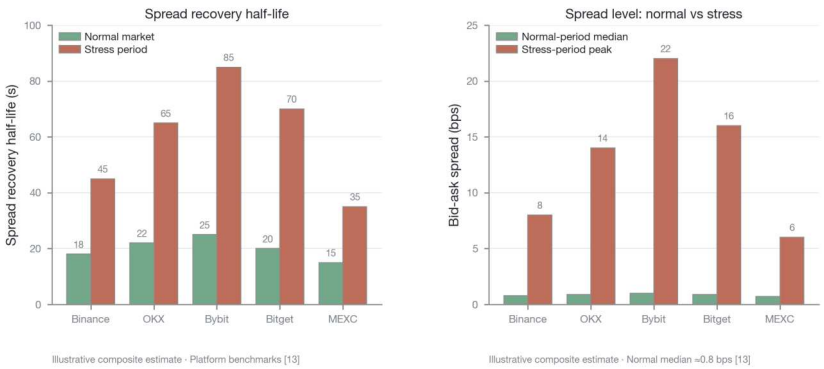


Figure 25-4. Cross-platform comparison of liquidity-resilience indicators (illustrative composite estimate, not direct measurement): the left panel shows the spread-recovery half-life (normal market versus periods of stress), and the right panel shows the median spread in normal periods (≈0.8 basis points) versus the peak spread during stress; recovery diverges markedly under stress and the peak spread widens substantially (data source: commercial-platform review [13] and exchanges' market data)

The cross-platform comparison in the figure clearly shows the differences in resilience across exchanges. Under normal market conditions, the Bitcoin perpetual futures spread on major exchanges is extremely small, with a median typically held at about 0.008% (about 0.8 basis points) [13]. During periods of stress, however, recovery speeds diverge markedly across platforms. Some can restore the spread to normal within 30 to 40 seconds, while others may need more than 1 minute [13]. This difference arises not only from the composition of each platform's market-maker community but, more fundamentally, from the efficiency of each exchange's liquidation engine and the design of its risk-management mechanisms. Platforms with a shorter recovery half-life generally offer traders a more reliable execution environment and thus attract greater liquidity inflows during a crisis.

25.4.3 The speed of order-book reconstruction

Depth resilience concerns the physical process by which the order book rebuilds after a shock. When a large market order sweeps through the order book and consumes liquidity across multiple price levels, it leaves a *hole* in the market. Depth resilience measures the speed at which new limit orders fill these holes and restore the original market depth. This process depends heavily on the behavioral logic of high-frequency traders and market makers.

Figure 25-10 presents the liquidity baseline before a shock: the order-book depth within a $\pm 0.1\%$ price band around the best price under normal market conditions on major exchanges, with most leading platforms in the range of $40 million to $50 million and a few notably higher. It is precisely this baseline that is punched through in an instant under a violent shock, and the speed of its reconstruction determines a platform's depth resilience.

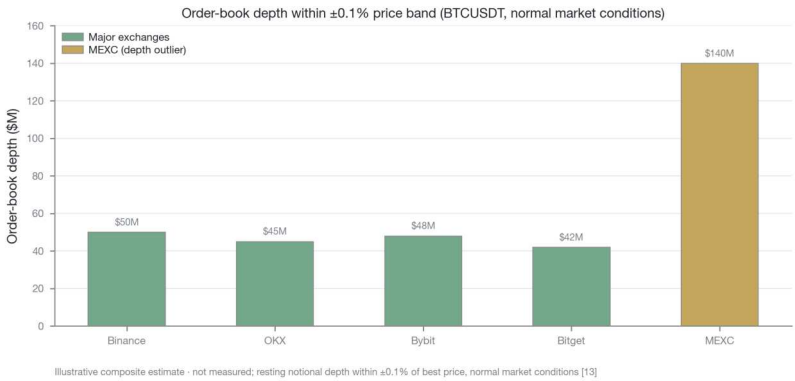


Figure 25-10. Order-book depth of Bitcoin perpetual futures on major exchanges (illustrative composite estimate, not direct measurement; notional resting depth within a ±0.1% price band around the best price, normal market conditions; data source: composite estimate from exchanges' market data [13])

After a shock, market makers face a dilemma. On one hand, the widened spread offers higher market-making profits; on the other, violent price swings entail enormous inventory risk. In crypto markets, market makers often adopt a *retreat-observe-return* strategy. The market crash of March 2020 offers an extreme case of depth-resilience failure. At the time, the Bitcoin price plunged in a short span, triggering a large-scale wave of liquidations [22]. Faced with enormous one-sided selling pressure and a cross-platform spread of more than $500, many market makers simply shut down their trading algorithms and stopped providing liquidity [22].

During that crisis, the order-book depth of one leading derivatives exchange dried up at one point to only about $20 million in buy orders, while the long positions awaiting liquidation reached as much as $200 million [22]. Had all these liquidation orders been dumped at market, the price of Bitcoin might have been driven down to near zero. This total collapse of depth resilience shows that when a market shock exceeds market makers' risk tolerance, the order book's reconstruction mechanism fails entirely. Only after the price stabilized at a new equilibrium, or the exchange intervened (for example, by halting trading), did market makers begin submitting orders again and the order-book depth slowly recover.

25.4.4 The capacity to withstand liquidation cascades

System resilience is the dimension of resilience unique to—and most critical for—perpetual futures markets. It measures the market's survival capacity when confronted with a liquidation cascade. A liquidation cascade is a vicious cycle in which a price decline triggers the liquidation of long positions, the market sell orders generated by those liquidations push the price down further, and the lower price triggers still more liquidations. In crypto markets, which lack circuit breakers and price limits, liquidation cascades are the primary threat leading to systemic collapse.

To withstand liquidation cascades, perpetual futures markets have developed a distinctive set of buffering mechanisms, of which the insurance fund is the most central.

When a liquidated position fills below its bankruptcy price, the resulting shortfall is covered by the insurance fund, protecting the profits of winning traders from being clawed back. The size and burn rate of the insurance fund are important indicators of system resilience. The insurance funds of major exchanges are not a single pool but are established separately by currency and contract type: BTC perpetual futures, ETH perpetual futures, and the various altcoin perpetual futures each have their own independent insurance fund. While this segmented structure provides risk isolation, it also means that the insurance-fund protection for long-tail asset contracts is far weaker than for major assets: in extreme conditions, the insurance fund for a given altcoin contract may already be exhausted and have triggered auto-deleveraging (ADL) while the fund for the BTC contract remains ample. If a single market shock consumes most or all of the insurance fund, the market's system resilience is near its limit. In the extreme, if the insurance fund is exhausted, the exchange must activate the auto-deleveraging mechanism to forcibly close profitable positions, which severely damages the market's credibility and liquidity.

The design of the liquidation mechanism has a profound effect on system resilience. Some early exchanges, for example, accepted only Bitcoin as margin, which meant that when the price fell, long traders faced not only losses on their positions but also a shrinking value of their margin—a nonlinear leverage effect that significantly increased the risk of a liquidation cascade [22]. As the market has evolved, modern exchanges have generally introduced multi-currency blended margin and tiered liquidation mechanisms, which effectively slow the instantaneous impact of liquidations on the market and thereby improve the system's overall resilience.

The architectural choice of the margin system has a decisive effect on the capacity to withstand liquidation cascades. The isolated-margin model isolates the risk of a specific position by allocating independent margin to each contract; its advantage is that it caps the loss from any single bad decision. Isolated margin, however, leaves each position with very little buffer, making it more prone to frequent micro-liquidations during violent volatility and thereby increasing trading friction and short-term volatility at the micro level [23]. The cross-margin model lets a trader use all available balance in the account as shared collateral, which significantly lowers the probability of liquidation under normal market conditions and allows the portfolio to better absorb local shocks [23]. But cross margin proves extremely fragile in extreme tail events: when the market suffers a broad sell-off, the synchronized decline of correlated assets can rapidly exhaust the entire account's margin, causing all positions to be force-liquidated at once and instantly releasing enormous selling pressure into the market. At the system level, when most traders use cross margin, this individually reasonable risk-control choice produces a fallacy of composition in the aggregate: the crash of a single asset liquidates not only that asset's position but also, by extension, all other positions in the same account, spreading a shock originally confined to one market into cross-asset systemic selling pressure.

Tiered liquidation mechanisms and the insurance fund form a second line of defense that buffers the shock of liquidations. Modern liquidation engines have evolved from the early days of simple market-order closeouts into complex multi-stage systems: the first stage attempts to liquidate with limit orders, placing the liquidated position into

the order book to await matching; the second stage, when the limit orders fail to fill, initiates counterparty matching to find market makers willing to take over the position at a discount; and only the final stage executes a market liquidation. Some exchanges have also introduced a liquidation-delegation mechanism that allows third-party market makers to proactively take over liquidated positions at a price better than the bankruptcy price, thereby reducing the impact on the market. Tiered liquidation refines this further by reducing positions gradually in batches according to market depth and price slippage, smoothing the selling pressure to some degree. The insurance fund serves as a pool for absorbing shortfall losses, using the penalties collected in routine liquidations to cover systemic losses in extreme situations; but its size is typically only a small fraction of total open interest [23].

25.4.5 The resilience function of the funding rate

The funding rate is the most central innovation of perpetual futures. It was designed as a price-anchoring mechanism that, through periodic funding exchanges between longs and shorts, forces the perpetual futures price back toward the spot index price. Under normal market conditions, the funding rate serves as an effective corrective mechanism: when the contract price is above spot, longs pay shorts, dampening excessively bullish sentiment, and vice versa. This anchoring mechanism continuously draws arbitrage capital into the market, sustaining pricing efficiency and forming an important micro-foundation of market resilience.

In extreme conditions, however, the funding-rate mechanism may not only fail to perform its resilience function but also exhibit strong procyclicality that amplifies market volatility. When the market swings violently in one direction, speculative order flow tends to overwhelm the capacity of arbitrage capital to absorb it [24]. The funding rate then spikes to extreme levels. Traders holding counter-trend positions must bear not only large paper losses but also steep funding payments. This dual negative effect accelerates the depletion of margin, forcing traders to close early or be force-liquidated, which in turn drives the price still further from fundamentals. The failure of the funding rate's resilience function is not a sudden jump from normal to collapse but a process of gradual exhaustion: during a spell of moderately high funding, arbitrageurs keep stepping in to compress the funding deviation, but their margin buffer is progressively consumed at each settlement cycle. When a truly extreme shock arrives, arbitrage capital has already been significantly weakened, so the corrective mechanism collapses far faster than it would from a fully normal starting state. This path dependence means that after a prolonged one-sided market, the market's resilience to a sudden shock is systematically lower than in its initial state.

Another key condition for the failure of the funding rate's resilience function is the exhaustion of underlying liquidity. When spot-market liquidity dries up, small movements in the index price are amplified, so the basis for calculating the funding rate loses its representativeness. Some exchanges, moreover, set a funding-rate cap to prevent the rate from excessively penalizing one-sided position holders. Under extreme sentiment, when the true degree of market imbalance far exceeds this cap, the funding rate loses its price-anchoring and supply-demand-adjusting functions, causing the perpetual futures

price to diverge from the spot price over a prolonged period and to a severe degree.

25.4.6 Cross-exchange contagion

The crypto-asset market is composed of many mutually independent exchanges, yet they are tightly connected through cross-exchange arbitrageurs, shared market makers, and highly overlapping pools of collateral assets. This network structure means that a crisis at any single node can rapidly evolve into a market-wide systemic crisis, constituting the ultimate test of market resilience.

The mechanism of cross-exchange price contagion works mainly through two channels. The first is the direct transmission of a liquidity shock. When a large exchange suffers a run or a liquidation cascade, asset prices on that platform plunge. Because arbitrage algorithms monitor price differences across platforms at the millisecond level, they quickly sell the same assets on other, still-functioning exchanges to lock in profits, transmitting the price collapse across the entire network in an instant. The second channel is balance-sheet cross-contagion. Many institutional investors hold exposure on multiple platforms at once; when a position on one platform is force-liquidated, they must liquidate healthy positions on other platforms to top up margin, and this cross-platform deleveraging spreads the liquidity crisis to assets and exchanges that were originally unrelated.

Beyond these two channels, the bidirectional contagion between decentralized protocols and centralized exchanges forms a third, increasingly important transmission channel. Large-scale liquidations on centralized exchanges transmit through oracle price feeds to on-chain lending and perpetual futures protocols, triggering on-chain liquidations and imbalances in liquidity pools; conversely, the collapse of an on-chain protocol (such as the Terra-Luna event of 2022) transmits the shock to centralized exchanges through arbitrageurs and shared collateral. A stablecoin depeg constitutes a particularly dangerous, standalone contagion channel: when the peg of USDT or USDC deviates, stablecoin-denominated margin shrinks simultaneously across all exchanges and protocols, triggering liquidations synchronously across venues and protocols, with a contagion reach far exceeding that of a crisis in a single asset or on a single platform.

The collapse of FTX in November 2022 offers a canonical case for studying systemic contagion. As the world's third-largest crypto derivatives exchange at the time, FTX's bankruptcy not only destroyed its native token but also produced a significant downside-risk spillover. Empirical research shows that during the FTX collapse, assets deeply tied to the platform bore the strongest risk contagion, and even the relatively independent Bitcoin and Ethereum underwent violent repricing as overall market liquidity was withdrawn [25]. More strikingly, the FTX crisis left many cross-exchange market makers badly wounded or bankrupt, and the network-wide liquidity baseline fell sharply over the following months, inflicting lasting damage on the market's overall resilience.

Figure 25-5 renders the contagion mechanisms above as a network structure, in which the nodes represent major exchanges and key institutional participants, and the thickness and direction of the edges reflect the intensity and pathways of liquidity-shock and balance-sheet contagion. The key amplifying nodes marked in the figure—large market makers and hedge funds connected to multiple exchanges at once—act as liquidity

bridges in normal times but become accelerators of systemic contagion during a crisis.

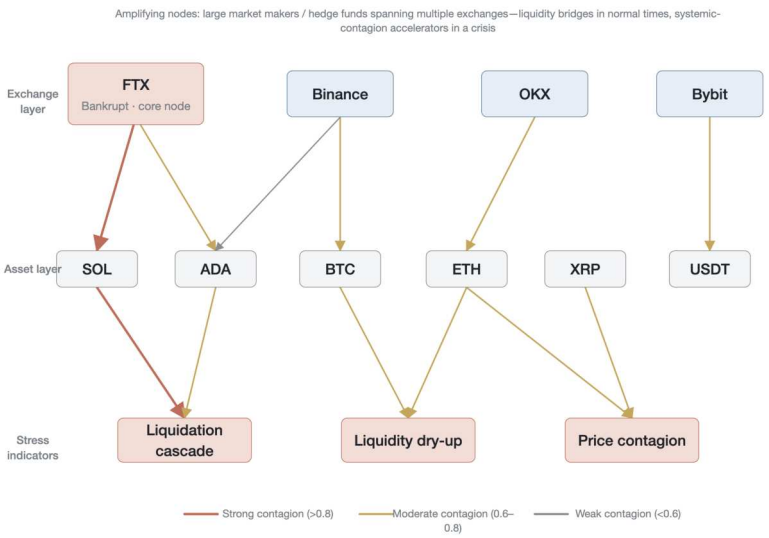


Figure 25-5. Network structure and transmission pathways of cross-exchange stability contagion (a representative reconstruction based on empirical research on the FTX collapse [25] and the October 2025 liquidation cascade [20][26]; edge strengths are qualitative gradations, not measured weights)

Similar network effects have recurred in subsequent episodes of market stress. In the liquidation cascade triggered by a macroeconomic shock in October 2025, for example, an unprecedented scale of liquidations occurred network-wide within about 36 hours (measured as notional value liquidated, not open interest wiped out) [20]. In this crisis, the intensity of cross-asset contagion even exceeded that of previous macro stress tests, demonstrating once again that, absent a network-wide coordination mechanism, the risk controls of individual exchanges cannot halt the spread of a systemic crisis.

25.4.7 Cross-platform comparison and evolution

Although the liquidity of perpetual futures markets has improved significantly over the past few years, resilience has been the slowest to improve. This is because liquidity can be enhanced through short-term incentives such as lower fees or market-maker rebates, whereas strengthening resilience requires systematic institutional reform and long-term investment in infrastructure. Viewed through the lens of historical evolution, the resilience of crypto markets has already undergone several major tests and rounds of evolution.

Comparing several major market crises reveals the trajectory of resilience's evolution clearly. In the crash of March 2020, market structure nearly collapsed entirely: network congestion prevented arbitrageurs from moving funds between exchanges, intensifying local liquidity exhaustion [22]. By the Terra-Luna collapse of May 2022, although the market again experienced violent price swings and large-scale liquidations, the systems of major exchanges did not suffer prolonged outages, and the recovery time for spread and depth had shortened from hours to tens of minutes. In the turmoil triggered by

FTX's bankruptcy in November 2022, although market confidence was badly shaken, the liquidity metrics of the remaining major platforms largely returned to normal within 10 to 20 minutes [27]. Attributing this recovery entirely to a rise in the maturity of market structure, however, calls for caution. The FTX collapse was essentially a credit event rather than a pure liquidity event, and the rapid recovery of liquidity on other platforms stemmed in part from the migration of FTX users' funds and trading activity to surviving platforms—essentially a mechanism of liquidity redistribution rather than the market's own capacity to regenerate liquidity. Distinguishing these two recovery mechanisms is essential to any accurate assessment of market resilience: liquidity redistribution works only when some platforms are in crisis and cannot be relied upon in a market-wide systemic shock.

The improvement of resilience, however, is not without cost; it often requires a trade-off against transaction costs. To maintain higher resilience, an exchange must build a larger insurance fund, implement stricter risk controls, and may impose stricter quoting obligations on market makers. These measures inevitably raise the system's overall friction cost, which is ultimately passed on to traders. Market resilience, moreover, comprises not only the liquidity- and liquidation-level recovery capacity described above but also technical reliability at the operational level. Exchange system outages, API interruptions, DDoS attacks, and security incidents all pose direct threats to market continuity. The 2024 Bybit security incident and the repeated system overloads at centralized exchanges during periods of extreme volatility show that a platform excelling in liquidity resilience but suffering frequent failures of its technical infrastructure will see its effective resilience sharply discounted. Assessing operational resilience should incorporate metrics such as system availability, the frequency of security incidents, and the speed of disaster recovery.

The future design of perpetual futures markets therefore needs to find an optimal balance between extreme liquidity and reliable resilience, so as to ensure that the market maintains its core functions in both normal and stressed states.

25.5 Fairness

Fairness is a dimension often neglected in traditional market-quality evaluation but especially prominent in crypto markets. In traditional financial markets, fairness is usually enforced by regulators through unified rules and severe penalties, so it is treated more as a *precondition* than as a market-quality metric requiring continuous assessment. In perpetual futures markets, however, the absence of a unified global regulatory framework and the fundamental architectural differences between centralized exchanges and decentralized protocols make fairness a highly variable core variable that directly affects participant welfare.

Fairness in perpetual futures markets is not a purely moral question but a profound question of market efficiency. When market rules are not applied equally, or when certain participants obtain systematic information privileges, the resulting unfairness translates directly into transaction costs for other participants. Information privileges and execution inequality amplify the advantage of informed traders, and market mak-

ers, to compensate for the risk of being adversely selected, inevitably widen the bid-ask spread or reduce liquidity supply. Ultimately, this unfairness—produced by institutional design or technical architecture—harms the welfare of all uninformed traders and lowers the overall quality of the market.

25.5.1 A multilayered structure

Market fairness is not a single-dimensional concept but is composed of several interwoven layers. To assess the fairness of perpetual futures markets systematically, we decompose it into three core layers: execution fairness, information fairness, and institutional fairness. These three layers correspond, respectively, to the unequal treatment a trader may encounter in *how they trade*, *on what information they trade*, and *under what rules they trade*.

Figure 25-6 presents the nested relationship among the three layers of execution fairness, information fairness, and institutional fairness, and their transmission paths to market-quality metrics; the three layers rise from the most micro-level execution quality up to the institutional layer that defines the boundaries of the rules of the game. The figure also shows how different architectures perform differently: centralized exchanges face greater challenges at the execution- and information-fairness layers (arising from market-maker privileges and colocation advantages) but can repair them at the institutional layer through internal governance; decentralized exchanges perform excellently at the institutional layer (with rules hardcoded into smart contracts) yet face a novel unfairness threat in maximal extractable value (MEV), which has the dual properties of both information and execution.

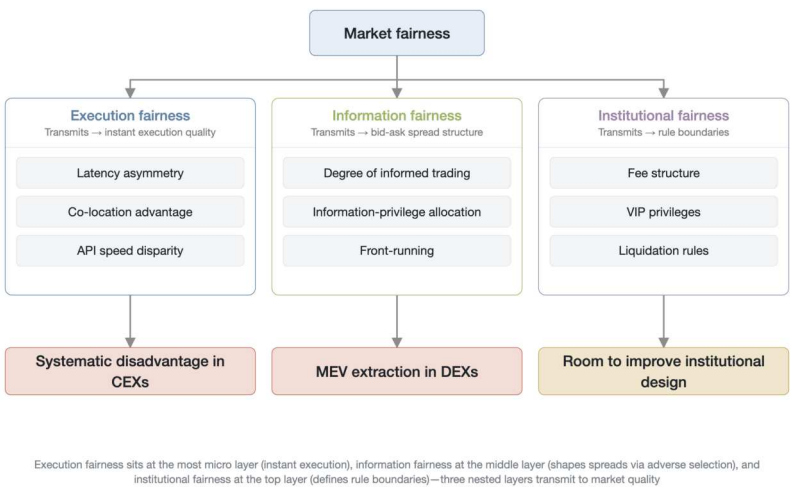


Figure 25-6. The multilayered structure of fairness and its market-impact pathways (author's integrated conceptual illustration, not measured data; theoretical basis: Angel & McCabe, 2013 [28]; Lin et al., 2019 [29])

Execution fairness concerns differences among participants in the speed of order submission and processing. When certain participants obtain a systematic speed advantage through technical means or special channels provided by the exchange, execution fairness is undermined, and the orders of ordinary traders always lag behind those of

high-frequency traders (the mechanism is elaborated in Section 25.5.2).

Information fairness concerns the distribution of market data and order-flow information. In an ideal market, all public information should be available to all participants simultaneously. In real perpetual futures markets, however, the degree of informed trading and the distribution of information privileges vary significantly. Market makers or high-frequency trading firms are often able to obtain deeper order-book data or faster market-data feeds. The classic model of Glosten and Milgrom (1985) shows that, under information asymmetry, a market maker's bid-ask spread is a rational compensation for adverse-selection risk [30]. When information privileges are monopolized by a tiny number of participants, this asymmetry forces other liquidity providers to widen their spreads, raising transaction costs across the whole market.

Institutional fairness is the highest layer of fairness, concerning how equitably an exchange or protocol makes and enforces its rules, and it is also the most destructive source of unfairness—it not only distorts price discovery but can also give rise to conflicts of interest and even a collapse of market trust (the mechanism is elaborated in Section 25.5.3).

25.5.2 Execution fairness

Unfairness at the execution level stems mainly from high-frequency traders' speed advantage in order processing and information acquisition. In perpetual futures markets, price discovery and liquidity supply depend heavily on market makers' continuous quoting. This microstructure, however, also creates room for rent-seeking by participants with ultra-low-latency connections. Angel and McCabe (2013), in examining market fairness in the age of high-frequency trading, note that although high-frequency trading can improve market liquidity in some cases, the microsecond-level speed advantage it relies on essentially constitutes an implicit cost levied on other, slower participants [28].

On centralized exchanges, this execution unfairness typically manifests as differences in physical location and API access. Well-funded trading firms can use server-colocation services to place their trading servers in the same data center as, and as close as possible to, the exchange's matching engine. This shortened physical distance lets them receive price updates earlier than ordinary traders and send orders faster. Exchanges, moreover, often provide dedicated API channels to top market makers, further amplifying this latency asymmetry. When new market information arrives, firms with the speed advantage can cancel stale quotes or establish new positions ahead of ordinary traders; this *front-running* behavior often leaves ordinary traders facing wider slippage or rejected orders.

Beyond a pure speed advantage, market makers also frequently obtain information privileges. On many centralized perpetual futures exchanges, market makers can obtain order-flow information through specific data-sharing agreements that ordinary users cannot see. This privileged data access lets them predict short-term price movements more accurately. Although exchanges typically argue that they provide this data to help market makers manage risk and thus quote more tightly, in practice it creates an unfair information hierarchy. Informed traders use this privileged information to earn excess profits in the market, and those profits are ultimately paid by uninformed

traders (mainly retail) through wider effective spreads and higher slippage.

Under extreme conditions, execution-level unfairness gives rise to more serious problems. When the market swings violently and many positions face liquidation risk, firms that obtain liquidation information in advance and execute faster can establish positions near the liquidation price ahead of time and thus profit when the liquidation cascade unfolds. This behavior not only intensifies the market's excess volatility but also causes those being liquidated to suffer additional losses far exceeding their margin shortfall.

25.5.3 Institutional fairness

Unfairness at the institutional level stems from the deliberate favoritism of those who make the rules. In centralized perpetual futures markets, the exchange is not only the provider of the trading venue but often also the rule-maker, the referee, and even a participant. This multiplicity of roles creates severe conflicts of interest and gives rise to opaque market-maker agreements and extreme institutional favoritism.

The opacity of market-maker agreements is a typical expression of institutional unfairness. To attract liquidity, exchanges generally implement a maker-taker fee structure. Malinova and Park (2015) show that this fee structure can indeed significantly affect market quality and traders' behavioral choices [31]. In crypto markets, however, these agreements are often not disclosed. Top market makers can obtain not only a full rebate of maker fees but even additional cash subsidies or token rewards. These *VIP privileges* let large market makers trade at costs far below those of ordinary participants, giving them a decisive advantage in strategic competition. Although this price discrimination helps sustain nominal liquidity to some degree, it essentially constitutes an asymmetric cost burden on small and medium traders.

A more severe form of institutional unfairness arises from an exchange's proprietary market making and related-party trading. When an exchange operates its own market-making team or is closely tied by interest to a particular hedge fund, fairness is fundamentally compromised. The collapse of the FTX exchange in 2022 offers an extreme case of institutional unfairness. According to the complaint filed by the Commodity Futures Trading Commission (CFTC), FTX's affiliated hedge fund, Alameda Research, not only obtained special trading priority as the platform's principal market maker but also received a systematic exemption from risk controls at the code level [32].

In the FTX case, Alameda was able to bypass the automatic liquidation mechanism and continue trading even when its account balance was negative. This privilege let Alameda draw without limit on other customers' funds for high-risk speculation. Alameda also had faster API access than other institutional clients, letting it stay ahead of the market at the millisecond level [33]. This extreme institutional favoritism was not only unfair to all other traders but also fundamentally undermined the basis of trust in the exchange as a neutral matching engine. When the rule-maker can modify system parameters at will for particular participants, so-called market efficiency and price discovery degenerate into a channel through which internal stakeholders siphon off outside funds.

25.5.4 The fairness of DEXs

In response to the execution and institutional unfairness prevalent on centralized exchanges, decentralized exchanges offer an alternative built on blockchain architecture. One of the core value propositions of a DEX is to fundamentally eliminate human rule-favoritism and secret agreements through the properties of code-is-law and permissionlessness. Under this architecture, the fairness problem shifts from a *game between people and institutions* to a *game between people and code*.

DEXs enjoy a significant structural advantage in institutional fairness. In protocols based on automated market makers or on-chain central limit order books (CLOBs), all trading rules, fee structures, and liquidation parameters are hardcoded into open-source smart contracts. There are no opaque market-maker agreements, no hidden VIP channels, and no risk exemptions for an exchange's proprietary team. All participants—whether an institution with hundreds of millions of dollars in capital or a retail trader with only a few hundred dollars—must follow exactly the same execution logic. This on-chain transparency ensures the absolutely equal application of the rules. McMenamin et al. (2022) note that decentralized architectures have the potential to significantly reduce the value extraction of traditional markets through innovative protocol designs such as frequent batch auctions [34]. Yet *hardcoding* is not absolutely immutable: most DEX protocols adopt an upgradeable-contract model in which core parameters can be adjusted through governance votes or a multisignature wallet, and in practice governance power is often concentrated in the hands of a few large token holders. More critically, the sequencers of today's leading app-chain perpetual futures platforms (Hyperliquid, dYdX v4) are all run centrally by a single entity, which means that the sequencer operator theoretically holds trade-ordering power similar to that of a centralized exchange, including the ability to reorder transactions or selectively censor them. These factors show that although the DEX advantage in institutional fairness is real in terms of rule transparency, it still rests on trust assumptions—regarding how far power is actually dispersed—that are not fully consistent with its decentralization narrative.

The transparency and openness of DEXs, however, also introduce a new type of fairness challenge, the most prominent being the problem of maximal extractable value. In a blockchain network, because all pending transactions are publicly exposed in the mempool and block builders hold the power to decide the order in which transactions are packed, an unfairness arises that has the dual properties of both information and execution. Daian et al. (2019) describe in detail the phenomena of front-running and transaction reordering on DEXs, noting that these cause instability at the consensus level and directly harm the interests of ordinary traders [35].

On perpetual futures DEXs, MEV typically takes the form of a sandwich attack. When a searcher observes in the mempool a large order likely to cause significant price slippage, the searcher pays a higher priority fee to submit a same-direction trade ahead of that order, pushing the price up, and then immediately closes in the opposite direction for a profit once the order executes. This attack forces the victim to fill at a worse price. According to Uniswap's analysis, although this value extraction is a byproduct of blockchain transparency, it is essentially a hidden tax levied on ordinary users by exploiting the power to order transactions [36].

Beyond MEV attacks, liquidity providers on decentralized exchanges face a more insidious structural loss. The loss-versus-rebalancing framework proposed by Milionis et al. (2022) reveals a fundamental fairness defect in automated market makers: when the price on a centralized exchange updates first, informed arbitrageurs can execute trades against the stale quotes on a decentralized exchange, systematically extracting value from liquidity providers [37]. The existence of loss-versus-rebalancing (LVR) shows that even in a decentralized architecture that has eliminated institutional privileges, information asymmetry still imposes a hidden tax on liquidity providers through the channel of cross-market arbitrage—a fairness problem distinct from the market-maker privileges of centralized exchanges but no less deserving of attention.

Fairness on DEXs is further constrained by technical capability and capital scale. Although the rules are equal for everyone, the technically adept who can write complex smart contracts to arbitrage clearly hold an enormous advantage over ordinary users who can only trade manually through a web front end. At the same time, in protocols based on token governance, well-capitalized large holders can change protocol parameters by voting, achieving rule changes that serve their own interests. Thus, although DEXs eliminate the privileges of centralized institutions, they do not achieve absolute fairness; rather, they redefine the sources of advantage under transparent rules.

25.6 Accessibility

Accessibility measures the extent to which a market allows a broad range of participants to enter trading across the geographic, capital, and technical dimensions. In traditional financial markets, eligibility to participate is set in advance by regulatory rules, and the barrier to entry is treated as an exogenous institutional constraint rather than an endogenous variable of market quality. In crypto perpetual futures markets, however, permissionless decentralized architectures coexist with centralized platforms that impose strict identity screening, making accessibility a core quality dimension that directly determines the diversity of participants and the depth of liquidity pools. A market that performs excellently on transaction costs and price discovery but sets high barriers to entry for most potential participants will see its overall quality sharply discounted by the narrowness of its liquidity base.

25.6.1 Definition and measurement framework

Accessibility can be decomposed into three interrelated sub-dimensions. Geographic accessibility measures how open the market is to users in different jurisdictions, including regional bans, tiered Know Your Customer (KYC) screening regimes, and the coverage of fiat on- and off-ramp channels. Capital accessibility concerns the screening effect that minimum deposit thresholds, VIP volume tiers, and margin requirements have on participants of different capital sizes. Technical accessibility assesses the operational complexity users face in wallet setup, gas-fee payment, smart-contract interaction, and cross-chain bridging.

To quantify these three sub-dimensions, the following operational indicators can be constructed: a geographic coverage index (the share of the global active crypto market

represented by the jurisdictions a platform can serve), minimum viable capital (the smallest amount of funds needed to open and maintain a meaningful position on a given platform), and a technical-barrier score (the number of steps and the level of technical knowledge required from wallet creation to completing a first trade) [38][39]. These indicators turn accessibility from a qualitative concept into a quantifiable dimension of market quality.

25.6.2 The entry barriers of CEXs

Centralized exchanges construct an implicit hierarchy of access through a tiered identity-verification system. At the most basic level, KYC screening excludes users who cannot provide valid proof of identity. Geographically, U.S. residents are excluded from the vast majority of offshore perpetual futures platforms, while users in several Asian and African countries face partial or complete service restrictions [38][40]. At the capital level, the VIP tier system constitutes structural discrimination against retail traders: as noted in Section 25.2.1, the $100 million monthly-volume threshold required to reach Binance's VIP 3 tier makes competitive fees structurally unattainable for the vast majority of retail users [10].

Users in developing economies face even harsher friction in fiat on- and off-ramps. The lack of bank-transfer channels, the high fees charged by payment processors, and the exchange-rate losses on local-currency conversion make the real cost of entering the crypto derivatives market from fiat far exceed explicit trading fees [41]. This friction not only limits individual participants' market access but also systematically weakens the ability of particular geographic regions to contribute to the global price-discovery process.

25.6.3 The permissionless access and technical barriers of DEXs

One of the core value propositions of decentralized exchanges is permissionless access: any user holding a digital wallet can participate in trading without identity verification [8][42]. This architecture fundamentally eliminates the geographic restrictions and identity-screening barriers of centralized platforms, achieving the greatest possible openness at the institutional level. Technical barriers, however, largely replace institutional barriers as the main constraint on the accessibility of decentralized markets.

Wallet management, the secure storage of seed phrases, gas-fee optimization, smart-contract risk assessment, and cross-chain asset bridging form a series of operational steps of considerable complexity for non-technical users. Smart-contract security risk deserves particular attention: in interacting with a DEX, non-technical users face not only operational complexity but also the security risk that carelessly authorizing a malicious contract can lead to the theft of their assets. This security barrier partly offsets the permissionless-access advantage of DEXs in practice. On the Ethereum mainnet, gas fees during periods of network congestion can make small trades economically unviable, creating a de facto minimum trade-size threshold. The rise of app-chain architectures exemplified by Hyperliquid and dYdX v4 has significantly eased this technical barrier by migrating trade execution to a dedicated chain and drastically lowering trading fees [43]. While lowering trading fees, however, app-chains also introduce a new barrier:

cross-chain asset bridging. Users must bridge assets from the Ethereum mainnet or another chain to the app-chain, a process that involves trust assumptions about the bridge's security, a wait of several minutes to tens of minutes, and additional bridging fees. For non-technical users, the cognitive complexity and security risk of bridging may be higher than gas-fee optimization itself. Account abstraction and mobile-first interface design have further reduced the complexity of wallet management, so that the user experience of decentralized markets is gradually converging toward that of centralized platforms.

25.6.4 The feedback loop between accessibility and market quality

Accessibility is not an isolated dimension of market quality but forms a deep feedback loop with the other dimensions through the intermediate variable of participant diversity. Broader access means a more diverse body of participants—with different risk preferences, different information sets, and different trading time zones—converging in the same market, which deepens liquidity pools, improves price-discovery efficiency, and strengthens the market's resilience to local shocks. Restricted access, by contrast, concentrates participation among professional institutions; while this may raise execution efficiency in the short run, it also leaves the market overly dependent on a few homogeneous actors for liquidity supply and risk-bearing, so that it proves more fragile in periods of stress.

The marked climb in DEX trading volume during regulatory crackdowns on centralized exchanges provides natural-experiment evidence of accessibility's effect on market quality. When a particular jurisdiction restricts users' access to centralized platforms, trading activity migrates rapidly to permissionless decentralized protocols, revealing participants' strong preference for accessibility [44]. This migration pattern shows that the accessibility constraint is not merely a question of welfare distribution but a structural force shaping the global allocation of liquidity. Section 25.8.1 further examines the regulatory dilemma between consumer protection and market accessibility.

25.6.5 Cross-platform comparison and evolution

On the accessibility dimension, centralized and decentralized exchanges present a complementary pattern of strengths and weaknesses. Through smooth user interfaces, well-developed fiat channels, and professional customer service, centralized exchanges greatly lower the technical barrier, letting users without a technical background participate in trading with ease. Their geographic coverage, however, is strictly limited by regulatory-compliance requirements, and the tiered KYC system, while protecting investors, also constitutes an institutional constraint on global participation. Decentralized exchanges, by contrast, have a structural advantage in geographic accessibility—their permissionless architecture means any internet user can participate—but their higher technical complexity forms a de facto entry barrier at the level of user experience.

These two architectures are undergoing marked convergent evolution. Centralized exchanges have begun to introduce on-chain settlement and non-custodial wallet options, seeking to grant users greater autonomy over their assets within a compliance frame-

work. Decentralized exchanges, through technical innovations such as account abstraction, social login, and mobile-native applications, keep lowering the technical barrier [43][45]. This bidirectional convergence shows that market competition is driving an overall improvement in the accessibility dimension. The focus of future competition will shift from *whether access is possible* to *at what cost access is possible*—that is, how to minimize the combined friction that users face across the geographic, capital, and technical sub-dimensions while maintaining broad access [41].

25.7 Integrated assessment

A single-dimensional market-quality metric is inevitably one-sided. As noted earlier, an extremely narrow bid-ask spread or extremely high informational efficiency alone cannot establish that a market possesses high overall quality. Only by integrating the five dimensions of transaction costs, price discovery, resilience, fairness, and accessibility into a composite index can we capture the full picture of market quality. This integration, however, itself faces the fundamental challenges of weight selection and trade-offs among dimensions. This section sets out in detail the methodology for constructing a five-dimensional market-quality index and applies it to comparative analysis across platforms and across architectures.

25.7.1 Construction methodology

The core objective of the five-dimensional market-quality index is to convert microstructure indicators of different units and properties into comparable, standardized scores. This process requires establishing a unified measurement benchmark while preserving the independent characteristics of each dimension.

Construction of the index begins with the standardization of base indicators. The transaction-cost dimension relies chiefly on the reciprocal of the effective spread and market-impact cost; the price-discovery dimension uses the Hasbrouck information share or the variance-ratio-test score; the resilience dimension is quantified as the spread-recovery half-life and depth-recovery rate after a liquidity shock; the fairness dimension is based on inverse indicators of market-maker-agreement transparency and the probability of informed trading; and the accessibility dimension jointly assesses the minimum capital threshold and the degree of global access (see Section 25.6). These base indicators are mapped onto a 0-to-100 score range through min-max normalization, turning originally incommensurable physical quantities into quality scores on a common scale [7]. Min-max normalization is highly sensitive to extreme values, and extreme values occur frequently in crypto markets—the bid-ask spread of a given altcoin, for example, may suddenly spike to several hundred basis points during a liquidity crisis. In practice, one can use quantile normalization or winsorize extreme values to enhance robustness. Moreover, the choice of reference sample for normalization (whether to benchmark against all assets or only against major assets) fundamentally changes the score ranking of the platforms, so the reference scope must be stated explicitly in any specific application.

Once standardization is complete, computing the composite index confronts the core

difficulty of weight selection. An equal-weighting scheme assumes that the five dimensions contribute identically to overall market quality; this treatment is mathematically simpler but often lacks solid theoretical support in economic terms. A functional-weighting scheme, by contrast, has greater explanatory power. For a high-frequency arbitrageur, for example, the transaction-cost and price-discovery dimensions carry far greater weight than the others, whereas for a long-term institutional investor or a regulator, the resilience and fairness dimensions dominate. The practical application of the five-dimensional index therefore does not pursue a single universal total score but emphasizes a weighted result under a specific evaluation objective.

Although the five-dimensional model offers a more comprehensive evaluative perspective, its limitations must not be ignored. The incommensurability of the dimensions means that a higher score on one dimension cannot fully compensate for a lower score on another. The score gain from lowering transaction costs through extreme leverage, for example, cannot offset the collapse in the resilience score caused by the resulting liquidation cascade. This nonlinear relationship requires that, in applying the five-dimensional index, we not look at the weighted total alone but examine the distribution of scores across dimensions and the trade-offs inherent among them.

25.7.2 A cross-sectional comparison of CEXs

Figure 25-7 uses a heatmap to compare the quality scores of Binance, OKX, Bybit, and the decentralized exemplar Hyperliquid across the five dimensions of transaction costs, price discovery, resilience, fairness, and accessibility. The figure exhibits a striking diagonal pattern of divergence: centralized exchanges lead on the efficiency-oriented dimensions, whereas the decentralized exchange shows an advantage on the fairness and accessibility dimensions.

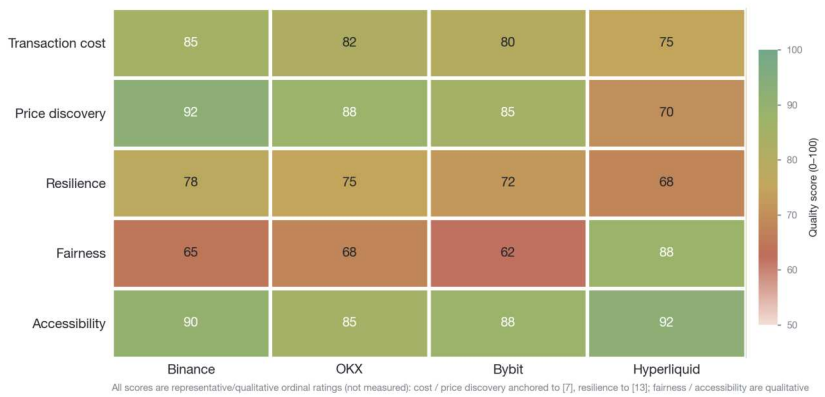


Figure 25-7. Heatmap of five-dimensional market-quality scores for major exchanges (the scores on all five dimensions are representative/qualitative ordinal ratings, not precise direct measurements; the transaction-cost and price-discovery scores are based on the methodology of Barbon and Ranaldo [7], the resilience score is estimated from platforms' publicly disclosed insurance-fund data and spread-recovery speed [13], the fairness score is a qualitative rating based on market-maker-agreement transparency and API-access tiering, and the accessibility score is based on geographic access restrictions and the minimum capital threshold)

On the two *efficiency-oriented dimensions* of transaction costs and price discovery, Binance shows a marked lead. Its deep order-book liquidity and vast trading-volume base keep the bid-ask spread at an extremely low level over the long run while it also dominates the price-discovery process. OKX and Bybit follow close behind on these dimensions, and the gap among the three leading platforms is gradually narrowing, reflecting a homogenization trend driven by market-maker competition and infrastructure upgrades [46].

On the *institutional-safeguard dimensions*, however, the platforms perform with different characteristics. On resilience, leading centralized exchanges, backed by large insurance funds and mature liquidation engines, can restore liquidity relatively quickly when facing medium-sized market shocks. On the fairness dimension, however, centralized exchanges generally score low. This stems mainly from the opacity of their market-maker agreements and the execution privileges enjoyed by VIP clients. Although this institutional design effectively lowers overall transaction costs, it inevitably harms the relative fairness experienced by ordinary traders.

25.7.3 A comparison of CEXs and DEXs

Centralized and decentralized exchanges represent two sharply different paradigms of market architecture, and this underlying architectural difference profoundly shapes their quality characteristics within the five-dimensional model.

Decentralized exchanges hold a structural, absolute advantage in fairness and transparency. On platforms exemplified by Hyperliquid or dYdX, the core logic rests on the enforced execution of on-chain rules. All liquidation parameters, fee structures, and order-matching logic are fully public to every participant, thoroughly eliminating the private market-maker agreements and information privileges prevalent on centralized exchanges [42]. The permissionless nature of the decentralized architecture, moreover, gives it an equally strong performance on the accessibility dimension: any user with a digital wallet can participate in the market without hindrance.

This advantage in fairness and accessibility, however, comes at the expense of some transaction cost and resilience. Decentralized exchanges can often offer competitive execution costs for large trades [7] but still face a cost disadvantage for small trades. For AMM-type decentralized exchanges on the Ethereum mainnet, the fixed gas fee is the main reason small trades are too costly; for app-chain-architecture perpetual futures platforms (such as Hyperliquid and dYdX v4), trades themselves incur no gas fee, and the cost disadvantage for small trades comes mainly from the larger price impact caused by relatively limited liquidity depth rather than from a fixed fee. More critical is the resilience dimension: decentralized platforms face a greater risk of liquidity exhaustion under extreme market stress. When the market swings violently, a decentralized liquidity pool that lacks the backstop of centralized market makers may dry up rapidly, causing the spread to widen sharply and the price-discovery function to stall briefly. This shows that CEXs and DEXs are not simple substitutes in market quality but are highly complementary.

25.7.4 Dynamic evolution

Market quality is not static cross-sectional data but a dynamic process that evolves continually over time. Looking back over the development from 2019 to 2025, the perpetual futures market shows a significant but uneven trajectory of improvement across its dimensions.

Over the past 6 years, the transaction-cost and price-discovery dimensions have improved the fastest. With the full-scale entry of high-frequency trading firms and upgrades in matching-engine technology, the bid-ask spread on major assets has fallen by nearly an order of magnitude [47]. At the same time, the information share of perpetual futures in price discovery has risen steadily, transforming them from an early spot follower into the primary pricing venue. This rapid improvement in the *efficiency-oriented dimensions* is due chiefly to the direct drive of fierce commercial competition and technical innovation.

By contrast, the two *institutional-safeguard dimensions* of resilience and fairness have improved at a relatively slow pace. Although the major platforms have progressively refined their liquidation mechanisms and expanded their insurance funds, momentary liquidity vacuums and liquidation cascades continued to occur frequently during the repeated extreme market shocks of 2022 and 2025 [48]. This shows that technical upgrades alone cannot fundamentally resolve the endogenous fragility produced by high leverage and procyclical mechanisms. Progress on fairness has been equally slow, with market-maker privileges and information asymmetry remaining persistent problems in the market microstructure. The future direction of market-quality improvement will necessarily require a shift from purely technical efficiency competition toward deeper optimization of institutional design and architectural innovation.

25.8 Regulatory implications

Measuring market quality is not merely an academic description of the trading environment but a scientific basis for making regulatory policy. The regulatory frameworks of traditional financial markets are largely built on intervening in specific market-quality defects—for example, improving price-discovery efficiency by requiring disclosure, or strengthening system resilience through capital-adequacy requirements. Transposing this regulatory logic to perpetual futures markets, however, confronts regulators with entirely new challenges. The decentralized character, cross-border nature, and around-the-clock trading of the crypto derivatives market often cause interventions to deviate from their intended effect. More importantly, regulatory intervention itself reshapes the market microstructure, so that resolving a quality problem on one dimension may trigger a deterioration on others. Understanding how each dimension of market quality maps onto regulatory objectives, and the effects of regulatory arbitrage across jurisdictions, is therefore a prerequisite for finding a reasonable balance among efficiency, fairness, and stability.

25.8.1 The correspondence with regulatory objectives

The five dimensions of market quality are not isolated academic metrics; they correspond directly to the core policy objectives of financial regulation. A regulator's intervention in the market is, in essence, a correction and reinforcement of one or more of these dimensions. Figure 25-8 presents, in matrix form, the mapping between the five dimensions of market quality and the main regulatory objectives, and marks the direct strength of each regulatory tool's effect on each dimension along with its potential negative spillovers. From this a key insight emerges: there is almost no regulatory tool that affects only a single dimension. Each intervention, while improving its target dimension, produces positive or negative spillovers onto other dimensions through complex transmission chains—and this multidimensional interconnection is precisely the core difficulty in designing regulatory policy.



Figure 25-8. The mapping between regulatory objectives and market-quality dimensions (a 5×5 qualitative mapping matrix constructed by the author based on Harris [4] and various regulatory frameworks, not direct measurement): the horizontal axis is the five quality dimensions and the vertical axis is the five categories of regulatory objective—competition policy, information regulation, macroprudential regulation, anti-manipulation, and consumer protection; the principal effects lie along the diagonal, and the off-diagonal cells are marked with plus/minus signs for positive and negative spillovers (per Section 25.8.1)

The transaction-cost dimension is directly linked to competition policy. Regulators lower explicit and implicit transaction costs through antitrust measures and by promoting market transparency. In perpetual futures markets, the intensity of competition among market makers and the fee structures of exchanges directly determine investors' transaction costs [17]. Regulatory intervention in transaction costs typically takes the form of requiring exchanges to disclose order-book data and reduce information asymmetry, thereby promoting full competition among market makers. Pushing transaction costs down excessively, however, may squeeze out marginal market makers, concentrating liquidity among a few leading institutions and thereby harming the market's overall resilience.

The price-discovery dimension corresponds to information regulation. A basic objective

of financial regulation is to ensure that market prices reflect all available information accurately and in a timely manner [4]. To this end, regulators typically require issuers to make periodic and ad hoc disclosures and crack down hard on insider trading. In perpetual futures markets, for lack of corporate fundamental information in the traditional sense, price discovery relies more on macroeconomic data, on-chain activity indicators, and the signal transmission of the funding rate. The challenge for regulators is how to define *material nonpublic information* for crypto assets and how to govern the process by which oracle quotes are generated, so as to prevent the price-discovery mechanism from being maliciously distorted.

The resilience dimension is the central concern of systemic-risk regulation. Through macroprudential tools such as capital-adequacy and liquidity-coverage ratios, regulators aim to strengthen financial institutions' ability to withstand external shocks [49]. In perpetual futures markets, for the general absence of a central clearing house's credit guarantee, resilience relies mainly on exchanges' auto-deleveraging mechanisms and insurance funds. The objective of regulatory intervention on this dimension is to prevent a single participant's default from evolving into a systemic liquidation cascade. Requiring exchanges to maintain adequate insurance-fund reserves and to stress-test their risk-management models is a key regulatory means of enhancing market resilience.

The fairness dimension points directly to anti-market-manipulation regulation. A fair market environment requires all participants to enjoy equal opportunity in trade execution and information acquisition. In the crypto derivatives market, latency arbitrage, front-running, and exchanges' VIP privileges for particular market makers constitute the main threats to fairness [50]. Regulators safeguard market fairness by prohibiting specific trading behaviors and requiring exchanges to allocate order flow fairly. Strict fairness regulation, however, may dampen market makers' willingness to provide liquidity, thereby pushing transaction costs up to some degree.

The accessibility dimension is closely tied to consumer protection and access regulation. As analyzed in Section 25.6, regulators decide who may participate in the market through customer-identification and anti-money-laundering (AML) requirements and through assessments of investor suitability [38]. The high leverage of perpetual futures markets makes them enormously attractive to retail investors but also carries a very high risk of liquidation. Regulators face a hard trade-off here: raising the barrier to entry can effectively protect investors who lack the capacity to bear risk, but it also runs counter to the *permissionless* founding ideal of crypto markets and lowers the market's overall accessibility.

25.8.2 The applicability of existing frameworks

Applying traditional securities and derivatives regulatory frameworks directly to perpetual futures markets exposes significant misfits. The traditional framework is built on centralized clearing, clear jurisdictional boundaries, and a standardized product life cycle—assumptions that often do not hold in crypto markets.

The European Union's Markets in Financial Instruments Directive II (MiFID II) is one of the most comprehensive financial-market regulatory frameworks in the world. Under the directive, crypto derivatives may be classified as financial instruments and thus

become subject to complex reporting, transparency, and best-execution requirements [51]. Yet MiFID II was not designed with the features of perpetual futures—no expiry date and a continuous funding rate—in mind. Its pre-trade and post-trade transparency requirements, when applied to the high-frequency, volatile crypto market, may have unintended negative effects on market makers' quoting strategies. MiFID II also relies on clear clearing and settlement infrastructure, which fundamentally conflicts with the logic of many decentralized exchanges that execute automatically through smart contracts. The European Union's Markets in Crypto-Assets Regulation (MiCA), which formally took effect in December 2024, explicitly excludes crypto derivatives from its direct scope, pushing them back under MiFID II for regulation. This classification decision creates a significant regulatory gap: perpetual futures, the most innovative and also the most risky of crypto derivatives, in practice remain subject to a traditional framework that was not designed for them.

The U.S. Regulation National Market System (Reg NMS) aims to enhance the overall efficiency and fairness of the equity market through unified market-data distribution and order-protection rules [52]. Its core order-protection rule requires exchanges to execute orders at the best price available marketwide. Introducing this principle into perpetual futures markets could, in theory, resolve the spread problems caused by liquidity fragmentation. The reality of crypto markets, however, is that exchanges lack a unified clearing system and interoperability protocols, so best execution across platforms faces extremely high technical and fund-transfer costs. Mandating a Reg NMS-like rule would not only be technically difficult to implement but could also trigger severe execution delays that instead harm market quality.

The distinctive features of perpetual futures markets leave large gaps in the traditional regulatory framework. The first is the absence of a regulatory subject under a decentralized architecture. When a trading protocol runs on immutable smart contracts with no clear centralized operating entity, the traditional model of regulation based on institutional licensing becomes difficult to apply. The second is jurisdictional conflict arising from cross-border liquidity. Crypto assets can be transferred worldwide in an instant, so a single jurisdiction's regulatory ban can often only push trading activity toward more lenient jurisdictions rather than eliminate the risk at its root. The third is the new type of risk introduced by the funding-rate mechanism. Traditional regulation lacks tools for assessing this continuous leverage cost and struggles to effectively measure the hidden exploitation it inflicts on investors under extreme conditions.

25.8.3 The design of regulatory tools

Given the limitations of the traditional regulatory framework, regulators need to design more targeted, dynamic tools—tools that anchor directly to specific market-quality indicators rather than relying solely on static compliance checks.

A liquidity-based leverage cap is an important tool for regulating market resilience. Margin requirements in traditional derivatives markets are typically static or adjusted only periodically according to historical volatility. In perpetual futures markets, there is a tight nonlinear relationship between the leverage ratio and liquidity depth [53]. When market depth is insufficient to absorb large-scale liquidation orders, high lever-

age readily triggers a liquidation cascade. Regulators could therefore require exchanges to implement a dynamic leverage cap based on real-time order-book depth: permitting higher leverage in the liquidity-rich normal state to lower transaction costs, and automatically lowering the maximum permitted leverage in a liquidity-exhausted stressed state, thereby enhancing the market's shock-absorbing capacity without entirely sacrificing capital efficiency.

A resilience-based insurance-fund requirement is another key mechanism for guarding against systemic risk. At present, the size of most exchanges' insurance funds is determined by their internal models, without transparency or external oversight. Regulators could introduce a mechanism similar to stress testing in traditional banking, requiring an exchange's insurance fund to be large enough to cover an extreme price shock at a specified confidence level [54]. Such a requirement directly enhances the market's institutional resilience, ensuring that in a black-swan event the profits of winning traders are not subject to a forced clawback because of a counterparty's shortfall.

Market-quality-based regulatory intervention, however, must be wary of unintended consequences. When a jurisdiction imposes strict leverage limits or high capital requirements, the most immediate market reaction is often liquidity migration. Market makers and high-frequency traders quickly move their funds and trading volume to more lightly regulated offshore platforms. This loss of liquidity widens spreads and raises impact costs in the regulated market, ultimately putting it at a disadvantage on the transaction-cost and price-discovery dimensions. The design of regulatory tools must therefore strike a delicate balance between enhancing market safety and preserving market competitiveness.

25.8.4 Regulatory arbitrage and international coordination

The global connectivity of the crypto derivatives market and the fragmentation of national regulatory policies have together given rise to widespread regulatory arbitrage, with a significant effect on global market quality.

Different jurisdictions display sharply different attitudes toward regulating perpetual futures, and those attitudes are themselves evolving rapidly [40]. The U.S. regulatory stance has shifted from aggressive enforcement to cautious openness: between 2022 and 2024, the SEC and CFTC filed suits against numerous crypto exchanges, treated most crypto assets as securities, and cracked down hard on unregistered derivatives platforms; but after a new administration took office in early 2025, the SEC withdrew its suits against some exchanges, and the CFTC began exploring a more inclusive framework for regulating derivatives. This policy shift may significantly improve the accessibility dimension of the U.S. market, but its long-term effect on resilience and fairness remains to be seen. The European Union, for its part, has sought to establish a unified compliance framework through the Markets in Crypto-Assets Regulation, though ambiguity remains in the classification of derivatives. By contrast, some offshore financial centers in Asia and the Caribbean have adopted a more lenient regulatory stance, attracting the registration of many crypto-native exchanges. The direct consequence of these regulatory differences is that global perpetual futures liquidity is highly concentrated on a few offshore platforms.

The effect of regulatory arbitrage on market quality is two-sided. On one hand, it gives participants a channel to escape stringent regulation, letting the perpetual futures market achieve very high accessibility and low transaction costs at low compliance cost. On the other, it severely harms the market's overall fairness and resilience. Offshore platforms often lack rigorous investor-protection mechanisms and transparent market-maker rules, exposing retail investors on these platforms to very high risks of manipulation and systemic collapse. When a crisis strikes an offshore platform, moreover, the absence of a lender of last resort and clear bankruptcy-resolution procedures allows its negative effects to spread rapidly to global markets through the price-transmission mechanism.

International coordination is the fundamental route to resolving regulatory arbitrage, but it faces enormous limits in practice. Although international bodies such as the Financial Stability Board and the International Organization of Securities Commissions have issued high-level recommendations on crypto-asset regulation, a binding, globally unified regulatory framework remains out of reach, owing to fundamental disagreements among countries over financial sovereignty, tolerance for innovation, and philosophies of investor protection [55]. Absent global coordination, a single country's stringent regulation can often only cost its own investors the opportunity to participate in global pricing, without truly eliminating the structural risks inherent in crypto markets. Future regulatory evolution may therefore rely more on bilateral agreements among major economies and on coordinated control over the fiat on- and off-ramp channels for key fiat currencies.

25.9 Technological change and institutional evolution

The evolutionary trajectory of market quality is not linear growth along a single dimension but a complex process in which technical innovation and institutional regulation interact. Technical progress can significantly lower trading friction and improve the efficiency of information processing, but it does not automatically bring an all-around improvement in market quality. On the contrary, the introduction of new technology, while solving old frictions, often creates new structural fragility or fairness imbalances. Assessing the future of market quality turns on discerning whether technical change can reach the fundamental defects of the current market structure and whether institutional evolution can establish a new equilibrium between innovation and stability.

25.9.1 AI market makers

The application of artificial intelligence and machine learning at the microstructure level is reshaping the basic model of liquidity supply. Traditional algorithmic market makers rely on preset parameters and static inventory-control models, whereas AI market makers based on reinforcement learning can dynamically adjust their quoting strategies in response to high-frequency order-flow data [56]. This technical upgrade displays a marked efficiency advantage in normal markets: AI models can predict short-term price direction more accurately, thereby narrowing the bid-ask spread and lowering adverse-selection costs [57]. In inventory management, intelligent algorithms dynamically adjust their position preferences, providing deeper liquidity while maintaining a

lower rate of capital utilization, which directly improves the market's overall trading efficiency [58].

The proliferation of AI market makers, however, also introduces new systemic risk. When most liquidity providers in the market adopt similar machine-learning architectures and training datasets, the problem of strategy homogenization becomes acute. Faced with an extreme shock absent from their training data, homogeneous intelligent algorithms may produce highly correlated order cancellations or aggressive, unidirectional closing of positions [59]. This behavioral pattern drains order-book depth during periods of stress, causing liquidity to dry up in an instant and thereby amplifying local price fluctuations into a systemic liquidity crisis.

The behavioral logic of AI market makers during stress reveals an intrinsic conflict between technical efficiency and market resilience. Although intelligent algorithms can compress transaction costs to extremely low levels in calm periods, their risk-avoidance mechanisms often retreat faster in moments of crisis. This means that a liquidity environment dominated by AI may exhibit a highly fragile character: extremely smooth under normal conditions but highly prone to structural rupture in tail-risk events. Assessing the ultimate effect of AI on market quality must therefore go beyond spread metrics under normal conditions and focus on the willingness to supply liquidity under extreme conditions.

25.9.2 On-chain derivatives

Fully on-chain derivatives protocols represent decentralized finance's (DeFi) latest attempt at complex trading structures. Architectures exemplified by dYdX and Hyperliquid seek to reproduce the execution efficiency of centralized exchanges in a permissionless blockchain environment [43] (this dimension is analyzed systematically in Section 25.6). The core advantage of on-chain derivatives lies in the systematic improvement in market fairness that their transparency brings. On centralized platforms, market makers' privileged agreements and internal liquidation parameters are often hidden from ordinary participants; in an on-chain environment, by contrast, the liquidation logic, funding-rate calculation, and order-matching rules are all executed publicly by smart contracts [8]. This architecture eliminates the space for information privileges and opaque operations, placing all market participants on absolutely equal footing at the level of the rules.

Although fairness improves markedly, on-chain derivatives still face the constraints of physical law in execution efficiency and transaction costs. The consensus mechanism of a blockchain network inevitably introduces a delay in state updates, making on-chain order books less efficient than centralized matching engines at processing high-frequency information [45]. This delay not only widens market makers' adverse-selection risk but also creates room for maximal extractable value. When a trader's intent is exposed before it is packed into a block, arbitrageurs can extract rents through sandwich attacks or front-running, which in effect constitutes a hidden trading tax that erodes the price-discovery quality of on-chain markets.

Whether on-chain derivatives can compete with centralized exchanges on liquidity depth depends on whether their architecture can effectively balance decentralization against

capital efficiency. Current data indicate that although leading on-chain protocols can support the opening of positions on the order of tens of millions of dollars while maintaining reasonable price impact [60], the variability of their liquidity is significantly higher than that of centralized platforms [43]. Under extreme conditions, constrained by the throughput bottleneck of the underlying chain, the on-chain liquidation mechanism may fail because of network congestion, triggering a more severe systemic cascade. The future evolution of on-chain derivatives therefore needs to find a new design space between technical breakthroughs at the consensus level and risk isolation at the application level.

25.9.3 Cross-chain liquidity integration

The proliferation of independent blockchains has made liquidity fragmentation the most severe structural obstacle to crypto-asset market quality. The liquidity of a single asset is dispersed across many independent blockchain networks and protocols, thinning the depth of local markets, enlarging price impact, and thereby pushing up overall transaction costs [61]. This physical isolation impedes the flow of capital to the most efficient trading venues, leaving the price-discovery processes on different chains mutually independent and weakening the pricing efficiency of the global market [62].

The development of cross-chain interoperability protocols offers a technical path out of this predicament. Infrastructure exemplified by LayerZero and Chainlink CCIP establishes standardized cross-chain message-passing mechanisms that let assets and state transfer securely between networks [63]. Such liquidity integration not only lets traders access the globally best quote but also greatly reduces the friction of cross-chain arbitrage. When arbitrageurs can iron out price differences across chains with lower latency and cost, the market's overall informational efficiency improves significantly, and the price-discovery process evolves from a locally fragmented state into a globally coordinated whole [64].

Cross-chain liquidity integration, however, likewise faces serious technical and regulatory challenges. On the technical side, the security model of cross-chain bridges often becomes the weakest link in the entire system; once validator nodes are compromised or a smart contract contains a vulnerability, the integrated liquidity pool faces the risk of devastating asset loss. On the regulatory side, the flow of cross-chain assets blurs jurisdictional boundaries, making anti-money-laundering and compliance monitoring exceptionally difficult. How to build a technical firewall strong enough to withstand systemic contagion, and to meet compliance requirements, while achieving unified liquidity is an obstacle that cross-chain integration must clear on its way to maturity.

25.9.4 Long-term equilibrium

The long-term evolution of market quality can be conceptualized as a dynamic game among competition, innovation, and regulation. Competitive pressure is the most direct force driving improvements in market quality. To compete for market share and liquidity providers, exchanges continually compress fees, optimize matching engines, and offer richer product lines [65]. This commercially driven competition has produced an order-of-magnitude decline in transaction costs over the past 10 years while also

driving continual growth in order-book depth.

Technical innovation, in turn, plays the role of breaking the existing equilibrium. Whether through innovations in the mathematical models of automated market makers or through architectural breakthroughs in cross-chain protocols, innovation provides new tools for resolving old frictions. Yet innovation is often unpredictable: while raising efficiency in one respect, it may harm market fairness or introduce new fragility [66]. Regulatory pressure then becomes the necessary constraint that steers market quality toward the socially optimal direction. Through minimum-capital requirements, mandatory disclosure, and limits on conflicts of interest, the regulatory framework aims to protect participants at an information disadvantage and to guard against the accumulation of systemic risk [67].

Competition, innovation, and regulation together shape the trade-off frontier of market quality. Excessively strict regulation may suppress technical innovation and entrench the competitive advantages of existing oligopolies, impeding further declines in transaction costs; wholly unrestrained competition and innovation, on the other hand, may tilt the market toward an extreme pursuit of efficiency at the expense of fairness and stability. The future of market quality will depend on how these three forces reach a new institutional equilibrium. Only when technical innovation is steered toward resolving fundamental defects, and the regulatory framework can adapt to the features of decentralized architectures, can the perpetual futures market build the deep resilience to withstand extreme shocks while maintaining efficient price discovery.

Figure 25-9 uses a competition-innovation-regulation triangular-game framework to present how these three forces pull differentially on the five dimensions of market quality: the efficiency-oriented dimensions (transaction costs, price discovery) are pulled by the competition-innovation axis and improve rapidly, the institutional-safeguard dimensions (resilience, fairness) strengthen as they draw closer to the regulation vertex, and accessibility occupies a bridging position. The long-term evolution of market quality cannot be dominated by any single force but depends on the equilibrium position of the three forces within a specific institutional environment: when competitive and innovative pressures are too great and regulatory constraints too weak, the efficiency-oriented dimensions are pushed to an extreme and resilience and fairness may deteriorate severely; conversely, excessive regulation may lock the market into a low-efficiency but superficially stable equilibrium.

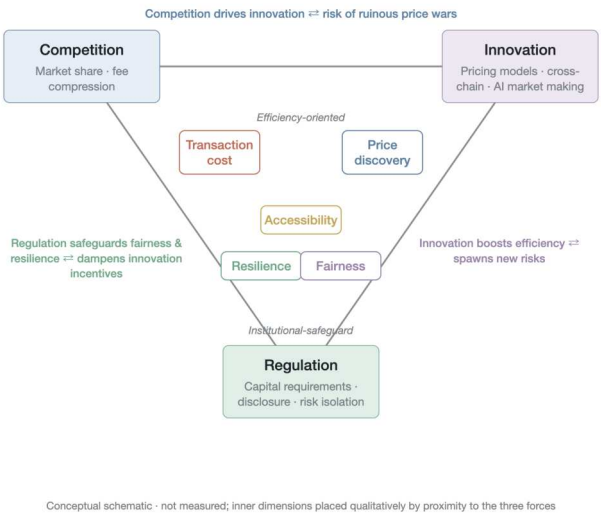


Figure 25-9. The triangular-game framework of market-quality evolution (conceptual illustration, not measured data; an integrated analysis based on market microstructure theory [66][67])

25.10 Chapter summary

Market quality is a multidimensional, dynamically evolving systemic concept. Traditional market microstructure research has tended to focus on single-dimension analysis—using the bid-ask spread as a proxy for liquidity, for example, or the information share as a measure of price discovery [3]. A single-dimension assessment, however, cannot capture the global characteristics of market structure, especially in an innovative derivatives market like perpetual futures that lacks centralized clearing and permits high leverage. The five-dimensional model constructed in this chapter (transaction costs, price discovery, resilience, fairness, and accessibility) provides a comprehensive evaluation framework that reveals the structural characteristics of the perpetual futures market across its different dimensions.

These five dimensions are not mutually independent but harbor intrinsic tensions: improvements in liquidity often come at the cost of declining resilience—the extremely narrow bid-ask spreads of high-frequency market makers lower transaction costs in normal times, but their rapid order cancellations during extreme volatility also amplify the risk of liquidity exhaustion [68]; and between fairness and efficiency, and between accessibility and fairness, there exist the same systematic trade-offs described in Section 25.1.4. These trade-offs show that no market attains the optimum on all dimensions, and constructing a composite index necessarily involves assigning weights to the different dimensions—an assignment that essentially reflects the value preferences of a particular set of market participants or regulators.

Over the past 10 years of evolution, the improvement in perpetual futures market quality has been markedly asymmetric. As analyzed in Section 25.7.4, the dimensions driven

directly by competition have improved fastest—transaction costs have fallen sharply and the leadership position in price discovery has grown ever firmer [69]—whereas dimensions that require institutional change to support them, such as resilience and fairness, have progressed slowly, and market-maker privileges and information asymmetry still leave ordinary traders at a structural disadvantage [4]. This asymmetric evolution shows that the expansion of market size and the deepening of liquidity cannot automatically eliminate structural defects; improving the institutional-safeguard dimensions must rely on deliberate institutional design and external regulatory intervention.

Assessing market quality is not only a diagnosis of the current state of the market but also an integrative framework for market microstructure research. Liquidity sets the baseline for transaction costs, arbitrage efficiency shapes the boundary of price discovery, the characteristics of volatility reflect the level of market resilience, and governance mechanisms fundamentally define the market's fairness and accessibility. The perpetual futures market's lead on the efficiency-oriented dimensions (transaction costs and price discovery) demonstrates the value of its mechanism innovation, but its persistent lag on the institutional-safeguard dimensions (resilience and fairness) exposes the systemic risk of unconstrained development. As the bridging dimension linking efficiency and institutions, accessibility shows a structural advantage in decentralized architectures but still needs continual improvement in user experience and technical barriers. Chapter 26 will approach, from the standpoint of regulatory economics, how to find the optimal balance between innovation and stability.

References

[1] Ruan, Q., & Streltsov, A. (2025). Perpetual futures contracts and cryptocurrency market quality: Insights from emerging markets. Cornell SC Johnson College of Business. https://business.cornell.edu/article/2025/02/perpetual-futures-contracts-and-cryptocurrency/

[2] Stigler, G. (1961). The economics of information. *Journal of Political Economy*, 69(3), 213-225.

[3] O'Hara, M. (1995). *Market Microstructure Theory*. Blackwell Publishers.

[4] Harris, L. (2003). *Trading and Exchanges: Market Microstructure for Practitioners*. Oxford University Press.

[5] Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5(1), 31-56. https://doi.org/10.1016/s1386-4181(01)00024-6

[6] Ruan, Q., & Streltsov, A. (2022). Perpetual futures contracts and cryptocurrency market quality. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4218907

[7] Barbon, A., & Ranaldo, A. (2024). On the quality of cryptocurrency markets: Centralized versus decentralized exchanges. *arXiv preprint arXiv:2112.07386v7*. https://doi.org/10.48550/arXiv.2112.07386

[8] Capponi, A., & Jia, R. (2025). Liquidity provision on blockchain-based decentralized

exchanges. *The Review of Financial Studies*, 38(10), 3040–3085. https://doi.org/10.1093/rfs/hhaf046

[9] Alexander, C., Choi, J., Massie, H. R. A., & Sohn, S. (2020). Price discovery and microstructure in ether spot and derivative markets. *International Review of Financial Analysis*, 71, 101506. https://doi.org/10.2139/ssrn.3511533

[10] Binance. (2026). *USD(S)-M Futures Trading Fee Rate*. Retrieved from https://www.binance.com/en/fee/futureFee

[11] OKX. (2026). *Trading Fee | Fee Rate | Crypto Exchange Fees*. Retrieved from https://www.okx.com/fees

[12] Petersen, M. A., & Fialkowski, D. (1994). Posted versus effective spreads: Good prices or bad quotes? *Journal of Financial Economics*, 35(3), 269-292.

[13] Jain, A. (2025, September 5). *Comparing hidden trading costs: Liquidity analysis across six major platforms* [Commercial review]. Analytics Insight. Retrieved June 24, 2026, from https://www.analyticsinsight.net/

[14] Kasinski, D. (2025). Digital asset funding rates as a distinct risk factor: Evidence from a systematic harvesting approach. *Available at SSRN 5551719*.

[15] Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *Journal of Finance*, 50(4), 1175-1199. https://doi.org/10.1111/j.1540-6261.1995.tb04054.x

[16] Gonzalo, J., & Granger, C. (1995). Estimation of common long-memory components in cointegrated systems. *Journal of Business & Economic Statistics*, 13(1), 27-35.

[17] He, S., Manela, A., Ross, O., & von Wachter, V. (2022). Fundamentals of perpetual futures. *arXiv preprint arXiv:2212.06888*. https://doi.org/10.48550/arXiv.2212.06888

[18] Robertson, K., & Zhang, R. (2025). Price discovery in bitcoin spot and futures markets. *Journal of International Money and Finance*, 159, 103415. https://doi.org/10.1016/j.jimonfin.2025.103415

[19] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315-1335. https://doi.org/10.2307/1913210

[20] Ali, Z. (2025). *Anatomy of the Oct 10-11, 2025 crypto liquidation cascade: Macroeconomic triggers, market microstructure, and systemic risk lessons*. Available at SSRN 5611392.

[21] Large, J. (2007). Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, 10(1), 1-25. https://doi.org/10.1016/j.finmar.2006.09.001

[22] Samani, K. (2020). March 12: The day crypto market structure broke (Part 1). Multicoin Capital.

[23] Chen, Z., He, Z., & Wei, W. (2024). Margin rules and margin trading: Past, present, and implications. *Annual Review of Financial Economics*, 16, 153–177. https://doi.org/10.1146/annurev-financial-110921-102509

[24] Gornall, W., Rinaldi, M., & Xiao, Y. (2025). *Funding payments crisis-proofed bitcoin's perpetual futures* [Working paper]. European Corporate Governance Institute

(ECGI).

[25] Bouri, E., Kamal, E., & Kinateder, H. (2023). FTX collapse and systemic risk spillovers from FTX Token to major cryptocurrencies. *Finance Research Letters*, 56, 104099. https://doi.org/10.1016/j.frl.2023.104099

[26] Corbet, S., & Muñiz, J. A. (2025). *When cryptocurrency collapses: Systemic risk transmission and corporate finance*. Available at SSRN 5904822.

[27] Vidal-Tomás, D., Briola, A., & Aste, T. (2023). FTX's downfall and Binance's consolidation: The fragility of centralised digital finance. *Physica A: Statistical Mechanics and its Applications*, 625, 129044. https://doi.org/10.1016/j.physa.2023.129044

[28] Angel, J. J., & McCabe, D. (2013). Fairness in financial markets: The case of high frequency trading. *Journal of Business Ethics*, 112(4), 585-595.

[29] Lin, Y., Swan, P. L., & Harris, F. H. deB. (2019). Why maker-taker fees improve exchange quality: Theory and natural experimental evidence. *Available at SSRN 3034901*. https://doi.org/10.2139/ssrn.3034901

[30] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1), 71-100. https://doi.org/10.1016/0304-405x(85)90044-3

[31] Malinova, K., & Park, A. (2015). Subsidizing liquidity: The impact of make/take fees on market quality. *The Journal of Finance*, 70(2), 509-536. https://doi.org/10.1111/jofi.12230

[32] Commodity Futures Trading Commission. (2022). *Complaint for Injunctive and Other Equitable Relief and for Civil Monetary Penalties: CFTC v. Samuel Bankman-Fried, FTX Trading Ltd., and Alameda Research LLC*. https://www.cftc.gov/media/8021/enfftxtradingcomplaint122122/download

[33] Securities and Exchange Commission. (2022). *Complaint: SEC v. Samuel Bankman-Fried*. https://www.sec.gov/files/litigation/complaints/2022/comp-pr2022-219.pdf

[34] McMenamin, C., Daza, V., & Fitzi, M. (2022). Fairtradex: A decentralised exchange preventing value extraction. *Proceedings of the 2022 ACM CCS Workshop on Decentralized Finance and Security*, 21-30.

[35] Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., … & Juels, A. (2019). Flash boys 2.0: Frontrunning, transaction reordering, and consensus instability in decentralized exchanges. *arXiv preprint arXiv:1904.05234*. https://doi.org/10.48550/arXiv.1904.05234

[36] Uniswap Labs. (2025). *What is Maximal Extractable Value (MEV)?* https://blog.uniswap.org/maximal-extractable-value-mev

[37] Milionis, J., Moallemi, C. C., Roughgarden, T., & Zhang, A. L. (2022). Automated market making and loss-versus-rebalancing. *arXiv preprint arXiv:2208.06046*. https://doi.org/10.48550/arXiv.2208.06046

[38] Johnson, K. N. (2021). Decentralized finance: Regulating cryptocurrency exchanges. *William & Mary Law Review*, 62(6), 1911. https://doi.org/10.2139/ssrn

.3831439

[39] Demirgüç-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2022). *The Global Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19*. World Bank. https://doi.org/10.1596/978-1-4648-1897-4

[40] Dell'Erba, M., & Vianelli, A. (2025). Crypto-derivatives regulation(s). *Cornell International Law Journal* (forthcoming). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5895082 https://doi.org/10.2139/ssrn.5895082

[41] Chainalysis. (2025). *The 2025 Geography of Cryptocurrency Report*. Retrieved June 24, 2026, from https://go.chainalysis.com/2025-geography-of-cryptocurrency-report.html

[42] Hasbrouck, J., Rivera, T. J., & Saleh, F. (2024). An economic model of a decentralized exchange with concentrated liquidity. *Management Science*. Advance online publication. https://doi.org/10.1287/mnsc.2024.04510

[43] Ved, T. (2025). Inside dYdX: On-chain perpetual futures. *Coin Metrics State of the Network*. https://coinmetrics.substack.com/p/state-of-the-network-issue-318

[44] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 135(2), 293-319.

[45] Chainlink. (2026). DeFi derivatives: The future of onchain trading. https://chain.link/article/defi-derivatives

[46] Madhavan, A. (2000). Market microstructure: A survey. *Journal of Financial Markets*, 3(3), 205-258. https://doi.org/10.1016/s1386-4181(00)00007-0

[47] Bongaerts, D., & Van Achter, M. (2016). *High-frequency trading and market stability* [Working paper]. SSRN. https://doi.org/10.2139/ssrn.2698702

[48] International Monetary Fund. (2015). *Market liquidity – Resilient or fleeting?* Global Financial Stability Report.

[49] Financial Stability Board. (2025). *Recommendations to address financial stability risks created by leverage in nonbank financial intermediation*. https://www.fsb.org/2025/07/

[50] Easley, D., O'Hara, M., & Yang, L. (2016). Differential access to price information in financial markets. *Journal of Financial and Quantitative Analysis*, 51(4), 1071-1110. https://doi.org/10.1017/s0022109016000491

[51] Financial Conduct Authority. (2018). *Cryptocurrency derivatives statement*. https://www.fca.org.uk/news/statements/cryptocurrency-derivatives

[52] Securities and Exchange Commission. (2005). *Regulation NMS*. Release No. 34-51808.

[53] International Swaps and Derivatives Association. (2025). *Policy ideas to enhance market liquidity and risk management*. https://www.isda.org/

[54] Aramonte, S., Huang, W., & Schrimpf, A. (2021). DeFi risks and the decentralisation illusion. *BIS Quarterly Review*, December 2021.

[55] Draganidis, S. (2023). Jurisdictional arbitrage: Combatting an inevitable by-

product of cryptoasset regulation. *Journal of Financial Regulation and Compliance*, 31(2), 170-185.

[56] Kearns, M., & Nevmyvaka, Y. (2013). Machine learning for market microstructure and high frequency trading. *High Frequency Trading: New Realities for Traders, Markets, and Regulators*, 1-22.

[57] Chakrabarty, B., & Pascual, R. (2023). Stock liquidity and algorithmic market making during the COVID-19 crisis. *Journal of Banking & Finance*, 147, 106415. https://doi.org/10.1016/j.jbankfin.2022.106415

[58] DWF Labs. (2025). The impact of AI on crypto trading and market making. https://www.dwf-labs.com/news/the-impact-of-ai-on-crypto-trading-and-market-making

[59] Bogoev, D. (2025). Liquidity pricing and crisis: New metrics from traditional and blockchain-based markets. *Durham University Theses*.

[60] GMX. (2026). GMX | Decentralized perpetual exchange. https://gmx.io/

[61] Sim, J. Y., Chun, H., & Park, J. (2025). Toward unified liquidity in cryptocurrency markets: A comparative survey and architectural perspective. *International Journal of Advanced Smart Convergence*, 14(2), 33–45. https://doi.org/10.7236/IJASC.2025.14.2.33

[62] Johnson, W. C., & Scharnowski, S. (2025). Price discovery through wrapped tokens. *Economics Letters*, 257, 112703. https://doi.org/10.1016/j.econlet.2025.112703

[63] Chainlink. (2026). Multi-chain liquidity integration: Unifying fragmented assets. https://chain.link/article/multi-chain-liquidity-integration

[64] Lao, X., Chen, S., & Wang, Q. (2025). FLUXLAYER: High-performance design for cross-chain fragmented liquidity. *IEEE International Conference on Blockchain and Cryptocurrency*.

[65] OECD. (2023). Competition and innovation, Part I: A theoretical perspective. *OECD Competition Committee*.

[66] Auer, R., Farag, M., Lewrick, U., Orazem, L., & Zoss, M. (2022). Banking in the shadow of Bitcoin? The institutional adoption of cryptocurrencies. *BIS Working Papers No. 1013*.

[67] Rana, H. A., Audi, M., & Ali, A. (2025). Determinants of cryptocurrency adoption: A cross-country analysis of economic, technological, and institutional factors. *Journal of Social Signs Review*, 3(8), 58–76.

[68] Alexander, A. (2025). The impact of market-making incentives on crypto liquidity & price stability. *Available at SSRN 5143194*. https://doi.org/10.2139/ssrn.5143194

[69] Griffith, T., & Clancey-Shang, D. (2023). Cryptocurrency regulation and market quality. *Journal of International Financial Markets, Institutions and Money*, 84, 101744. https://doi.org/10.1016/j.intfin.2023.101744

The Economics of Regulation Between Innovation and Stability

In the first quarter of 2024, the average weekday trading volume of Bitcoin perpetual futures (measured as a 30-day rolling average) reached $57.7 billion [1]. That figure reflects perpetual futures' roughly 90% share of the crypto-derivatives market by open interest [2], and it has made them a central concern for financial regulators worldwide. Faced with this large-scale market—one that runs around the clock, spans national borders, and has no central clearinghouse—regulators confront a severe classification dilemma. Perpetual futures represent no ownership claim over any entity and therefore fail to satisfy the core "investment contract" element of the Howey test [3], which makes them difficult to classify simply as securities. At the same time, they have no expiration date and no physical-delivery mechanism, which places them outside the definitional framework of traditional commodity futures [4] and likewise challenges the jurisdiction of the Commodity Futures Trading Commission (CFTC).

The difficulty of classification is merely a surface symptom; the deeper predicament lies in the failure of traditional regulatory tools. Historical experience has shown that a simple, blanket ban cannot eliminate market demand and only invites regulatory arbitrage. Indeed, capital flight was among the regulatory motives well before any ban: Chainalysis estimates that from mid-2019 to mid-2020, more than $50 billion in crypto-assets (of which roughly $18 billion was USDT) moved from East Asian addresses to destinations outside the region, forming an important part of the backdrop to China's comprehensive 2021 ban [5]. After the ban took effect, trading activity migrated rapidly to offshore exchanges, while domestic traders continued to participate through technical workarounds. This outcome indicates that the only substantive consequence of prohibition was to strip regulators of all visibility into the market and any capacity to intervene in it, all without reducing systemic risk. Yet the complete absence of regulation carries equally serious consequences. The November 2022 collapse of the FTX exchange exposed a shortfall of roughly $8 billion in customer funds [6], and millions of users suffered substantial asset losses. This episode demonstrates just how severe information asymmetry and agency moral hazard can become in a "free market" that lacks mandatory asset segregation and external audit.

Regulators therefore can neither force perpetual futures into an ill-fitting traditional framework, nor ignore their existence, nor retreat to a posture of laissez-faire. This predicament raises the central question of the chapter: in the specific context of the

perpetual futures market, how should regulation intervene so as to repair the market's defensive deficiencies without destroying its innovative advantages in price discovery and accessibility? Answering this question requires moving beyond a moral or political stance on whether regulation is "good" or "bad" and adopting instead the cost-benefit analytical framework of economics. Regulation is, in essence, a tool for intervening in market failure, and every specific regulatory policy—whether capping leverage, mandating disclosure, or requiring capital segregation—inevitably carries both benefits and costs.

To assess the net effect of these policy tools systematically, this chapter proposes and applies three interrelated analytical frameworks. The five-dimensional regulatory impact matrix evaluates, one by one, the specific effect of each regulatory tool on five dimensions—transaction costs, price discovery, market resilience, fairness, and accessibility—providing a microfoundation for policy comparison. Building on this, the regulatory "impossible trinity" reveals that under a globalized and decentralized technical architecture, no regulatory regime can simultaneously maximize global consistency, innovation-friendliness, and consumer protection, thereby explaining the structural roots of today's fragmented global regulation. That fragmentation, in turn, motivates a paradigm contrast between functional and entity-based regulation, which argues why traditional institution-type regulation is markedly constrained in the face of cross-boundary innovation and prone to regulatory vacuums, and why regulation organized around function and risk characteristics is the more adaptive alternative.

Building on these frameworks, the chapter proceeds to a detailed analysis of the economics of regulation. It first defines four categories of market failure in the perpetual futures market—externalities, information asymmetry, and others—to delineate the legitimate boundaries of regulatory intervention, and it constructs a net-benefit evaluation model that incorporates compliance burdens and innovation suppression. It then sets out an inventory of regulatory targets for perpetual futures, prioritizing systemic risk, market manipulation, consumer protection, and market integrity according to the severity of each risk and the market's capacity for self-repair. On this basis, it dissects the five-dimensional impact of specific regulatory tools such as leverage limits and market-maker obligations, and it examines the economic consequences of the global contest over jurisdiction. Finally, the chapter extends to the regulatory boundaries of decentralized exchanges, exploring the possibility and the technical limits of compliance-by-code. After reading this chapter, readers will be able to step outside the black-and-white debate over regulation, use structured economic tools to assess the true cost of any regulatory proposal, and clearly delineate which market deficiencies can be repaired through legal rules and which must be left to the decentralized engineering discussed in later chapters.

26.1 The legitimacy of regulation

The legitimacy of regulatory intervention rests on market failure: when the market mechanism cannot on its own reach a Pareto-optimal allocation, external intervention may improve aggregate social welfare. In the context of the crypto-derivatives market, not every market deficiency constitutes a failure that warrants regulatory interven-

tion. Some deficiencies can repair themselves through competition among protocols or through underlying technological innovation, whereas excessive regulatory intervention may not only impose high compliance costs but also trigger innovation suppression and regulatory arbitrage. Assessing the soundness of a regulatory policy is therefore, in essence, a complex cost-benefit analysis: intervention is legitimate in the economic sense only when the marginal benefit of repairing a market failure exceeds the sum of the regulatory costs.

26.1.1 Four categories of market failure

In traditional financial theory, market failure is usually grouped into four basic forms: externalities, information asymmetry, undersupply of public goods, and market power [7]. Within the distinctive microstructure of perpetual futures, these four failures display characteristics quite different from those of traditional assets.

Externalities are the most damaging form of failure in the perpetual futures market, typically manifested as the systemic risk arising from liquidation cascades. When a single trader's over-leveraged position is forcibly liquidated, the liquidation engine's dumping of assets onto the market depresses the price, which in turn triggers the liquidation of additional highly leveraged positions [8]. This positive-feedback loop harms not only the liquidated trader but also imposes the negative externality of evaporating liquidity on all market participants. As illustrated by the October 2025 liquidation cascade detailed in Section 26.2.1, this endogenous fragility constitutes the core rationale for regulatory intervention [9].

Information asymmetry is especially pronounced between centralized exchanges and retail traders. Exchanges hold complete data on order flow, liquidation parameters, and asset reserves, whereas ordinary traders are often unable to assess counterparty risk accurately. The collapse of FTX laid bare the extreme consequences of this asymmetry: management's misappropriation of customer assets went undetected by the market for a long period, ultimately producing billions of dollars in customer losses [10]. This collapse of trust, driven by informational opacity, not only destroyed a single platform but also triggered cross-market financial contagion [11].

The undersupply of public goods manifests in the perpetual futures market as a tragedy of the commons in liquidity under extreme conditions. Liquidity is, in essence, a public resource with positive externalities. In calm markets, market makers earn the bid-ask spread by supplying liquidity; but in periods of high volatility or crisis, market makers, out of their own risk-management considerations, often choose to withdraw their quotes. This individually rational risk-averse behavior produces, at the aggregate level, a collectively irrational and sharp decline in liquidity, further exacerbating price volatility and liquidation risk [12].

Market power arises from the monopolistic position of the leading exchanges. In a market where liquidity exhibits strong network effects, trading volume tends to concentrate on a few large platforms. This market concentration allows the leading exchanges to exercise unilateral pricing power when setting market-maker agreements, fee schedules, and listing standards. Moreover, some exchanges combine the dual roles of matching venue and proprietary market maker, a conflict of interest that, absent external

oversight, can readily evolve into an implicit shifting of costs onto ordinary traders.

26.1.2 The boundary of self-repair

Identifying a market failure does not mean that regulation must intervene immediately. In crypto markets, many phenomena regarded as defects are in fact improving on their own through market mechanisms. The central criterion for judging whether regulation is necessary is whether the defect has the potential to correct itself naturally through competition or technological evolution.

Competition constitutes a highly effective self-correcting mechanism. Consider transaction costs: as decentralized derivatives protocols have proliferated and competition among centralized exchanges has intensified, the all-in trading cost of perpetual futures—including explicit fees and implicit slippage—has trended steadily downward. This competition-driven efficiency gain sharply reduces the need for direct price controls on fee rates or market-maker spreads. When the market can maximize welfare on its own through competition, forcible regulatory intervention may instead distort price signals.

Technological innovation likewise performs a regulation-substituting function. Early liquidation mechanisms often used a crude, whole-position approach that readily produced severe price shocks. To attract and retain users, however, many exchanges voluntarily introduced gradual liquidation algorithms and more refined risk engines. This technologically driven improvement in resilience demonstrates that, in certain domains, code and protocol rules can constrain high-risk behavior more effectively than traditional administrative directives. In addition, the maturation of cross-market arbitrage has caused many early pricing anomalies to fade, as arbitrageurs, in pursuing profit maximization, in effect perform the function of correcting price deviations.

Regulatory resources should therefore concentrate on the structural failures that the market cannot resolve on its own. The enforcement of asset segregation and the verification of reserve transparency, for example, cannot be achieved through exchanges' self-disciplinary pledges alone; they require external audit and mandatory legal constraints [13]. Conversely, for the micro-frictions that are naturally easing as competition intensifies or technology iterates, regulatory restraint is often the superior strategy.

26.1.3 The costs of regulation

Every form of regulatory intervention carries non-negligible costs, and in the crypto-derivatives market these costs are exceedingly complex.

Direct compliance costs are the most visible price of regulation. To satisfy the anti-money laundering (AML), know-your-customer (KYC), and periodic audit requirements of different jurisdictions, exchanges must maintain large legal and compliance teams. Research indicates that the average annual federal regulatory compliance cost per employee across all U.S. industries is approximately $12,800, rising to roughly $14,700 for small firms with fewer than 50 employees [14]. In crypto markets, these high operating costs are ultimately passed on to traders through higher fees or wider bid-ask spreads, directly eroding market accessibility and operational efficiency.

Innovation-suppression costs are a hidden but far-reaching negative effect of regulation. Financial innovation is often born in the gray areas of existing regulatory frameworks. Perpetual futures were themselves invented to circumvent the liquidity fragmentation caused by traditional futures delivery. Had this product been brought under a strict traditional-derivatives regime in its infancy, its distinctive funding-rate mechanism might never have been validated and developed. Surveys of regulation-intensive industries such as finance and utilities find that compliance costs can exceed 2.5% of a firm's total payroll, while the aggregate cost of incremental social regulation is estimated at roughly 5% of U.S. GDP [15]. An excessive compliance and regulatory burden can crowd out the resources firms devote to experimentation and R&D, although the net effect of regulation on innovation remains contested in the literature.

Regulatory-arbitrage costs are especially salient in a globalized, decentralized network environment. When a given country or region imposes stringent leverage limits or entry standards, capital and trading activity migrate rapidly to more lightly regulated offshore platforms. When Japan cut the leverage cap on crypto-asset derivatives sharply from 25x to 2x in 2020, for example, it nominally protected domestic investors but in practice drove large trading volumes to overseas exchanges beyond its jurisdiction (Section 26.7.1 documents this trade-off with policy data). Such regulatory arbitrage not only fails to reduce global systemic risk at its root but also strips the strictly regulating jurisdiction of any real pricing power over, or supervisory capacity within, the market.

Moreover, transplanting rules designed for the traditional financial system directly onto crypto architecture generates enormous friction costs. The non-custodial nature of decentralized exchanges and the logic of automatic smart-contract execution conflict fundamentally with the traditional intermediary-based regulatory paradigm. Forcing decentralized protocols to discharge the obligations of a traditional broker is not only technically difficult but may also fundamentally undermine the permissionless nature of decentralized finance.

26.1.4 The net-benefit analysis framework

Building on the preceding analysis, regulatory policy must be formulated on a rigorous cost-benefit framework [16]. The net benefit of a regulatory measure can be formalized as the increase in social welfare from repairing a market failure, minus direct compliance costs, innovation-suppression costs, and the efficiency losses induced by regulatory arbitrage.

Figure 26-1 characterizes the nonlinear relationship between net benefit and regulatory intensity along two dimensions, benefit and cost: the intersection of the two curves marks the critical threshold at which net benefit turns from positive to negative, while the point of maximum vertical distance between them corresponds to the optimal regulatory intensity that maximizes net social welfare.

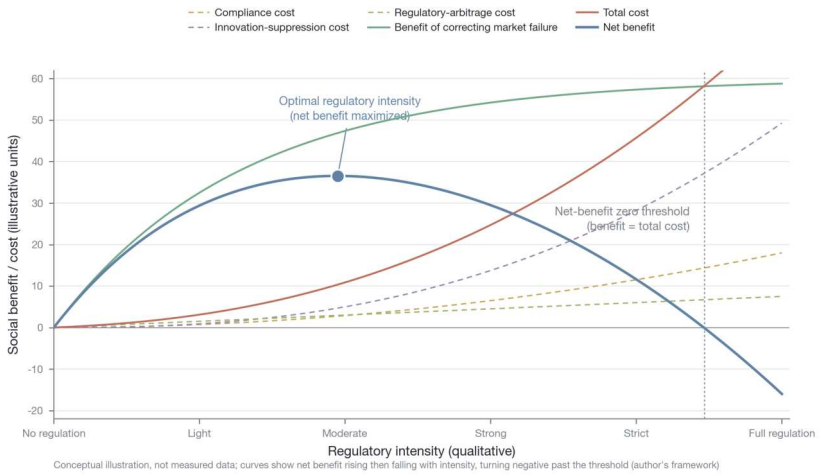


Figure 26-1. The net-benefit analysis framework for regulation

The shapes of the two curves follow from basic regulatory economics. The benefit of repairing market failure rises with regulatory intensity but exhibits pronounced diminishing marginal returns: basic early-stage regulation—such as mandatory disclosure and asset segregation—can sharply reduce systemic risk at relatively low cost. Once regulatory intensity crosses the critical point, however, further intervention such as extreme leverage limits or product bans yields negligible marginal safety benefits, even as compliance and innovation-suppression costs accelerate markedly.

Regulatory-arbitrage costs must be understood along two dimensions. Along the single-jurisdiction regulatory intensity depicted on the horizontal axis of Figure 26-1, arbitrage costs rise monotonically with intensity: the more stringently one jurisdiction regulates, and the larger the rule gap relative to other jurisdictions, the stronger the incentive for capital flight and risk transfer; hence this component is increasing in the figure. Along the dimension of global regulatory coordination, by contrast, arbitrage costs follow an inverted-U shape: under the extreme assumptions of a complete absence of regulation or a perfectly uniform global regime, there is no exploitable rule differential across jurisdictions and thus no room for arbitrage; but in the realistic state where jurisdictions each impose moderate, heterogeneous regulation, the rule differentials across jurisdictions maximize the arbitrage incentive, driving capital flight and risk transfer. This inverted-U shape is a theoretical inference: its peak depends on the degree of rule differential across jurisdictions, and it currently lacks direct empirical calibration.

Regulatory intervention is justified only when the total-benefit curve lies above the total-cost curve—that is, when net benefit is positive. The optimum in the figure indicates the regulatory intensity that maximizes net social welfare. This framework rules out two extreme policy positions: it rejects the complete laissez-faire that ignores the risks of liquidation cascades and asset misappropriation, and it equally rejects the over-regulation that ignores the value of innovation and the native characteristics of crypto. In the regulatory practice of perpetual futures, finding this optimal equilibrium requires policymakers to understand the underlying technical mechanisms thoroughly

and to weigh investor protection against the maintenance of market vitality with care.

26.2 Regulatory targets and prioritization

The scarcity of regulatory resources demands that intervention be selective. Not every deficiency in the perpetual futures market requires regulatory intervention; the core task is to identify the areas where the marginal benefit of intervention is highest and to prioritize among them. In evaluating each market failure, the market's own capacity for repair must be taken into account, because competitive mechanisms and technological evolution can often resolve some problems endogenously. Regulatory intervention is fully justified in the economic sense only when the market mechanism has entirely failed and the negative externality or information asymmetry causes a significant loss of social welfare. This section systematically reviews four potential regulatory targets in the perpetual futures market—systemic risk, market manipulation, consumer protection, and market integrity—and, based on the severity of the market failure, the market's capacity for self-repair, and the net benefit of intervention in each area, provides an operational framework for prioritization.

26.2.1 Systemic risk

Systemic risk is the most prominent source of negative externalities in the perpetual futures market, and its core mechanism is precisely the positive-feedback loop between the forced liquidation of highly leveraged positions and the evaporation of liquidity described in Section 26.1.1. This self-reinforcing liquidation cascade harms not only overleveraged users but, through liquidity evaporation and price contagion, strikes every participant in the ecosystem—including spot holders who never used leverage. The liquidation cascade of October 10–11, 2025 is an instructive case: within 36 hours, roughly $19 billion of open interest was forcibly liquidated. The event was triggered mainly by a macro-policy shock, but the structural accumulation of high leverage was the fundamental reason this shock escalated (base estimates, cross-asset contagion during the episode was roughly $20+) approached 0.90, indicating that the effect of the systemic shock was highly persistent [9].

Individual exchanges lack the economic incentive to unilaterally limit systemic risk—a classic prisoner's dilemma. In a highly competitive environment, exchanges tend to offer very high leverage to attract traders seeking high-risk returns. If one platform unilaterally lowers its leverage cap or tightens margin requirements, its market share will quickly flow to more lightly regulated competitors, whether centralized exchanges in other jurisdictions or permissionless decentralized protocols. This leverage arms race causes systemic risk to accumulate across the entire market, while no individual platform bears the full cost of the systemwide damage caused by a liquidation cascade: the negative externality of excessive leverage is externalized to the whole market, whereas its private benefits (trading fees and funding-rate revenue) are internalized by the exchange. This mismatch between costs and benefits is a clear manifestation of market failure and the fundamental reason regulatory intervention is justified.

Regulatory intervention against systemic risk carries a fairly high marginal benefit, but

the challenges of implementation are not to be underestimated. By imposing a market-wide uniform leverage cap, dynamic margin requirements, and standardized liquidation mechanisms, regulation can effectively interrupt the transmission path of the positive-feedback loop and markedly improve the market's overall resilience. Such intervention, however, also faces complex trade-offs. Strict leverage limits inevitably reduce capital efficiency, raise traders' implicit costs, and may drive high-leverage demand to unregulated offshore markets or decentralized platforms, producing a regulatory-arbitrage effect. The setting of regulatory parameters must therefore strike a delicate balance between reducing systemic risk and maintaining market liquidity, and the effectiveness of regulatory action by a single jurisdiction will be greatly diminished without international coordination.

26.2.2 Market manipulation

Market manipulation systematically undermines market fairness by distorting the price-discovery mechanism and inflating the appearance of liquidity. In the perpetual futures market, manipulation takes more varied forms than in traditional financial markets, and because of the regulatory vacuum, its execution cost is extremely low. One typical pattern is cross-market spot-perpetual manipulation: the manipulator buys the spot asset heavily during periods of concentrated liquidity, pushing up the index price on which the mark price depends and thereby triggering the forced liquidation of a large number of short positions in the contract market; the manipulator then rapidly sells in the opposite direction once the liquidations are complete, profiting from the artificially induced price swing. Empirical research shows that this manipulation pattern leaves identifiable statistical traces in data from major exchanges such as Binance: the price crashes that follow manipulative spikes produce significantly lower subsequent returns, and the position changes of top traders are highly correlated with price movements [17]. This strategy is especially frequent in the less liquid altcoin perpetual futures market, because a thin order book means the manipulator can generate a significant price deviation with relatively little capital.

Wash trading is another pervasive and far-reaching form of manipulation, designed to manufacture a false appearance of market activity through self-trading or offsetting trades between affiliated accounts. A systematic empirical analysis of 29 cryptocurrency exchanges found that wash trading accounts, on average, for more than 70% of the reported volume on unregulated exchanges, with such fictitious volume reaching trillions of dollars annually [18]. Exchanges have strong economic incentives to engage in or tolerate wash trading: inflated volume raises their ranking on third-party data platforms, attracts genuine liquidity providers and institutional clients, and provides stronger bargaining leverage in listing negotiations. The harm of wash trading lies not only in the direct price distortion but, more importantly, in its systematic pollution of the market's information environment. When investors cannot distinguish genuine liquidity from fictitious liquidity, their judgments about market depth and liquidity risk become systematically biased, leading them to make suboptimal investment decisions. Wash trading also provides cover for implicit collusion between market makers and exchanges, making conflicts of interest difficult for outside observers to detect.

Regulators play a critical role in combating market manipulation, but the effectiveness of enforcement depends heavily on their ability to obtain trading data. Although the market's endogenous arbitrage mechanism can dampen price manipulation to some degree, it is often powerless against complex cross-market manipulation and systematic wash trading, because the arbitrageurs themselves may become the injured party of a manipulation strategy. By mandating strict trade surveillance, suspicious-activity reporting regimes, and clear definitions and severe penalties for manipulative conduct, regulation can significantly raise the cost of violations and thereby reduce the frequency of manipulation at the margin. Yet the cross-border nature of crypto markets means that enforcement by a single jurisdiction often runs into the limits of its jurisdictional boundary, which is an important reason the net benefit of regulating market manipulation is partly diluted.

26.2.3 Consumer protection

Consumer protection is fundamentally about correcting the information asymmetry among market participants, and in particular about shielding inexperienced retail traders from the disproportionate harm of complex financial derivatives. The mechanism design of perpetual futures contains several elements that are highly opaque to retail participants: the dynamic adjustment of the funding rate, the calculation of the mark price from a multi-exchange index price, the nonlinear liquidation function, and the operating logic of the insurance fund. Most retail traders do not fully grasp the destructive power of these mechanisms under extreme conditions, especially how the multiplier effect of high leverage and volatility can drive initial margin to zero in a short time. The European Securities and Markets Authority's (ESMA) analysis of the contracts-for-differences (CFDs) market shows that 74% to 89% of retail investor accounts lose money in leveraged trading, with average per-investor losses ranging from €1,600 to €29,000 [19]. Although CFDs and perpetual futures differ structurally—in particular, the B-book market-making structure of CFDs differs from the order-book matching model of perpetual futures, so this figure can be treated as an upper-bound estimate of the retail loss rate—the two are highly similar in their leverage effect and their pattern of retail losses, and this figure provides an important reference benchmark for assessing the consumer-protection needs of the perpetual futures market.

In the regulatory design of consumer protection, there is a fundamental clash of values between libertarianism and paternalism, and this divergence has no objective solution at the economic level. The libertarian view holds that adult investors have the right to bear risk on the basis of their own judgment, including the right to sustain financial losses, and that the regulator's responsibility is limited to ensuring adequate disclosure rather than substituting for the investor's decisions. The economic rationale for this position is that mandatory protective measures generate a paternalism cost: they deprive investors of tools for reasonable risk management with derivatives and may push retail participants toward more lightly regulated offshore platforms, thereby reducing, in substance, the protection they receive. The paternalist view argues that, given retail participants' inherent limitations in cognitive capacity, risk tolerance, and access to information, market outcomes do not truly reflect their preferences, and that mandatory measures must therefore restrict their participation in high-risk trading even at the cost

of their formal freedom to invest. This chapter takes no position, but it notes that this divergence directly shapes the choice of regulatory tools—from mere risk warnings to strict accredited-investor entry standards and tiered leverage regimes.

The behavioral-economics framework of the "nudge" offers a third path beyond this binary opposition [20]. The core idea of a nudge is to guide decision-makers toward better outcomes by altering the choice architecture without prohibiting any option. Applied to the perpetual futures market, regulators can design a set of non-coercive interventions. For example, they can set the default leverage on the trading interface to a low level (such as 5x); users can still raise it manually, but the default effect significantly lowers a novice trader's initial risk exposure. As another example, before a user submits a high-leverage order, they can be required to complete a visual simulation of liquidation risk, so that they intuitively understand the probability distribution of having margin wiped out under extreme conditions. As a further example, a pre-trade cooling-off period can be introduced, imposing a brief delay on newly opened high-leverage positions during periods of extreme volatility to curb emotionally driven impulsive trades. These measures respect the trader's right to autonomous choice while using behavioral intervention to reduce the systematic losses caused by cognitive biases, thereby opening a pragmatic middle ground between libertarianism and paternalism.

The economic trade-offs of implementing consumer protection are especially pronounced, which is the main reason it ranks last in the prioritization. Strengthening disclosure is the lowest-cost intervention and can effectively reduce information asymmetry without directly interfering in market operation, but its actual effect is severely limited by retail participants' attention and comprehension. More aggressive measures—such as sharply limiting retail leverage or outright banning retail participation in certain products—can directly reduce retail financial losses, but they also sharply reduce market accessibility and may drive demand toward regulatory vacuums. Consumer-protection regulation therefore needs to be carefully calibrated between preventing exploitation and preserving market openness, and its net benefit depends to a large extent on the regulator's value preferences rather than on pure economic calculation.

26.2.4 Market integrity

Market integrity is the foundation on which any financial system exists, and its core requirement is the safety of customer assets and the transparency of intermediary operations. If an exchange can arbitrarily misappropriate customer assets without oversight, then any rule design at the trading level and any optimization of market microstructure become meaningless, because users' fundamental interests have already been violated at the most basic level. The collapse of FTX is the definitive cautionary case for the cryptocurrency market. The CFTC's complaint states that the platform entirely failed to implement basic corporate controls, failed to segregate customer assets from its own funds and from the funds of affiliated entities, and failed to keep proper records of intercompany fund flows [21]. The CFTC's 2022 complaint found customer deposit losses at FTX exceeding $8 billion; in August 2024, the Commission further obtained a court-approved monetary-relief judgment of approximately $12.7 billion (about $8.7

billion in restitution to customers and victims and about $4 billion in disgorgement and penalties) [21]. This episode reveals a structural loophole: in the absence of mandatory asset-segregation requirements, the moral hazard of centralized exchanges is systemic rather than incidental.

The proof-of-reserves mechanism that the industry has explored on its own attempts to prove an exchange's asset solvency through cryptographic means, but it has fundamental limitations as a safeguard of market integrity. Proof-of-reserves systems focus mainly on verifying the existence of on-chain assets, yet they often cannot reveal, fully or in real time, an exchange's hidden liabilities and complex off-balance-sheet leverage [22]. Specifically, proof of reserves can show that an exchange controls certain wallet addresses, but it cannot prove that those assets have not been secretly pledged to third-party lenders, nor that the exchange holds no hidden short positions established through derivatives. This incomplete transparency not only fails to prevent the covert misappropriation of assets but may even mislead investors by providing a false sense of security, causing them to form overly optimistic judgments about the exchange's actual solvency. Relying solely on technical means and industry self-discipline therefore cannot fully resolve the moral hazard of centralized intermediaries; it must be supplemented by legally binding external oversight.

Mandatory asset segregation and external audit are the core means of safeguarding market integrity. Regulators must require perpetual futures exchanges to maintain minimum capital to absorb operating losses, to enforce strict physical and legal segregation between customer assets and proprietary assets, and to undergo regular, comprehensive audits by independent third parties, with the audit scope extending to off-chain liabilities and related-party transactions. The compliance cost of this kind of regulatory intervention is relatively manageable—consisting mainly of audit fees and the cost of tied-up capital—but its benefit in averting a multibillion-dollar disaster is enormous. More importantly, safeguarding market integrity has significant positive externalities: when investors have full confidence in an exchange's solvency, the liquidity and participation of the entire market improve.

26.2.5 Prioritization

Based on the net-benefit analysis framework, we can construct a systematic prioritization of the four regulatory areas above. This ranking depends not only on the severity of the market failure but must also weigh the market's capacity for self-repair and the expected net benefit of regulatory intervention. By structuring these dimensions, we can clearly identify the optimal allocation of regulatory resources and provide a theoretical basis for regulators' decisions under resource constraints.

Table 26-1 provides a systematic prioritization of the four regulatory areas along three dimensions: the severity of the market failure, the market's capacity for self-repair, and the net benefit of regulation. Market integrity is placed in the first priority. Its extremely high failure severity has already been demonstrated by the FTX episode discussed above (Section 26.2.4), and the market is entirely unable to guarantee the safe segregation of assets on its own in the absence of external compulsion. Because the compliance cost is relatively low and the intervention effectively prevents catastrophic

outcomes, its net regulatory benefit is extremely high. Moreover, safeguarding market integrity has a foundational status: the other regulatory objectives are meaningful only once the safety of customer assets is assured.

Systemic risk follows in the second priority. Although its severity is slightly lower than that of a market-integrity failure, the market lacks any mechanism to reduce risk on its own because of the prisoner's dilemma created by the leverage arms race. Intervention against systemic risk can significantly improve market resilience and prevent the negative externality of a liquidation cascade from harming innocent participants, so its net benefit is likewise considerable. However, the effectiveness of systemic-risk regulation depends heavily on international coordination, and unilateral action may produce a regulatory-arbitrage effect, which to some extent dilutes its net benefit.

Market manipulation ranks third. Although price distortion and wash trading severely undermine fairness, the presence of arbitrage mechanisms can constrain the space for manipulation to some degree, and as the market matures, the profit margin of some manipulative behavior is narrowing naturally. Regulatory intervention in this area is necessary, but its net benefit is diluted by high enforcement costs owing to the complexity of cross-market and cross-jurisdictional enforcement. Consumer protection is placed fourth. Although retail participants face systematic losses in leveraged trading, the information-asymmetry problem has some room for self-improvement as market education spreads and information transparency increases. More importantly, consumer-protection measures often stand in direct negative trade-off with market accessibility, and their net benefit depends to a large extent on the regulator's value preferences rather than on pure economic calculation.

Regulatory area	Market-failure severity	Capacity for self-repair	Net regulatory benefit	Priority
Market integrity (asset segregation)	Very high (FTX: ~$8 billion in customer losses)	Very low (moral hazard cannot self-eliminate)	Very high	First
Systemic risk (leverage / liquidation)	High (cascade losses can reach tens of billions of dollars)	Low (arms race creates a prisoner's dilemma)	High (subject to international coordination)	Second
Market manipulation (wash trading / cross-market)	Medium to high (wash trading exceeds 70%)	Medium (arbitrage partly corrects; market is maturing)	Medium (enforcement costs are high)	Third
Consumer protection (information asymmetry)	Medium (74–89% of retail accounts lose money; based on ESMA CFD data, an upper-bound estimate for perpetuals)	Medium (information transparency is improving)	Medium (largest trade-off; value-dependent)	Fourth

Table 26-1. Prioritization matrix for regulating the perpetual futures market (Data source: constructed by the author)

This prioritization points to an important economic insight: in regulating the perpetual futures market, basic, innovation-neutral institution-building—such as asset segregation and financial audit—often yields the highest net benefit. Such regulatory measures require neither complex technical knowledge nor fine calibration of market microstructure, and their core logic is entirely consistent with the basic principles of traditional financial regulation. Conversely, the interventions that intuitively attract the most public attention—such as directly limiting retail leverage or banning specific products—have more complex economic consequences, accompanied by the largest unintended side effects and value controversies. A rational regulatory framework should therefore first shore up the foundation of market integrity and then extend gradually to the regulation of systemic risk and market conduct. This inside-out, foundation-to-complexity regulatory path is consistent both with the economic principle of net-benefit maximization and with the historical experience of the evolution of traditional financial regulation.

26.3 The regulatory toolbox

Having identified the failure areas in the perpetual futures market that require regulation and their priorities, the core challenge facing regulators is how to choose the appropriate policy tools. Every regulatory intervention not only generates direct compliance costs but also systematically alters the market's microstructure, thereby profoundly affecting market quality. This section provides an economic evaluation of four categories of core policy in the regulatory toolbox, using the five-dimensional market-quality model introduced in Chapter 25 (transaction costs, price discovery, resilience, fairness, and accessibility) to analyze systematically which dimensions these tools improve and on which dimensions they produce unavoidable side effects.

26.3.1 Leverage limits

Capping the maximum leverage on perpetual futures is the most intuitive response of regulators facing systemic risk. Absent external intervention, the crypto-derivatives market—whether centralized exchanges or decentralized protocols—generally offers extremely high leverage (often as high as 100x to 125x) to attract retail traders seeking high-risk returns [23]. This high-leverage environment, however, greatly increases the probability of liquidation cascades, allowing a single trader's insolvency to spread rapidly to the entire market through the automatic-liquidation mechanism and trigger liquidity evaporation and a price collapse. Regulators therefore generally favor cutting the leverage cap sharply, for example limiting it to the range of 20x to 50x.

Leverage limits have their most pronounced positive effect on market resilience. When the maximum leverage is forcibly reduced, traders must post a higher proportion of initial margin, which directly increases the thickness of a position's buffer. When the market faces sharp volatility, a thicker margin buffer can effectively delay the triggering of liquidation, thereby breaking the vicious circle within the liquidation-price-liquidity positive-feedback loop in which falling prices trigger liquidations and liquidations in turn push prices lower still. This intervention, in essence, provides a public good for

the entire market—systemic resilience against extreme shocks—by weakening individuals' capacity to take on excessive risk. At the same time, leverage limits improve market fairness to some degree, because they protect the retail traders who lack risk-management experience and are prone to the temptation of extreme leverage from severe losses [24].

This improvement in resilience, however, comes at the cost of market accessibility. When leverage is strictly limited, traders with smaller capital must commit more funds to establish a position of the same size, which in effect raises the barrier to market entry and excludes some participants. In addition, transaction costs are affected negatively. Although the explicit fee rate does not change, the reduction in capital efficiency means traders must leave more funds idle as margin, and this opportunity cost constitutes an implicit trading friction. Price discovery, meanwhile, remains relatively neutral in this process, because leverage limits mainly affect position size and do not directly impede the incorporation of new information into prices.

Figure 26-2 uses the leverage cap (2x to 125x) as its parameter axis, with the left and right panels tracking the improvement in market resilience and the loss of accessibility, respectively: when leverage is lowered from 125x to 50x, the resilience gain is substantial while the accessibility loss is limited; compressing it further below 10x flattens the marginal resilience gain but sharply accelerates the deterioration in accessibility, and the sweet spot marked jointly by the two panels (50x) is the core zone of the policy trade-off.

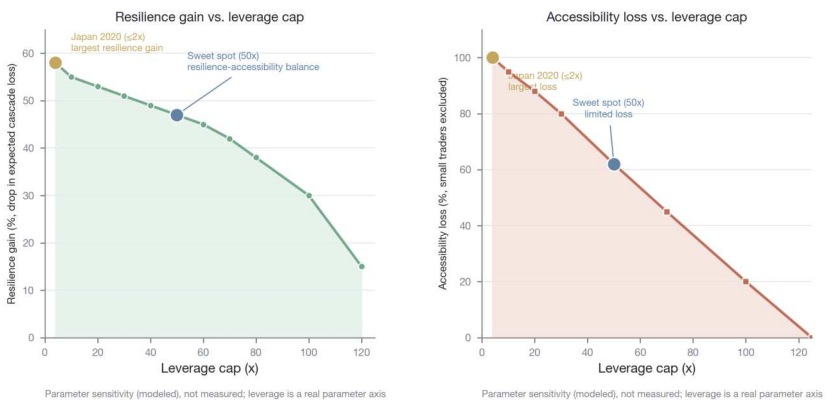


Figure 26-2. Parameter-sensitivity analysis of leverage limits

More complicated still, leverage limits face a serious risk of regulatory arbitrage in implementation. If a strict leverage cap is imposed within a single jurisdiction alone, traders demanding high leverage will quickly move to offshore exchanges or decentralized platforms. Japan's 2020 policy of cutting the leverage cap from 25x to 2x shows that overly stringent limits lead to a large-scale outflow of trading volume, so that the regulator instead loses its capacity to monitor the market (Section 26.7.1 documents this trade-off with policy data from Japan's Financial Services Agency (FSA)) [25]. The net effect of leverage limits therefore depends heavily on how the parameter is set: too loose (such as 100x) fails to guard effectively against liquidation cascades, whereas

too strict (such as 2x) triggers destructive regulatory arbitrage, and a moderate limit of around 50x may be the better compromise between resilience and accessibility.

26.3.2 Disclosure and transparency

Unlike leverage limits, which directly intervene in market trading behavior, disclosure and transparency requirements are a non-intrusive regulatory tool. This kind of policy does not prohibit any particular business model or product structure but instead mandates that exchanges publicly disclose key information about their internal operations, including the state of asset reserves, the summary terms of market-maker agreements, liquidation parameters and mark-price algorithms, and any conflicts of interest arising from proprietary market-making. The core of this regulatory philosophy is that, by eliminating information asymmetry, market participants can price risk on the basis of full information, so that the market mechanism itself spontaneously penalizes platforms that take excessive risk or act in bad faith.

Judged against the same five dimensions, the most direct and significant contribution of disclosure requirements is to improve market fairness. As described in Section 26.1.1, the information gap between exchanges and retail participants leaves ordinary users in a state of blind trust regarding core data such as order flow, liquidation logic, and market-maker privileges [26]. Mandatory disclosure of market-maker agreements and proprietary-trading positions can effectively curb platforms' use of their informational advantage for insider trading or market manipulation, ensuring that all participants stand on an equal footing at the level of the rules. In addition, greater transparency in the liquidation mechanism also helps improve the efficiency of price discovery, because market participants can more accurately anticipate the system's response under extreme conditions and thus make more rational pricing decisions.

The direct impact of disclosure policy on the dimensions of resilience, transaction costs, and accessibility is relatively neutral. It does not directly change the economic cost of trading, nor does it raise the barrier to market entry. In the long run, however, greater transparency can strengthen public trust in the market and reduce participants' concerns about credit risk, and this decline in the trust premium may translate indirectly into lower implicit transaction costs. At the same time, under market stress, a transparent picture of asset reserves can effectively prevent an irrational, panic-driven bank run, providing psychological support for market resilience.

Although disclosure is regarded as the lowest-cost regulatory tool with the smallest side effects, its effectiveness faces one key limitation: the disclosure itself must be truthful and verifiable. The limitations of proof of reserves (see Section 26.2.4—it can prove on-chain assets but cannot reveal off-chain liabilities and off-balance-sheet leverage [22]) are a case in point. Before its collapse, FTX also presented seemingly healthy financial statements to investors and venture-capital firms, but this independently unverified information ultimately proved to be gravely inaccurate. Disclosure therefore translates into effective market transparency only when it is combined with strict external audit and legal liability; otherwise it may merely produce a false sense of security.

26.3.3 Market-maker obligations

Liquidity is the foundation on which the perpetual futures market keeps running, and market makers are the core force supplying it. In an unregulated crypto market, most market makers operate on a voluntary liquidity-provision model that is highly flexible but leads them, when the market suffers an extreme shock, to withdraw en masse out of rational individual risk-management considerations—producing the tragedy of the commons in liquidity described in Section 26.1.1, in which individually rational risk-averse behavior causes liquidity to evaporate instantly at the aggregate level, aggravating one-sided crashes and liquidation cascades [12]. To address this undersupply of a public good, regulators typically introduce a designated-market-maker regime.

The core logic of the designated-market-maker regime is the "parity of rights and obligations." The exchange grants specific market makers privileges such as fee rebates, dedicated API channels, or lower latency; in exchange, these market makers must bear explicit liquidity-provision obligations. These obligations typically include maintaining minimum order depth within a specified price band, adhering to a maximum bid-ask spread, and meeting minimum uptime requirements. More critically, regulation requires that these obligations remain in force even under "abnormal market conditions" and cannot be waived at will simply because volatility has intensified [27].

Market-maker obligations, by contrast, are aimed mainly at improving market resilience. By requiring core liquidity providers to maintain quotes in times of crisis, the market can retain a minimal capacity to absorb selling pressure under extreme conditions, avoiding a disorderly price collapse. At the same time, this regime improves market fairness, because it makes public and standardizes the private incentive agreements that were originally hidden between exchanges and market makers, eliminating opaque interest arrangements.

The introduction of market-maker obligations, however, inevitably has a negative impact on transaction costs. When market makers are compelled to provide liquidity under extreme conditions, they face significant adverse-selection risk and inventory-devaluation risk. To compensate for this potential extreme loss, market makers must widen their bid-ask spreads under normal conditions, passing the cost of bearing the obligation on to ordinary traders. Moreover, the design of this regime faces a difficult balancing challenge: how to define "abnormal market conditions." If the definition is too broad, market makers can easily trigger the exemption clause, rendering the obligation a dead letter; if it is too strict, market makers may, in a genuine systemic crisis, suffer losses beyond their capital capacity and ultimately go bankrupt. Once a core market maker fails, it not only cannot provide liquidity but, as a large defaulting counterparty, further aggravates the market's systemic risk.

26.3.4 Capital requirements and asset segregation

If leverage limits and market-maker obligations address risks at the trading level, then capital requirements and asset segregation strike directly at the core of an exchange's governance structure. The misappropriation of FTX customer assets described in Section 26.2.4 shows that when an exchange's internal governance fails completely, any discussion of liquidation-engine efficiency or order-book depth loses its meaning [28].

In the post-FTX era, therefore, establishing strict capital-adequacy standards and a customer-asset-segregation regime has become the top-priority agenda for regulators worldwide.

Capital-requirement policy stipulates that an exchange must hold a certain proportion of its own funds as a buffer to absorb operating losses, technical failures, or cyberattacks. This is analogous to the Basel Accords framework in traditional banking, which aims to ensure that financial institutions have sufficient solvency when facing unexpected shocks. Asset-segregation requirements are more direct still: they mandate that an exchange strictly separate customers' deposited funds from the platform's own operating funds at the legal and physical levels, and they prohibit the exchange from using customer assets for proprietary trading, staking and lending, or as collateral [29].

Capital requirements and asset segregation deliver their largest gains on the resilience and fairness dimensions. Asset segregation fundamentally cuts off the exchange's path to misappropriating customer funds, eliminating the possibility of a systemic collapse triggered by internal moral hazard on the platform. This not only protects the absolute safety of customer assets but also restores participants' basic trust in centralized trading infrastructure, greatly strengthening the market's resilience against endogenous risk. At the same time, higher capital requirements ensure that an exchange has sufficient own funds to meet its payout obligations under extreme market stress, preserving the fairness of trade settlement.

The distinctive feature of this regulatory intervention is that its enormous benefits come with relatively limited negative side effects, concentrated mainly on accessibility. It does not interfere with traders' leverage choices or change the microstructure of the order book, so it has no direct effect on price discovery or transaction costs. The only dimension that may be slightly affected is market accessibility. Strict capital-adequacy standards and regular external-audit requirements substantially raise the fixed cost of opening and operating a cryptocurrency exchange, which may squeeze some smaller, less well-capitalized platforms out of the market and increase industry concentration. From the perspective of protecting investors, however, eliminating the marginal platforms that cannot provide basic safety guarantees is itself the intended outcome of regulatory selection, and this reduction in accessibility is an acceptable cost in the overall welfare analysis.

26.3.5 The five-dimensional impact matrix

Through the analysis of the four categories of core regulatory tools above, we can clearly see that no single policy can resolve all market failures at once. Every intervention aimed at repairing a specific market failure triggers a chain reaction in the complex market ecosystem, improving certain quality dimensions while inevitably causing friction or loss on others. Table 26-2 compresses these complex trade-offs into a five-dimensional impact matrix of regulatory tools, to facilitate comparison of the net effects of different policies.

Table 26-2 evaluates, one by one, the direction and intensity of the impact of seven categories of regulatory tools along five dimensions: transaction costs, price discovery, resilience, fairness, and accessibility. In the table, an upward arrow (↑) denotes improve-

ment on that dimension, a downward arrow (↓) denotes deterioration, and a rightward arrow (→) denotes roughly no change, with the number of arrows representing the magnitude of the impact. This standardized scoring framework makes horizontal comparison across policies immediately clear and also reveals the unavoidable trade-offs each intervention entails across the dimensions. Of these, the two tools of mandatory KYC/AML and an outright trading ban are not treated separately in the preceding subsections; their five-dimensional scores are based on their typical effects in the discussions of jurisdiction and anti-money laundering (see Sections 26.4.2 and 26.6.2): the former trades accessibility for gains in legitimacy and fairness, while the latter, by driving trading underground, produces negative effects on most dimensions.

Regulatory tool	Transaction costs	Price discovery	Resilience	Fairness	Accessibility	Net assessment
Leverage limit (to 20x)	↑	→	↑↑↑	↑	↓↓	Very large trade-off; highly prone to regulatory arbitrage
Leverage limit (to 50x)	→	→	↑↑	↑	↓	A better compromise; balances resilience and accessibility
Disclosure and external audit	→	↑	→	↑↑	→	Positive net benefit; very low cost; strongly recommended
Market-maker obligations	↑	→	↑	↑	→	Moderate net benefit; highly dependent on the design of exemption clauses
Capital requirements and asset segregation	→	→	↑↑	↑↑	↓	Very high net benefit; a core tool against severe systemic risk
Mandatory KYC/AML	↑	→	→	↑	↓↓	Large trade-off; touches on the decentralization values of crypto markets
Outright trading ban	N/A	N/A	N/A	N/A	↓↓↓	Negative net benefit; drives the market underground; counterproductive

Table 26-2. Five-dimensional market-quality impact matrix of the regulatory toolbox (Data source: constructed by the author)

This matrix distills a core insight of regulatory economics: excellent regulatory design

does not seek perfect improvement on every dimension but instead, on the basis of a deep understanding of a policy's side effects, chooses the tool combination that "maximizes net benefit" or "keeps side effects most manageable." For example, although capital requirements and asset segregation slightly reduce accessibility, their enormous benefit in preventing an FTX-style disaster makes them a primary policy. Conversely, although an extreme leverage limit (such as 2x) can significantly improve resilience in theory, its severe damage to accessibility and the regulatory arbitrage it triggers often make it a failed policy in practice. The core of regulatory design lies in finding, on this multidimensional trade-off frontier, the equilibrium point best suited to the market's current stage of development.

26.4 The jurisdiction problem

The global, decentralized character of perpetual futures raises a question thornier than how to regulate them: who regulates. Regulatory jurisdiction in traditional financial markets rests on the principle of territoriality and on entity registration, and a regulator holds exclusive jurisdiction over financial activity within its sovereign borders. The servers of a perpetual futures trading platform, however, may be deployed across multiple jurisdictions, the core development team may be distributed around the world, and liquidity providers and traders connect from everywhere via the internet. This borderless technical architecture conflicts fundamentally with a legal framework based on sovereign borders. The fragmentation of global regulation is not the incidental result of regulators' lack of willingness to cooperate but reflects a structural contradiction: global consistency, innovation-friendliness, and consumer protection cannot all be achieved at once.

26.4.1 The impossible trinity

The formulation of international financial-regulatory policy is subject to a structural trade-off framework that can be formalized as the regulatory "impossible trinity." This concept derives from the classic macroeconomic theory that a fixed exchange rate, free capital mobility, and independent monetary policy cannot all be achieved simultaneously [30]. This framework is not a strict mathematical impossibility theorem but an analytical tool that reveals the structural tension among policy objectives. In the context of regulating crypto-asset derivatives, the three vertices of the triangle are global consistency, innovation-friendliness, and consumer protection. Global consistency means that countries implement uniform regulatory standards to eliminate cross-border arbitrage space; innovation-friendliness means reserving sufficient room for experimentation for emerging technologies and business models; and consumer protection means reducing retail investors' risk exposure through entry standards, disclosure, and leverage limits. The mutual exclusivity among the three arises from resource constraints and incentive incompatibility: when the international community pursues global consistency and consumer protection at once, it must formulate and enforce strict uniform standards, which inevitably compresses the room for each country to experiment with locally tailored innovation. When global consistency and innovation-friendliness are prioritized together, by contrast, countries can agree only on a low standard, and the strength of consumer

protection is necessarily sacrificed.

Figure 26-3 shows the inherent conflict among these three policy objectives: pursuing global consistency and consumer protection simultaneously requires the enforcement of strict uniform standards, which suppresses early-stage technological innovation in lightly regulated jurisdictions—Europe's mature financial infrastructure can still bear the strict compliance costs, but fintech firms in emerging-market countries may be forced to abandon innovation projects as a result; whereas pursuing global consistency and innovation-friendliness simultaneously can only reach consensus on a loose minimum standard, which cannot provide effective consumer protection and may allow systemic risk and market manipulation to spread.

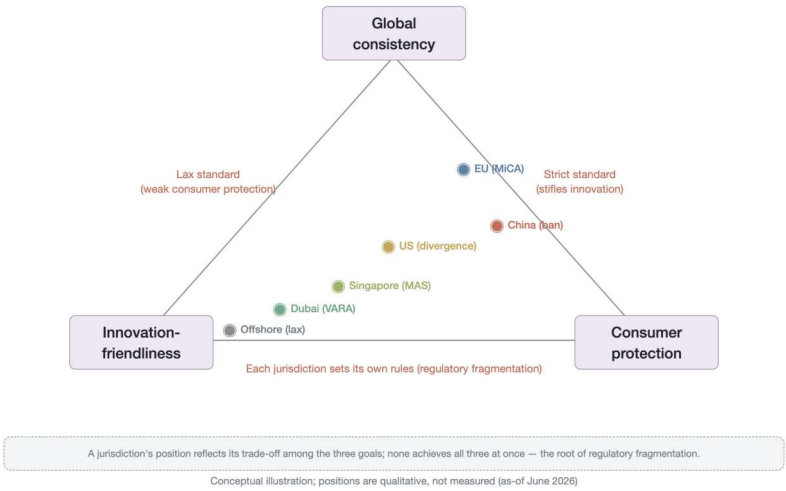


Figure 26-3. The regulatory "impossible trinity"

Faced with this impossible trinity, national regulators can only choose to compromise among the three objectives. The choice in reality is usually to give up global consistency and instead seek a balance between innovation-friendliness and consumer protection according to a country's own economic-development strategy and financial-stability concerns. This policy choice, grounded in sovereign autonomy, directly produces the inconsistency of global rules. The current regulatory fragmentation is therefore an inevitable product of structural constraints, and each country's or region's different position in the triangle reflects its different pricing of the benefits and risks of financial innovation.

From an economic perspective, this trade-off reflects different countries' different valuations of the benefits and risks of innovation. Developed economies generally have stronger financial-regulatory capacity and investor-protection mechanisms and can therefore bear higher innovation risk. By contrast, some developing countries may face more severe financial-stability challenges and therefore tend to adopt a more conservative stance. In addition, each country's industrial policy also influences its regulatory choice. Countries that hope to attract fintech talent and firms may choose

a more relaxed regulatory environment, whereas those that rely mainly on traditional financial institutions may take a more cautious attitude toward emerging technologies.

26.4.2 The global regulatory map

Based on their different trade-offs within the impossible trinity, the world's major economies have taken sharply different policy stances on regulating perpetual futures. This divergence of stances is not a simple binary opposition but forms a continuous spectrum from outright prohibition to active embrace. On this spectrum, the policy choices of different jurisdictions determine not only their share of the global crypto-derivatives market but also significantly affect the geographic distribution of trading activity.

Figure 26-4 depicts the current spectral distribution of global regulation. At the far left of the spectrum are the jurisdictions that adopt outright prohibition. China imposed a strict ban on cryptocurrency trading in 2021, aiming to sever completely the link between crypto-assets and the fiat-currency system in order to guard against capital flight and financial risk. Empirical research shows, however, that the enforcement effect of this extreme policy is limited: trading activity did not disappear but shifted to offshore platforms through intermediaries such as stablecoins, and Chinese investors continue to exert significant influence on Bitcoin price formation [31]. This finding carries an important policy implication: even the strictest regulatory measures cannot fully halt trading activity and may instead cause regulators to lose visibility into and control over that activity.

To the right of the prohibition end are the jurisdictions that impose strict regulation. Japan's FSA strictly limits the leverage cap on crypto-derivatives to 2x, and this extremely conservative consumer-protection measure forces traders seeking high leverage toward unrestricted overseas platforms, thereby weakening the domestic regulator's influence and the competitiveness of Japan's fintech industry (Section 26.7.1 documents this trade-off with FSA policy data).

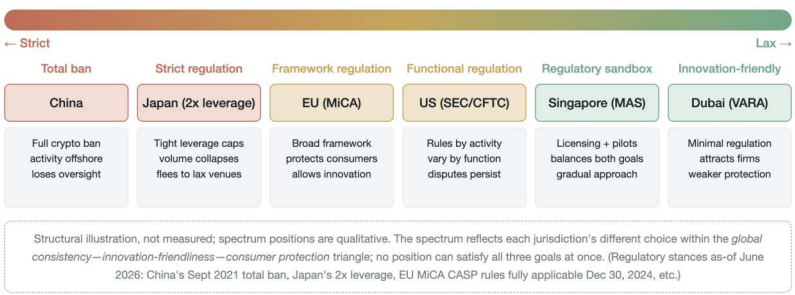


Figure 26-4. The global regulatory map and policy spectrum

The middle of the spectrum comprises the regions that attempt to build comprehensive

frameworks. Through the Markets in Crypto-Assets (MiCA) regulation, the European Union established the world's first cross-national comprehensive regulatory framework; although the regulation targets mainly spot markets and stablecoins, it provides a foundational classification and entry standard for derivatives regulation. The core feature of the MiCA framework is that it establishes uniform EU market-access rules, allowing a crypto-asset service provider licensed in one member state to provide services in other member states. This "single passport" mechanism substantially lowers cross-border compliance costs but also requires all participants to comply with relatively strict consumer-protection standards. The United States, by contrast, is in a state of functional-regulatory divergence: the U.S. Securities and Exchange Commission (SEC) and the CFTC have a long-standing dispute over jurisdiction of crypto-derivatives, and this enforcement-based rather than legislation-based regulatory model brings extremely high uncertainty, which has in effect entrenched the market dominance of offshore exchanges. The U.S. divergence reflects a deeper problem: different regulators may have different classificatory understandings of the same category of financial product, creating opportunities for regulatory arbitrage.

At the right end of the spectrum are the jurisdictions inclined toward innovation-friendliness. The Monetary Authority of Singapore (MAS) combines a regulatory sandbox with a licensing regime, attempting to strike a balance between attracting fintech innovation and guarding against money-laundering risk. The MAS framework allows qualified firms to test new products and services in a restricted environment while maintaining basic consumer protection. The advantage of this approach is that it can provide space for innovation without entirely abandoning regulation. Dubai's Virtual Assets Regulatory Authority (VARA) offers a more light-touch regulatory framework, attracting a large number of crypto firms through clear and flexible rules. VARA's success lies in providing clear legal status and relatively low compliance costs, which has led many global crypto firms to establish regional headquarters in Dubai. At the far right of the spectrum, certain offshore jurisdictions impose almost no substantive regulation on crypto-derivatives; they offer the highest degree of freedom to innovate and the lowest compliance costs but also entirely abandon consumer protection.

26.4.3 The economic consequences of fragmentation

The significant differences in global regulatory policy directly give rise to regulatory arbitrage. Regulatory arbitrage refers to market participants exploiting the inconsistency of rules across jurisdictions to shift business or trading activity to the regions with the lowest compliance costs and the fewest restrictions. In the perpetual futures market, this arbitrage takes three main forms: arbitrage of exchange domicile, circumvention of users' geographic location, and liquidity fragmentation. Arbitrage of exchange domicile means that an exchange chooses to register in the most lightly regulated region and then provides services to users worldwide via the internet. Circumvention of users' geographic location means that users employ technical means such as VPNs to bypass geographic restrictions and access trading platforms banned in their home country. Liquidity fragmentation means that, because of regulatory restrictions, trading liquidity is dispersed across multiple platforms, resulting in insufficient liquidity in the global market.

When a country imposes strict regulation, the direct consequence is often not the elimination of high-risk trading but the loss of jurisdiction over it. The migration of trading volume to offshore platforms after China's 2021 ban, described in Section 26.4.2, is a paradigmatic case of regulatory arbitrage causing a country to lose its regulatory capacity [31]: the effectiveness of regulation depends not only on the strictness of the rules but, more importantly, on their enforceability. This substantive regulatory failure instead increases systemic risk, because regulators cannot obtain trading data or conduct risk monitoring.

A deeper problem triggered by regulatory arbitrage is the loss of global market efficiency. The Financial Stability Board's (FSB) assessment report notes that significant gaps and inconsistencies in the implementation of crypto-asset regulation are undermining the interconnectedness of the global financial system [32]. At the level of micro-market structure, this fragmentation manifests as a pricing-island effect. Because of compliance requirements, fiat on- and off-ramp restrictions, and region-specific trading bans, the funds and positions of cross-market arbitrageurs cannot flow freely between exchanges in different jurisdictions. The blocking of arbitrage paths allows cross-market price differences—which should converge instantly—to persist, reducing the overall efficiency of price discovery in the global market. For example, if the perpetual futures price on a European exchange differs significantly from that on an Asian exchange, traditional arbitrageurs should eliminate the difference by buying in the low-price market and selling in the high-price market. But if the European exchange is prohibited from serving Asian users, or the Asian exchange is prohibited from serving European users, this arbitrage mechanism breaks down and the price difference may persist for a long time.

Moreover, regulatory fragmentation may trigger a race to the bottom. To attract the tax revenue, employment, and technology spillovers that crypto firms bring, some jurisdictions may face political pressure to lower their regulatory thresholds. If this competition evolves into a scramble to relax standards, the level of consumer protection worldwide will decline overall. When large numbers of retail investors conduct high-leverage trading on offshore platforms that lack asset-segregation and disclosure requirements, systemic risk is not reduced by some countries' strict regulation but instead becomes harder to monitor and guard against, because risk concentrates in opaque regions. This phenomenon has precedents in traditional finance as well: in certain historical periods, some countries competed to loosen regulatory standards in order to attract financial institutions, which is thought to have aggravated the subsequent fragility of the financial system.

26.4.4 The path to coordination

Faced with the economic costs of regulatory fragmentation, establishing some form of international coordination mechanism becomes an inevitable choice. Because of the structural constraint of the impossible trinity, however, pursuing a fully unified global regulatory framework is unrealistic in the short term. A more feasible path is to seek a compromise between core consensus and regional differences through bilateral mutual recognition and the setting of minimum standards.

Bilateral mutual-recognition agreements are an effective mechanism for lowering cross-border compliance costs. When the regulatory objectives and rule intensity of two jurisdictions are sufficiently close, they can mutually recognize each other's regulatory licenses. This means that a perpetual futures exchange licensed in one country can legally operate in another without reapplying. This mechanism can locally alleviate liquidity fragmentation without requiring global uniformity. The precondition for implementing bilateral mutual recognition, however, is that both sides must be highly consistent on core issues such as asset segregation, anti-money laundering, and the prevention of market manipulation. Successful examples of such agreements include the financial-services passport regime within the European Union, which allows a financial institution licensed in one member state to provide services in other member states, substantially lowering the cost of cross-border financial services.

At the multilateral level, setting international minimum standards is key to preventing a race to the bottom. In 2023, the International Organization of Securities Commissions (IOSCO) issued policy recommendations for crypto and digital-asset markets, providing benchmark regulatory guidance for its member countries [33]. This minimum-standard model is analogous to the role of the Basel Accords in traditional banking: it requires all participating countries to meet basic requirements on customer-asset protection and conflict-of-interest management but allows each country to add stricter rules on top (such as different leverage caps) according to its national circumstances. The advantage of this approach is that it can maintain a minimum protection standard while allowing each country to innovate according to its own situation.

IOSCO's recommendations cover core areas such as anti-money laundering, know-your-customer, the prevention of market manipulation, and customer-asset protection. Through these minimum standards, the international community can prevent regulatory competition from evolving into a total collapse of consumer protection. At the same time, countries can still maintain differences in leverage limits, product classification, and entry standards to suit their stage of economic development and risk appetite. This approach has already achieved some success in other financial domains; for example, the recommendations of the Financial Action Task Force (FATF) on anti-money laundering have been adopted by most countries.

Although international organizations are actively promoting coordination, the surrender of sovereignty in reality still faces complex political-economy constraints. Global regulatory unification involves not only the alignment of technical standards but also touches on each country's financial sovereignty, tax interests, and industrial policy. Over the next 5 to 10 years, therefore, the most likely regulatory configuration for the perpetual futures market combines a global consensus on bottom-line issues such as anti-money laundering and asset custody with persistent regional differences on dimensions such as leverage limits, product access, and derivatives classification. This hybrid model of core consensus plus regional difference is the most realistic response to the regulatory impossible trinity.

26.5 Functional regulation versus entity-based regulation

The traditional financial-regulatory framework rests mainly on the logic of entity-based regulation. This paradigm assumes that financial activity is tightly bound to specific types of institution, so regulators achieve their objectives by licensing and imposing overall constraints on specific entities such as banks, securities exchanges, or commodity-futures commissions. The emergence of perpetual futures, however, poses a fundamental challenge to this paradigm. As a derivative that occupies a dominant position in the crypto-asset market, perpetual futures operate in a way that crosses the clear boundaries of traditional financial products, so a regulatory method based on entity classification faces significant limitations when confronting this kind of innovative instrument. This predicament has driven an evolution of the regulatory paradigm, prompting academics and policymakers to reexamine the theoretical value and practical potential of functional regulation.

26.5.1 The limits of entity-based regulation

Entity-based regulation works effectively only on the premise that the regulated object can be accurately assigned to a predefined legal and economic category. In the U.S. financial-regulatory system, this classification process usually determines whether a product falls under the jurisdiction of the SEC or the CFTC. For digital assets and their derivatives, this either-or classification framework has produced long-standing jurisdictional disputes and regulatory vacuums.

The legal nature of perpetual futures first faces the challenge of the definition of a security. The Howey test, established by the U.S. Supreme Court in 1946, is the core standard for judging whether a transaction constitutes an investment contract and therefore a security [34], and the SEC adopted this standard in its analytical framework for digital assets [3]. The test requires an investment of money, in a common enterprise, with a reasonable expectation of profits, where the profits derive entirely or predominantly from the efforts of others. As a derivative, perpetual futures anchor their value to the price of an underlying crypto-asset, and traders profit from price movements by going long or short. This mechanism represents no ownership of any issuing entity, and traders' expectation of profit derives from changes in the market price rather than from the operating efforts of a specific management team. Classifying perpetual futures directly as a security under the Howey test framework therefore faces significant legal obstacles.

Having ruled out classification as a security, bringing perpetual futures under the commodity-futures regulatory framework seems the more reasonable path, but this too faces a conflict of technical definition. A traditional futures contract is defined as a standardized agreement to buy or sell a specific quantity of a commodity at a predetermined price on a specific future date. The core innovation of perpetual futures is precisely that they eliminate the expiration date, using the funding-rate mechanism to make the contract price track the spot price closely [13]. This design without a delivery deadline places them, in legal definition, outside the category of traditional futures, so the CFTC faces a challenge of legal interpretation in establishing its jurisdiction. Although the CFTC has recently signaled that it may permit perpetual futures trading

in the United States, this process remains subject to a reinterpretation of the definition of derivatives [35].

This classificatory ambiguity directly leads to the formation of a regulatory vacuum. When a systemically important financial innovation cannot be accommodated by the existing regulatory framework, market participants often exploit the ambiguity for regulatory arbitrage. Capital and trading activity tend to flow to the jurisdictions with the most lenient regulatory requirements or where clear rules have not yet been established. This phenomenon is especially evident in the perpetual futures market, where large trading volumes concentrate on offshore exchanges, leaving the regulators of major economies without effective tools to intervene when facing potential systemic risk and market manipulation. This regulatory failure stems not from regulators' inaction but from the structural limitation of the entity-based regulatory paradigm in the face of cross-boundary financial innovation.

26.5.2 The principle of functional regulation

Faced with the limitations of entity-based regulation, functional regulation offers a more adaptive alternative paradigm. The theoretical basis of functional regulation can be traced to the financial-system design framework of Merton and Bodie [36]. That framework proposes that financial functions are more stable than financial institutions; the form of an institution evolves with changes in technological, political, and cultural norms, but the basic functions that a financial system must perform are enduring. The design and regulation of a financial system should therefore anchor on function rather than on institution.

Applying this theory to regulatory practice, the core logic of functional regulation can be summarized as "same function, same risk, same regulation." In discussing financial-regulatory frameworks, the Bank for International Settlements notes that activity-based or functional regulation is more justified when financial instability originates from a specific activity rather than a specific entity [37]. Under this paradigm, the focus of regulation shifts from "what type of institution is this" to "what economic function is this institution performing." If a decentralized protocol and a centralized exchange both perform the functions of order matching and price discovery, then, for that specific function, they should be subject to the same rules aimed at preventing market manipulation and ensuring transparency, regardless of their legal organizational form.

The economic advantage of functional regulation lies in its ability to effectively eliminate the incentive for regulatory arbitrage. Under entity-based regulation, innovators have a strong incentive to circumvent strict regulatory requirements by changing their organizational form or product name. Functional regulation, by seeing through the appearance of organizational form and imposing constraints directly on the substance of the economic activity, renders such formal circumvention ineffective. Moreover, this paradigm is more friendly to technological innovation. It does not need to create a new legal category for every newly emerging financial instrument or institutional form but instead decomposes the new thing into a combination of known financial functions and applies the corresponding existing rules.

Figure 26-5 contrasts the structural difference between the two regulatory paradigms:

entity-based regulation attempts to force perpetual futures into ill-fitting institutional categories, whereas functional regulation achieves precise rule-matching by decomposing the specific functions they perform.

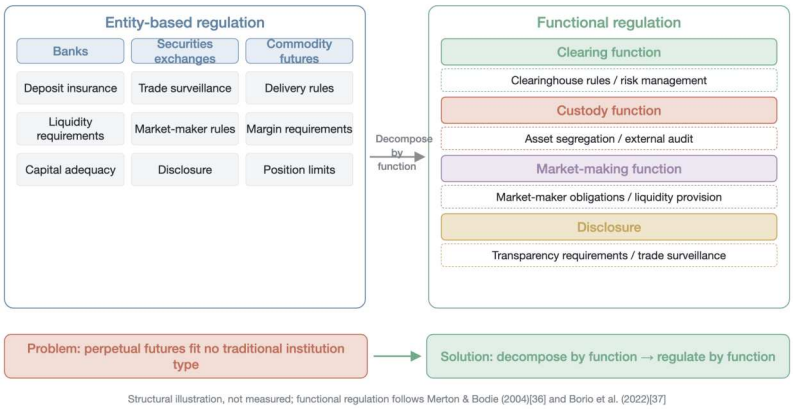


Figure 26-5. Structural comparison of entity-based and functional regulation

26.5.3 The functional decomposition of perpetual futures

Applying the functional-regulation paradigm to the perpetual futures market requires a deep functional decomposition of the platforms that offer perpetual futures trading. Whether a centralized exchange or a decentralized protocol, its operation can be deconstructed into four core financial functions: clearing, custody, market-making, and disclosure. By mapping these functions to their corresponding regulatory objectives, one can build a systematic and highly adaptive regulatory framework.

The clearing function is the primary source of systemic risk in the perpetual futures market. Because of the high leverage inherent in perpetual futures, the design of the liquidation mechanism directly bears on market resilience. Under a functional-regulation framework, any entity or protocol that performs the clearing function must be subject to rules analogous to those of a traditional clearinghouse. These include establishing a robust risk-management framework, setting reasonable initial and maintenance margin requirements, and maintaining an insurance fund sufficient to absorb the bad-debt losses of underwater positions under extreme market conditions. For decentralized exchanges, this means that the liquidation logic in their smart contracts must meet minimum risk-control standards to prevent a liquidation cascade from triggering a broad market collapse.

The custody function concerns the holding and safekeeping of customer assets. The lesson of the FTX episode is that when a trading platform performs both trade-matching and asset-custody functions, the risk of misappropriating customer assets is extremely high. Functional regulation requires strict asset-segregation rules for entities that perform the custody function. This means that customers' margin must be completely separated from the exchange's operating funds at the physical or cryptographic level

and must undergo regular external audit or provide verifiable transparency through cryptographic means such as proof of reserves.

The market-making function directly affects the supply of liquidity in the perpetual futures market but is also prone to conflicts of interest and market manipulation. When the exchange itself or an affiliated entity acts as a major market maker on the platform, the fairness problem comes to the fore. Functional regulation requires imposing explicit market-maker obligations on participants that perform the market-making function, such as maintaining minimum liquidity and limiting spreads under abnormal market conditions. At the same time, a strict firewall must be established to prevent the exchange performing the matching function from using non-public order-flow information to gain an improper advantage for its proprietary market-making business.

The disclosure function is the foundation for mitigating information asymmetry among market participants. Functional regulation requires that platforms offering perpetual futures trading, whatever their architecture, provide the market with sufficient transparency. This includes publishing trading data, disclosing the key terms of market-maker agreements, clarifying the calculation methods of liquidation parameters and the mark price, and providing regular trade-surveillance reports to deter manipulation.

Figure 26-6 systematizes the decomposition of the four functions above into a regulatory mapping, enabling regulators to identify precisely the functional nodes that require intervention rather than imposing homogeneous compliance obligations on an entire platform indiscriminately. For a centralized exchange that bears multiple functions at once, this framework requires the establishment of internal firewalls between the functions; for a decentralized architecture in which the functions are dispersed across different protocols, the compliance obligations can be assigned to the actual performer of each function.

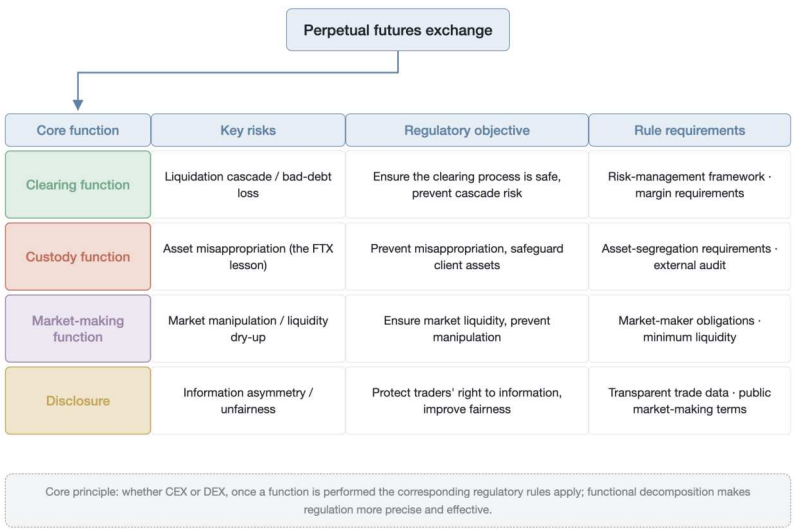


Figure 26-6. Functional decomposition of a perpetual futures exchange and the corresponding regulatory mapping

26.5.4 Implementation challenges

Although functional regulation offers an elegant solution in theory, its actual implementation in the perpetual futures market still faces major challenges. These challenges stem mainly from the special features of crypto-market microstructure and the inherent properties of decentralized technology.

The first is the problem of blurred functional boundaries. In traditional finance, clearing, custody, and trade-matching are usually performed separately by different specialized institutions. In the crypto-asset market, however—especially within centralized exchanges—these functions are highly concentrated in a single entity. This cross-functional coupling makes it exceedingly difficult to apply segregated regulation to a single function. For example, when an exchange controls both the liquidation engine and the market-making funds, requiring it to build an absolute internal firewall often lacks a credible verification mechanism. It is very difficult for regulators to force the separation of these deeply coupled functions without disrupting the exchange's existing business model and efficiency advantages.

The second is the challenge of enforcement difficulty—namely, "who determines that a certain function has been performed" and "against whom is accountability pursued." In the context of decentralized exchanges, this problem is especially acute. A decentralized protocol is, in essence, a set of automatically running smart-contract code that lacks a traditional legal-person entity as a regulatory object. When an automated market maker protocol performs the functions of price discovery and liquidity provision, it is difficult for regulators to determine whether the protocol's original developers, the holders of governance tokens, or the operators of the front-end interface that provides liquidity should bear compliance responsibility. This absence of a responsible subject greatly weakens the deterrent power and enforceability of functional regulation when it confronts a purely on-chain application.

Moreover, the scope of applicability of functional regulation has inherent limits. Although it can effectively handle activities highly similar to traditional financial functions, for mechanisms unique to crypto markets—such as algorithm-based automated liquidation and flash loans—the functional-regulation framework may lack ready-made rules to draw on. This means that regulators must not only decompose functions but also create entirely new regulatory standards for these new functions, which undoubtedly increases the complexity and time cost of regulatory design. Functional regulation is therefore not without limits; it requires continual adjustment and refinement in defining functional boundaries, adapting to cross-boundary coupling, and innovating enforcement means.

26.6 Regulating DeFi perpetual futures

The preceding analysis has systematically evaluated the various regulatory tools aimed at centralized exchanges. Assessing the actual effectiveness of these tools, however, requires considering one key variable: as the principal escape valve for regulatory arbitrage, decentralized exchanges' accessibility directly determines the net effect of regulating the centralized market. If traders can migrate almost frictionlessly to permissionless

decentralized protocols, the actual binding force of the leverage limits or compliance requirements imposed on centralized exchanges is greatly diminished. Understanding the regulatory boundaries of DeFi perpetual futures is therefore a prerequisite for accurately assessing the effectiveness of the entire regulatory toolbox.

Decentralized exchanges pose a fundamental challenge to the traditional regulatory paradigm. When an exchange is no longer an entity operated by a specific company but a set of smart contracts deployed on a blockchain and jointly maintained by nodes worldwide, the traditional "regulate the entity" approach faces the risk of complete failure. This failure stems not from a lack of regulatory will but from the physical limits of the technical architecture: code has no nationality, a smart contract has no legal personality, and a decentralized network has no chief executive who can be subpoenaed. This does not mean, however, that decentralized finance is unregulable; only that the entry point of regulation must shift. This section analyzes how a decentralized architecture dismantles traditional compliance tools and explores how, in an environment without a centralized controller, a new type of regulatory framework can be built through code audits, front-end restrictions, and liquidity-provider obligations.

26.6.1 The spectrum of decentralization

Decentralization is not a black-and-white binary state but a continuous spectrum. In the perpetual futures market, a protocol's degree of decentralization directly determines the difficulty of regulatory intervention. Depending on how many core components run on-chain, existing decentralized perpetual futures exchanges fall into three categories: nominal decentralization, hybrid architecture, and fully on-chain architecture [38].

The nominally decentralized architecture is typified by dYdX v3 (its form before migrating to the v4 standalone chain in 2023). The protocol deploys its smart contracts on an Ethereum layer-2 network, achieving decentralization of asset custody and settlement, but its order book and matching engine still run on centralized servers operated by dYdX Trading [39]. This design guarantees users' control over their assets while providing a trading experience close to that of a centralized exchange. From a regulatory standpoint, this architecture retains a clear operating entity and centralized infrastructure, and regulators can enforce compliance requirements by requiring the operating company to implement geo-blocking or identity verification.

The hybrid architecture is typified by GMX, whose core feature is that it abandons the order-book model in favor of a global liquidity-pool design [40]. In this architecture, users' trades are matched directly against the liquidity pool as counterparty, prices are supplied by a decentralized oracle network, and the liquidation logic is executed entirely and automatically by smart contracts [41]. This design markedly raises the degree of decentralization, with no centralized matching engine and no single controlling entity. The hybrid architecture, however, still depends on an external oracle network to feed in price data, which provides a potential intervention node for regulation.

The fully on-chain architecture represents the current endpoint of decentralized evolution, typified by Hyperliquid and dYdX v4. Such protocols are usually built on a dedicated layer-1 blockchain optimized for trading, integrating the order book, matching engine, oracle, and liquidation mechanism entirely into the consensus layer [39] [42].

In Hyperliquid's design, the validator network is responsible not only for producing blocks but also for maintaining the global order book directly in memory, with all state transitions confirmed through Byzantine fault-tolerant consensus [42]. This architecture greatly reduces reliance on centralized components and approaches end-to-end decentralization. It eliminates a certain class of systemic risk arising from single points of failure but at the same time introduces new risks at the consensus and oracle layers, leaving traditional regulatory tools almost nothing to grip onto.

Figure 26-7 places the three architectures above on a continuous spectrum of decentralization from low to high: the higher the degree of decentralization, the lower the regulatory accessibility, which means that the design of regulatory tools must adjust dynamically to the regulated object's position on the spectrum.



Figure 26-7. The decentralization spectrum of DeFi perpetual futures

26.6.2 The failure of traditional tools

Traditional financial regulation is built on the foundational assumption of "intermediary institutions," achieving its policy objectives by licensing entities such as banks, brokers, and exchanges and imposing obligations on them. When a decentralized exchange eliminates the intermediary through smart contracts, this regulatory toolbox encounters a structural failure.

Licensing requirements are the first line of defense in traditional regulation, but in a fully decentralized protocol this requirement loses its object. A smart contract deployed on a blockchain cannot apply for a business license, and there is no legal-person entity to bear the legal consequences of operating without one [43]. Even if regulators attempt to hold the protocol's developers accountable, the nature of open-source software allows the writing and deployment of code to be carried out completely anonymously. More importantly, once a smart contract is set to be non-upgradeable, even its original developers cannot modify or halt its operation, which makes traditional enforcement means such as "revoking the license" or "forcing a shutdown" physically impossible [44].

AML/KYC rules are another pillar of traditional regulation, but they likewise face an enforcement predicament in a permissionless blockchain network. The essential feature

of a decentralized exchange is that it allows any user holding a crypto wallet to interact directly with the protocol, and this interaction occurs in the underlying blockchain network without passing through any centralized gateway that could perform identity verification [45]. Requiring a smart contract to verify a user's real-world identity before executing a trade is not only technically difficult but also runs counter to the core spirit of blockchain pseudonymity. This architectural feature gives decentralized exchanges a natural capacity to evade traditional identity checks, thereby weakening the global anti-money-laundering system in practice [46].

Asset-segregation requirements are used in traditional regulation to prevent intermediaries from misappropriating customer funds, and this is the core mechanism for guarding against an FTX-like collapse. In a decentralized exchange, however, this problem is largely mitigated by technical means. Users' assets are locked in publicly transparent smart contracts, and their transfer and liquidation follow predetermined mathematical logic, making it difficult for any individual or organization to bypass the code rules and use those funds [47]. On the specific dimension of custodial misappropriation, the determinism of code substitutes to some degree for human trust in an intermediary, but this does not eliminate other, new risks—such as oracle manipulation and the inadequate absorption of liquidated positions (see Section 26.6.4).

Regulatory arbitrage is the inevitable result of the failure of traditional tools. Because blockchain networks transcend geographic borders, the strict regulation imposed by one country can often only drive trading activity to offshore hubs or decentralized networks rather than truly eliminate the risk [48]. When compliance costs make centralized exchanges uncompetitive, users seeking high leverage and permissionless trading migrate naturally to decentralized alternatives. This migration not only causes regulators to lose visibility into the market but may also expose users to informal markets that lack basic protection, producing consequences contrary to the original regulatory intent [45].

26.6.3 New regulatory tools

Faced with the failure of traditional tools, the regulatory paradigm must shift from "regulating the entity" to "regulating the code" and "regulating the access point." This shift has given rise to a set of new regulatory tools adapted to the decentralized architecture, which attempt to achieve public-policy objectives without disrupting the underlying technical logic.

Code-audit requirements are becoming a core pillar of decentralized-finance regulation. Since smart contracts substitute for the functions of traditional financial institutions, reviewing the security and logical correctness of the code is equivalent to reviewing the internal controls of a traditional financial institution. Regulators can require that all smart contracts offering financial services to the public undergo a comprehensive security audit by a certified third-party auditor before deployment [49]. Such an audit must cover not only the security vulnerabilities that guard against hacking but also economic stress tests of the liquidation logic, oracle dependencies, and governance permissions. By establishing a standardized audit framework and an entry regime for audit firms, regulation can significantly improve the market's overall resilience without directly interfering in the operation of the protocol [50].

Front-end regulation represents another pragmatic compromise strategy. Although regulators cannot prevent users from interacting directly with smart contracts on the blockchain, the vast majority of ordinary users still rely on a web-based graphical user interface (the front end) to access these protocols [51]. Regulation can therefore focus on the operators of these front-end interfaces, requiring them to implement geo-blocking based on users' IP addresses or to integrate compliance-screening tools at the front-end level [52]. This "regulate the access point" strategy has already been widely adopted in practice; for example, the front ends of mainstream protocols such as Uniswap actively block access from sanctioned regions. Although advanced users can bypass these restrictions by calling contracts directly or using alternative front ends, front-end regulation effectively raises the barrier to non-compliant participation and brings the vast majority of non-professional users within the scope of protection.

Liquidity-provider obligations are an innovative regulatory idea designed for the features of the decentralized architecture. In protocols that adopt an automated market maker or liquidity-pool model, although traders may be transient and anonymous, the participants who supply large-scale liquidity are often long-lived entities with large amounts of capital [38]. Regulators can shift their approach: rather than attempting to identify every trader, they can impose compliance obligations on liquidity providers above a certain capital scale [53]. For example, they can require institutional-grade liquidity providers to conduct anti-money-laundering screening of the sources of their funds or to follow specific risk-management standards when providing liquidity. Because the on-chain behavior of liquidity providers is continuous and traceable, this strategy of focusing on large participants is far more feasible to enforce than comprehensive surveillance of all users.

26.6.4 Case study: Hyperliquid

Fully on-chain protocols, in responding to extreme market events, have displayed a complex interplay between decentralized governance and the logic of traditional regulation. The JELLY-token manipulation incident on Hyperliquid in March 2025 offers a highly valuable natural experiment for understanding this tension [54].

JELLY is a low-market-cap perpetual futures asset listed for trading on Hyperliquid. Exploiting the asset's scarce spot liquidity, the attacker established a large short position on Hyperliquid and then maliciously ramped up the price in the spot market, attempting to drain the protocol's liquidation fund pool [54]. Because JELLY's price was manipulated upward by more than 400% in a short time, and the order book lacked sufficient liquidity to absorb the forcibly liquidated short position, the position was transferred to the protocol's liquidity-provider vault, driving the vault's unrealized loss to about $13.5 million at one point (about $12 million when the validators intervened) [54][55]. In this moment of crisis, the protocol's built-in auto-deleveraging (ADL) mechanism failed to activate in time because of a flaw in the parameter setting of its trigger condition, and the solvency of the entire protocol faced a severe test [54].

Faced with systemic risk, Hyperliquid did not rely on the intervention of an external regulator but instead activated a decentralized-governance mechanism. The validator network quickly reached consensus and executed an emergency intervention through an

on-chain vote: they overrode the manipulated oracle price, forcibly closed the relevant positions at the reasonable pre-manipulation price, and compensated the affected legitimate long traders from the protocol's reserves (excluding addresses associated with the attack) [54]. At the same time, the validators voted to permanently delist the JELLY contract from the trading platform [56].

This incident reveals the potential and the limits of decentralized autonomy as an endogenous regulatory mechanism. On the one hand, it demonstrated that even without a centralized entity, a decentralized network still has the capacity to act quickly in a crisis, protect the system from malicious attack, and compensate innocent users. This crisis-handling mechanism, driven by validator consensus, in effect partly achieved the market-integrity and majority-user-protection objectives that traditional regulation pursues [57]. On the other hand, this act of intervening in contract execution by modifying the oracle price essentially broke the absolute rigidity of "code is law" and introduced the subjectivity of human judgment.

From the perspective of regulatory economics, the Hyperliquid case shows that decentralized governance is not the opposite of traditional regulation but a complementary institutional arrangement. When a protocol's architecture technically rules out the possibility of external enforcement, cultivating a healthy, fast-responding decentralized-governance mechanism—one that can internalize and enforce the basic principles of maintaining market fairness and stability—will be the most pragmatic choice for regulators confronting the decentralization paradox. Future regulatory frameworks should not seek to destroy this capacity for autonomy but should instead work to set reasonable boundaries and transparency requirements for decentralized governance, so that it can resist risk without degenerating into a tool through which a few validators abuse their power.

26.7 Empirical evidence on regulation

The economic analysis of regulatory intervention cannot remain at the level of theoretical deduction alone. The distinctiveness of the crypto-asset market lies in its high degree of data transparency and the fragmentation of the global regulatory framework, which provide a rich array of natural-experiment scenarios for evaluating the market-quality effects of different regulatory tools. Empirical evidence shows that the impact of regulation on market quality is not a one-way cost or benefit but exhibits pronounced nonlinearity and dimensional differentiation. Every regulatory intervention, while improving certain market-quality dimensions, necessarily comes at the cost of others. This section quantitatively assesses the specific impact of regulatory intervention on the five market-quality dimensions—transaction costs, price discovery, resilience, fairness, and accessibility—by analyzing the leverage limits of Japan and the European Union, exchange compliance requirements, and the enforcement actions of the SEC.

26.7.1 Empirical evidence on leverage limits

Leverage limits are the intuitive tool with which regulators address systemic risk, and their core logic is to interrupt the transmission path of the liquidation cascade by

reducing individual participants' risk exposure. The net effect of leverage limits on market quality, however, depends heavily on the parameter setting and the enforcement environment. Japan's FSA leverage-limit policy of 2020 provides a clear empirical sample of this trade-off. Before the limit was implemented, Japanese crypto-asset trading platforms generally offered leverage as high as 25x, and annual margin-trading volume once reached the scale of $500 billion [58]. In May 2020, the FSA forcibly reduced the maximum leverage on margin trading to 2x, and this policy shock directly reshaped the microstructure of the Japanese crypto market.

Figure 26-8 uses two panels to show the change in market quality before and after Japan's leverage-limit policy. On the accessibility and liquidity dimension, average monthly margin-trading volume collapsed from about $42 billion before the limit to about $5 billion per month by the end of 2020, and market depth shrank significantly; on the fairness and resilience dimension, the overall profitability rate of retained users rose from about 17% to 32%. Two points must be emphasized here. First, this profitability figure is cited by the author from [59], has not yet undergone independent public verification, and should be regarded as cited evidence rather than a verified empirical conclusion. Second, this increase stems mainly from a composition effect—the change in the makeup of market participants after high-risk retail traders exited—rather than from an improvement in the profitability of any individual trader. The essence of this trade-off is that regulation, at the cost of liquidity and retail accessibility, weeded out the most fragile market participants, thereby improving the system's overall resilience and traders' average survival rate.

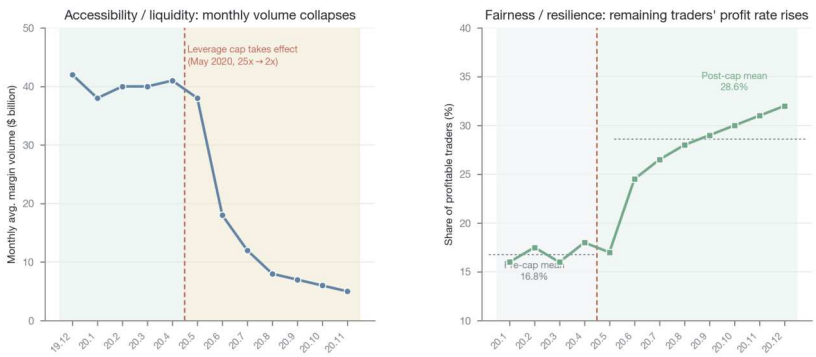


Figure 26-8. Change in market quality before and after Japan's 2020 leverage limit

Similar trade-offs also appeared in the regulatory practice of Europe and South Korea. Through its product-intervention measures for CFDs, ESMA strictly limited retail leverage on crypto-derivatives to 2x, which placed European domestic perpetual futures providers at a significant disadvantage in competing with offshore platforms offering 10x to 100x leverage [60]. South Korea's regulatory path was more aggressive: in September 2025, its Financial Services Commission (FSC) issued guidelines on crypto-asset lending that prohibit leveraged borrowing in excess of collateral value [61]. At

the same time, South Korea's long-standing controls on cross-border capital flows have impeded the free entry and exit of arbitrage capital, isolating the domestic market from the global liquidity pool and producing a structural pricing-island phenomenon, the so-called "kimchi premium" [62]. Together, these natural experiments demonstrate that, absent global coordination, unilateral leverage limits often lead to a geographic reconfiguration of liquidity rather than a fundamental elimination of systemic risk.

26.7.2 The market-quality effects of compliance requirements

An exchange's compliance status and choice of domicile directly determine the strength of the regulatory constraints it faces and, in turn, systematically affect its market-quality performance. In traditional financial-microstructure theory, stricter regulation is generally thought to improve market integrity and reduce adverse-selection costs. In the crypto-asset market, however, the relationship between compliance requirements and liquidity is more complex and dynamic. Jurisdictions that require exchanges to implement strict asset segregation, customer identity verification, and anti-money-laundering procedures often face pressure from the migration of trading volume to more lightly regulated regions.

Empirical research shows that exchanges registered in different jurisdictions exhibit significant divergence in market-quality indicators. Taking global platforms such as Binance and OKX as examples, their liquidity-provision patterns under different regulatory environments reflect the pass-through mechanism of compliance costs [63]. In markets with strict regulatory requirements, market makers face higher costs of tied-up capital and compliance-reporting burdens, and these costs are ultimately internalized as wider bid-ask spreads and thinner order-book depth. Conversely, in more lightly regulated offshore markets, lower entry barriers attract a large number of high-frequency traders and arbitrage capital, forming a highly competitive transaction-cost environment.

This liquidity stratification, driven by regulatory arbitrage, embodies the dual effect of compliance requirements. On the one hand, strict registration requirements and asset-segregation rules significantly improve market resilience and fairness and reduce the tail risk of an FTX-like collapse. On the other hand, high compliance costs constitute a substantive barrier to market entry, causing liquidity to concentrate among the few large compliant institutions that can bear these costs. This regulation-driven market concentration not only weakens competition but may also aggravate the fragility of liquidity supply under extreme market conditions. When compliant market makers retreat in unison during a crisis because of regulatory constraints or capital requirements, liquidity in the compliant market may even dry up faster than in the non-compliant market.

26.7.3 The cost of regulatory uncertainty

In assessing the economic consequences of regulation, one frequently overlooked dimension is the friction cost of regulatory uncertainty itself. When regulatory rules are ambiguous, classification standards unclear, or reliance placed on ex post enforcement rather than ex ante guidance, market participants must pay an additional premium for

the potential compliance risk. The enforcement actions the SEC took against major crypto-asset exchanges in 2023 provide an ideal window for quantifying this cost of regulatory uncertainty.

In June 2023, the SEC sued Binance and Coinbase in quick succession, alleging that they offered securities-trading services without registration. The event study by Saggu et al. shows that this strategy of establishing regulatory boundaries through enforcement caused immediate and far-reaching damage to market liquidity [64]. The event study, based on a full sample of the SEC's series of enforcement actions, finds that in the first three days after the enforcement announcements, the prices of the affected crypto-assets fell by an average of about 5.2%, with the decline widening further to about 17.2% within 30 days. More critically, trading volume dropped sharply, indicating that this was not merely a repricing of prices but a structural withdrawal of liquidity.

The damage that regulatory uncertainty does to market microstructure operates mainly through the defensive retreat of market makers. Facing ambiguous asset-classification standards and potential legal sanctions, the rational choice for compliant market makers is to widen spreads or stop providing liquidity for the relevant assets outright. The research finds that assets with smaller market capitalization and poorer initial liquidity suffered more severely under this regulatory shock, with the widening multiple of their bid-ask spreads significantly higher than that of the leading assets [64]. This liquidity evaporation, triggered by regulatory uncertainty, leaves the market highly fragile when facing an external shock. The empirical results indicate that, for market quality, an unclear regulatory framework may cause greater harm than clear but strict regulatory rules. Under ambiguous rules, the risk aversion of investors and market makers is amplified, leading to the inefficient idling of capital and the stalling of the price-discovery mechanism.

26.7.4 A comprehensive assessment of five-dimensional market quality

Based on the evidence from the preceding empirical research and natural experiments, we can systematically assess the comprehensive impact of the various regulatory tools on crypto-asset market quality. Table 26-3 aggregates the existing empirical findings and maps them onto the five-dimensional market-quality assessment framework proposed in this chapter. This matrix is not only a summary of historical data but also provides an economic benchmark for future regulatory design.

Table 26-3 reveals a core policy dilemma: in the existing regulatory toolbox, no single intervention can produce a purely positive effect across all five market-quality dimensions. Leverage limits greatly strengthen the system's resilience and protect inexperienced retail participants but inevitably harm market accessibility and overall liquidity. Mandatory customer identity verification and anti-money-laundering compliance improve market fairness and legitimacy but likewise come at the cost of accessibility and higher transaction costs. Disclosure and asset-segregation requirements are among the few regulatory tools with a high net benefit: while significantly improving fairness and resilience, they have relatively small negative spillovers to the other dimensions.

Type of regulatory intervention	Transaction costs	Price discovery	Resilience	Fairness	Accessibility	Empirical basis
Strict leverage limit (e.g., reduced to 2x)	Significant increase	Slight weakening	Significant improvement	Significant improvement	Significant decline	Japan FSA data [58][59]
Mandatory asset segregation and audit	Slight increase	Unchanged	Significant improvement	Significant improvement	Unchanged	Post-FTX comparative study [63]
Strict KYC/AML compliance	Significant increase	Unchanged	Unchanged	Significant improvement	Significant decline	Spread comparison between offshore and compliant platforms [63]
Regulatory uncertainty (enforcement actions)	Significant increase	Significant weakening	Significant decline	Unchanged	Slight decline	SEC litigation event study [64]
Isolationist policy (capital controls)	Significant increase	Significant weakening	Slight improvement	Unchanged	Significant decline	South Korea's "kimchi premium" phenomenon [62]

Table 26-3. Summary of the empirical effects of regulatory intervention on five-dimensional market quality (Data source: compiled by the author based on [58][59][62][63][64])

The ultimate lesson of this empirical matrix is that optimal regulatory design does not seek the absolute maximization of any single dimension but instead makes wise trade-offs among the five dimensions that suit specific policy objectives. If the regulator's primary objective is to guard against a systemic collapse, then accepting lower liquidity and higher transaction costs is the price that must be paid. If regulation relies excessively on ex post enforcement or imposes technically incompatible requirements, however, the regulatory uncertainty and friction costs it generates will cause market quality to deteriorate simultaneously across all dimensions. Future regulatory frameworks must therefore move from moral judgment toward cost-benefit analysis grounded in empirical data, seeking room for marginal improvement while acknowledging technical constraints.

26.8 The political economy of regulation

The analytical frameworks of the preceding sections carry an important implicit assumption: that the regulator is a benevolent social planner aiming to maximize social welfare. In reality, however, regulatory rules are not formulated in a vacuum but are the political-economy equilibrium of a game among many interests. The systematic deviation between the theoretical optimum and actual policy is precisely the object of this section's analysis. Understanding the direction of crypto-market regulation cannot proceed from the theoretical model of market failure alone; it must examine who has an incentive to push for regulation, who has an incentive to block it, and who has the ability to exert substantive influence over the rule-making process. At the intersection of decentralized finance and traditional finance, the final form of regulation is often the product of mutual compromise among stakeholders rather than the theoretical optimum in the economic sense. This section applies the theory of regulatory capture to analyze the evolution of the crypto industry's political influence, maps the stakeholders, and explores how the mechanism of regulatory competition among jurisdictions determines the ultimate quality of regulation.

26.8.1 The theory of regulatory capture

The theory of regulatory capture was advanced in 1971 by the Nobel laureate in economics George Stigler; its core thesis is that a regulatory agency will ultimately be controlled by the industry it regulates, causing regulatory rulemaking to favor industry interests over the public interest [65]. Peltzman (1976) developed this theory further, proposing that regulators are not simply captured by a single interest group but instead seek an equilibrium that maximizes political support between producers and consumers—a perspective that provides a theoretical basis for understanding the multi-stakeholder game in Section 26.8.2 [66]. In the evolution of the crypto market, this theory has attracted a growing body of empirical observation consistent with its predictions. As industry capital has accumulated, crypto firms have shifted from initially evading regulation to actively shaping regulatory rules.

The crypto industry's influence over the political process is most directly reflected in the exponential growth of its lobbying spending. Before 2020, the industry's lobbying spending in Washington was almost negligible. As the market grew and compliance pressure mounted, however, industry resources began to tilt on a large scale toward the political arena. In 2022, the crypto industry's federal lobbying spending reached $21.6 million, 2.6 times that of 2021 and 8.6 times that of 2020 [67]. By 2023, despite multiple major industry crises, this figure still set a new record of nearly $25 million, with Coinbase and the Blockchain Association among the largest spenders [68].

Figure 26-9 tracks the exponential growth of the crypto industry's federal lobbying spending from 2018 to 2024. Even amid multiple major industry crises in 2022 (the Terra/Luna collapse and the FTX bankruptcy), lobbying spending rose rather than fell—a sign that in a crisis industry participants step up their investment in the regulatory agenda, competing for a voice while the rules are being rewritten.

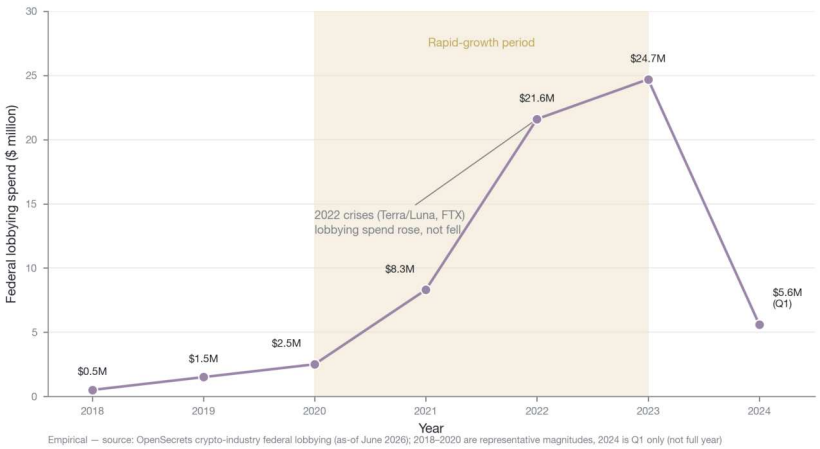


Figure 26-9. Annual trend of the crypto industry's federal lobbying spending

FTX and its founder, Sam Bankman-Fried, offer a classic instance of regulatory capture in the crypto domain. Before the company's collapse, Bankman-Fried was one of the largest donors in U.S. politics and, through his team, established direct contact with financial regulators at the highest level, seeking to advance the Digital Commodities Consumer Protection Act, which would have favored his platform [69]. This influence rested on more than capital alone; it also ran through the subtler "revolving door" effect, whereby crypto firms hired former regulatory officials and government staff at high salaries to work for their new employers using the connections and rule knowledge they had accumulated within the system.

This deep political engagement has produced an endogenous bias in regulatory policy. When the regulated hold an informational and resource advantage, they can steer legislators' attention toward specific types of risk while obscuring other risks that might touch their core business model. The observed regulatory framework therefore often reflects the strategic intentions of the dominant firms within the industry rather than an objective outcome based on a systemic risk assessment.

26.8.2 Stakeholder analysis

The driving force behind the evolution of crypto regulation originates in the ongoing game among different interest groups. In this complex ecosystem, the motives and strategies of each party determine the resistance to, and the direction of, policy progress.

The forces pushing for strict regulation consist mainly of traditional financial institutions, consumer-protection advocates, and victim groups. Traditional financial institutions play a dual role in this game: on the one hand, they view unregulated crypto platforms as a threat of unfair competition and demand that regulators apply the principle of "same business, same risk, same rules" to eliminate the space for regulatory arbitrage; on the other hand, they also hope that a clear compliance framework will remove the compliance barriers to their own entry into this emerging market. Consumer-protection advocates and the retail groups that suffered losses in market collapses constitute the social pressure pushing for regulation, demanding strict asset segregation, leverage limits,

and disclosure requirements to prevent a recurrence of an FTX-like episode.

The forces that block or seek to weaken regulation are led by crypto-native firms, industry lobbying groups, and advocates of technological innovation. The strategy of this camp is not to resist regulation outright but to advocate a "tailored," light-touch regulatory framework. They emphasize the special nature of decentralized technology, arguing that imposing traditional financial regulations designed for intermediaries on smart contracts would suppress innovation and force the industry to move overseas. Offshore jurisdictions, as this camp's implicit allies, attract capital and talent by offering a relaxed regulatory environment, which in effect raises the cost to major economies of imposing strict unilateral regulation.

Between these two opposing forces, institutional investors and large asset managers constitute a key middle force. This group manages vast amounts of capital and is eager to include crypto-assets in its portfolios but, constrained by its own compliance requirements, must operate in a regulated environment. It therefore neither wishes to see over-regulation that could shrink the market nor accepts complete laissez-faire that lacks basic property-rights protection. This demand for "moderate compliance" often becomes the decisive factor in bringing a pragmatic regulatory framework into being.

26.8.3 Regulatory competition

Because of crypto-assets' borderless nature and digital liquidity, regulatory jurisdiction faces a significant arbitrage challenge. Capital and business can cross borders easily, which puts countries and regions in direct competition over crypto-regulatory rules [70]. This competitive mechanism has a significant impact on the quality of global regulation.

In the initial stage of competition, the market widely feared a "race to the bottom." Under this pattern, in order to attract crypto firms and gain tax revenue and employment, some jurisdictions might competitively lower their compliance thresholds and relax their consumer-protection and anti-money-laundering requirements. This regulatory arbitrage not only weakens the overall safety of the global financial system but may also force countries that originally held to high standards to compromise under the pressure of capital outflow.

As the market matures and institutional investors enter, however, a "race to the top" has begun to emerge. After a series of trust crises at offshore platforms, high-quality crypto firms and large institutional investors increasingly prefer to operate in jurisdictions with a clear, rigorous, yet adaptive regulatory framework. This compliance premium has prompted financial centers such as Dubai, Singapore, and Hong Kong to compete in establishing comprehensive and clear virtual-asset regulatory regimes [71].

These emerging crypto hubs have not chosen to abandon regulation but instead compete by offering regulatory certainty. Hong Kong, for example, implemented a new licensing regime for virtual-asset trading platforms in 2023, setting extremely high requirements on asset segregation, conflict-of-interest management, and cybersecurity [72]. This competitive pattern shows that, in a crypto market with severe information asymmetry, high-quality regulation has itself become a competitive advantage: it provides reputational endorsement for the firms within a jurisdiction and lowers the

friction costs of connecting with the traditional financial system.

26.8.4 Factors affecting regulatory quality

Having identified the political-economy drivers of regulation, the core question turns to how to assess and improve the actual effectiveness of regulatory intervention. The effectiveness of regulation is not simply equivalent to the strictness of the rules but depends on a series of determinants of institutional design and enforcement capacity.

The independence and technical capacity of the regulator are the foremost factors determining regulatory quality. In a domain as highly complex and rapidly iterating as crypto, if a regulator lacks sufficient resources and expertise to understand the underlying technical mechanisms, it can easily fall into the trap of "formal compliance"—in which firms superficially meet all reporting requirements but systemic risk is not, in fact, effectively controlled. At the same time, an agency's independence determines whether it can resist the regulatory-capture pressures described above and maintain an objective balance between industry interests and the public interest.

The paradigm choice in regulatory design directly affects the adaptability of the rules. Rules-based regulation often struggles to keep pace with the innovation of smart contracts and decentralized finance and readily leads to obsolete rules or new arbitrage space. By contrast, a principles-based regulatory framework, combined with the concept of functional regulation, can better see through the technological surface and strike at the essence of financial risk. This design allows regulators to remain flexible about the technical implementation path while holding to core bottom lines (such as customer-asset segregation).

In addition, the degree of international coordination determines the size of the space for regulatory arbitrage. Absent a globally unified standard, any unilateral high-quality regulation may see its effect significantly reduced by business migration. Effective regulation must not only balance the interests of all parties domestically but also build a minimum international consensus on guarding against core risks. Only when improving regulatory quality no longer comes at the cost of a jurisdiction's competitiveness can the crypto market truly escape the prisoner's dilemma of regulatory arbitrage and achieve a long-term equilibrium between innovation and stability.

26.9 Chapter summary

Through the analytical framework of regulatory economics, this chapter has systematically answered the core questions of why to regulate, what to regulate, how to regulate, and who should regulate, and it has proposed an optimal regulatory framework based on net-benefit maximization.

The legitimacy of regulatory intervention rests on market failure, not on the market itself. The externalities, information asymmetry, undersupply of public goods, and market power in the perpetual futures market constitute four main categories of failure, but failure alone does not automatically constitute a sufficient condition for intervention—competitive mechanisms and technological innovation can repair some defects on their own, and intervention is economically justified only when the expected benefits of

regulation exceed costs such as the compliance burden, innovation suppression, and regulatory arbitrage [73]. Four conclusions can be drawn on this basis. First, asset segregation is the single tool with the highest net benefit: FTX's roughly $8 billion in customer-asset losses exposed the risk of misappropriation, and mandatory asset segregation combined with regular external audit can significantly improve resilience and fairness at relatively low compliance cost, which is why it ranks first in priority [10]. Second, functional regulation is superior to entity-based regulation: the cross-boundary nature of perpetual futures makes them hard to classify simply as a security or a commodity, whereas the functional principle of same activity, same risk, same regulation shifts the focus from institution type to specific economic function and better adapts to the rapid iteration of financial innovation [37]. Third, the cost of regulatory fragmentation may exceed that of regulation itself: the fundamental contradiction between a globalized market and fragmented sovereign jurisdiction produces, through the impossible trinity, a fragmented global regulatory map that pushes trading activity toward arbitrage space. Fourth, regulatory quality matters more than regulatory strength: disclosure and transparency requirements are the lowest-cost tools with the smallest side effects, and a well-designed framework should give priority to transparency mechanisms that promote market self-discipline.

At the practical level, an optimal regulatory framework should focus on guarding against tail risk and protecting market integrity, allocating regulatory resources first to areas such as asset segregation, reserve audit, and anti-manipulation. For centralized exchanges, compliance is no longer merely a passive cost of responding to external pressure but a strategic asset for building core competitiveness. For investors, understanding the boundaries of regulation's capability is a prerequisite for managing one's own risk, and investors need to develop the ability to assess counterparty risk independently.

Regulation plays a critical role in improving the defensive deficiencies of the perpetual futures market, but the boundary of its capability is equally clear. Legal and administrative means can mandate disclosure, set leverage caps, and regulate market-maker behavior, but they cannot change the physical mechanism of the liquidation cascade, nor can they eliminate the inherent trade-offs in a decentralized architecture. These domains that regulation cannot reach are precisely the space in which technical solutions come into play. Smart-contract audits, on-chain transparency mechanisms, and automated risk-control protocols are building an entirely new paradigm of embedded compliance [74], in which rules are hard-coded into the protocol's execution logic, thereby lowering the friction cost of compliance. Technology is not a substitute for regulation but its necessary complement. Having understood the economic logic of regulation and its limitations, the next chapter turns to a thought experiment, exploring the potential and the boundaries of perpetual futures as a universal pricing infrastructure.

References

[1] Coinbase Institutional. (2024). *A Primer on Perpetual Futures.* Coinbase Market Intelligence, June 10, 2024 (Data source: CoinMetrics and Coinbase; 30-day rolling average, weekday basis). Reprinted via Richey May: https://richeymay.com/resource/articles/unlocking-the-power-of-perpetual-futures-cryptos-game-changing-trading-

tool/

[2] Kavout. (2025). What Are Crypto Perpetual Futures and Why Do They Matter. https://www.kavout.com/market-lens/what-are-crypto-perpetual-futures-and-why-do-they-matter

[3] Securities and Exchange Commission. (2019). Framework for "Investment Contract" Analysis of Digital Assets. https://www.sec.gov/files/dlt-framework.pdf

[4] Ernst & Young. (2023). Crypto derivatives are becoming a major digital asset class. https://www.ey.com/content/dam/ey-unified-site/ey-com/en-us/insights/financial-services/documents/ey-crypto-derivatives-pov_final4.pdf

[5] Chainalysis. (2020). *East Asia: Pro Traders and Stablecoins Drive World's Biggest Cryptocurrency Market*. (Net outflow of over $50 billion in crypto-assets from East Asian addresses during 2019–2020, of which approximately $18 billion was USDT.) As recounted by the World Economic Forum (2022): https://www.weforum.org/stories/2022/01/what-s-behind-china-s-cryptocurrency-ban/

[6] BBC News. (2024). Customers of failed crypto firm FTX set for refunds. https://www.bbc.com/news/articles/c0qz3dg21vqo

[7] Stiglitz, J. E. (2024). *Government Failure vs. Market Failure*. Columbia University Academic Commons. https://ipdcolumbia.org/wp-content/uploads/2024/08/Govt_Failure_vs_Market_Failure_Stiglitz.pdf

[8] Auh, J. K., & Cho, W. (2023). Liquidation cascade and anticipatory trading: Evidence from the structured equity product market. *The Review of Asset Pricing Studies*, *13*(1), 53–94. https://doi.org/10.1093/rapstu/raac010

[9] Ali, Z. (2025). *Anatomy of the Oct 10–11, 2025 Crypto Liquidation Cascade: Macroeconomic Triggers, Market Microstructure, and Systemic Risk Lessons*. SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5611392 https://doi.org/10.2139/ssrn.5611392

[10] DeKlotz, A. (2023). *FTX and Failed Custody Rules: The Collapse that Shocked the World*. SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4544930 https://doi.org/10.2139/ssrn.4544930

[11] Jalan, A., & Matkovskyy, R. (2024). Systemic risks in the cryptocurrency market: Evidence from the FTX collapse. *Finance Research Letters*, *53*(C), 103632.

[12] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201-2238. https://doi.org/10.1093/rfs/hhn098

[13] Dell'Erba, M., & Vianelli, A. (2025). Crypto-Derivatives Regulation(s). *Cornell International Law Journal*. SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5895082 https://doi.org/10.2139/ssrn.5895082

[14] Crain, N. V., & Crain, W. M. (2023). *The Cost of Federal Regulation to the U.S. Economy, Manufacturing and Small Business*. National Association of Manufacturers. https://www.nam.org/wp-content/uploads/2023/11/NAM-3731-Crains-Study-R3-V2-FIN.pdf

[15] Bombardini, M., Trebbi, F., & Zhang, M. B. (2024). *Measuring the Costs and Benefits of Regulation* (NBER Working Paper No. 32955). National Bureau of Economic Research. https://www.nber.org/papers/w32955

[16] Coates, J. C. (2014). Cost-benefit analysis of financial regulation: Case studies and implications. *Yale Law Journal*, *124*, 882–1011.

[17] Wang, S., & Zhang, T. (2025). *Spot-futures manipulations in cryptocurrency markets* (SSRN Working Paper 5125326). https://ssrn.com/abstract=5125326

[18] Cong, L. W., Li, X., Tang, K., & Yang, Y. (2023). Crypto wash trading. *Management Science*, *69*(11), 6427–6454. https://doi.org/10.1287/mnsc.2021.02709

[19] European Securities and Markets Authority. (2018). *Additional information on the agreed product intervention measures relating to contracts for differences and binary options* (ESMA35-43-1000). https://www.esma.europa.eu/sites/default/files/library/esma35-43-1000_additional_information_on_the_agreed_product_intervention_measures_relating_to_contracts_for_differences_and_binary_options.pdf

[20] Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

[21] Commodity Futures Trading Commission. (2022). *Complaint: FTX Trading, Ltd., et al.* https://www.cftc.gov/media/8021/enfftxtradingcomplaint122122/download

[22] Lazirko, M., Appelbaum, D., & Vasarhelyi, M. (2025). Proof of reserves: A double-helix framework. *The British Accounting Review*, *57*(3), 101389.

[23] Aramonte, S., Huang, W., & Schrimpf, A. (2021). DeFi risks and the decentralisation illusion. *BIS Quarterly Review*, 21-36.

[24] Kahraman, B., & Tookes, H. E. (2017). Margin trading and leverage management. *Journal of Financial and Quantitative Analysis*, *52*(2), 481-512.

[25] Japan Cryptoasset Business Association. (2023). *Request for Revision of Leverage in Crypto Asset Margin Trading*. https://cryptocurrency-association.org/news/mainfo/20231017-01/

[26] Auer, R., Cornelli, G., Doerr, S., Frost, J., & Gambacorta, L. (2023). Crypto trading and Bitcoin prices: evidence from a new database of retail adoption. *BIS Working Papers*, No 1049.

[27] Bessembinder, H., Hao, J., & Lemmon, M. (2015). Why designate market makers? Affirmative obligations and market quality. *Journal of Financial Economics*, *116*(3), 536-549. https://doi.org/10.2139/ssrn.1007274

[28] Arner, D. W., Zetzsche, D. A., & Buckley, R. P. (2023). The financialization of crypto: Lessons from FTX and the crypto winter of 2022-2023. *Available at SSRN 4372516*. https://doi.org/10.2139/ssrn.4372516

[29] Securities and Exchange Commission. (2025). *Statement on the Custody of Crypto Asset Securities by Broker-Dealers*. https://www.sec.gov/newsroom/speeches-statements/trading-markets-121725-statement-custody-crypto-asset-securities-broker-dealers

[30] Aizenman, J. (2010). The impossible trinity (aka the policy trilemma). *Encyclo-*

pedia of Financial Globalization, 1-9.

[31] Chen, C., & Liu, L. (2022). How effective is China's cryptocurrency trading ban?. *Finance Research Letters*, *46*, 102429. https://doi.org/10.1016/j.frl.2021.102429

[32] Financial Stability Board. (2025). *Thematic Review on FSB Global Regulatory Framework for Crypto-asset Activities*. https://www.fsb.org/2025/10/thematic-review-on-fsb-global-regulatory-framework-for-crypto-asset-activities/

[33] International Organization of Securities Commissions. (2023). *Policy Recommendations for Crypto and Digital Asset Markets*. https://www.iosco.org/library/pubdocs/pdf/IOSCOPD747.pdf

[34] SEC v. W.J. Howey Co., 328 U.S. 293 (1946). United States Supreme Court.

[35] Selig, J. (2026). *The Next Phase of Project Crypto: Unleashing Innovation*. Commodity Futures Trading Commission. https://www.cftc.gov/PressRoom/SpeechesTestimony/opaselig1

[36] Merton, R. C., & Bodie, Z. (2004). *The Design of Financial Systems: Towards a Synthesis of Function and Structure* (Working Paper 10620). National Bureau of Economic Research. https://www.nber.org/papers/w10620

[37] Borio, C., Claessens, S., & Tarashev, N. (2022). *Entity-based vs activity-based regulation: a framework and applications to traditional financial firms and big techs* (FSI Occasional Paper No 19). Bank for International Settlements. https://www.bis.org/fsi/fsipapers19.pdf

[38] Heimbach, L., Wang, Y., & Wattenhofer, R. (2021). Behavior of liquidity providers in decentralized exchanges. *arXiv preprint arXiv:2105.13822*. https://doi.org/10.48550/arXiv.2105.13822

[39] dYdX Foundation. (2023). *v4 Technical Architecture Overview*. https://www.dydx.xyz/blog/v4-technical-architecture-overview

[40] GMX Contributors. (2023). *GMX: Decentralized Perpetual Exchange*. https://gmx.io/

[41] Capponi, A., & Jia, R. (2025). Liquidity provision on blockchain-based decentralized exchanges. *The Review of Financial Studies*, *38*(10), 3040-3080. https://doi.org/10.1093/rfs/hhaf046

[42] Hyperliquid Contributors. (2024). *Hyperliquid Docs: Order Book Architecture*. https://hyperliquid.gitbook.io/hyperliquid-docs/trading/order-book

[43] Ferreira, A. (2024). Decentralized finance (DeFi): the ultimate regulatory frontier? *Capital Markets Law Journal*, *19*(3), 242-260.

[44] Lorenz, G. (2024). Regulating decentralized financial technology: A qualitative study on the challenges of regulating DeFi with a focus on embedded supervision. *Stanford Journal of Blockchain Law & Policy*, *7*, 10-45.

[45] Financial Crimes Enforcement Network. (2023). *Illicit Finance Risk Assessment of Decentralized Finance*. US Department of the Treasury.

[46] Oye, E. (2025). *The Role of Decentralized Exchanges (DEXs) in Money Laundering Schemes* (working paper / preprint). ResearchGate.

[47] Bourveau, T., De George, E. T., & Macciocchi, D. (2024). Decentralized Finance (DeFi) assurance: early evidence. *Review of Accounting Studies*, 1-35. https://doi.org/10.1007/s11142-024-09834-8

[48] Priem, R. (2025). Regulatory Sandboxes for Security Tokens Trading: A Policy Instrument for Diverting Market Activity from Decentralised Exchanges (DEXs)? *SSRN Electronic Journal*.

[49] Landsman, W. R., Lyandres, E., Maydew, E. L., Rabetti, D., & Zhang, C. (2025). *Auditing Smart Contracts* (SSRN Working Paper No. 5198563). SSRN Electronic Journal.

[50] Rabetti, D. (2023). *Auditing Decentralized Finance (DeFi) Protocols* (SSRN Working Paper No. 4458298). SSRN Electronic Journal.

[51] Legal Nodes. (2024). *Legal Liability of Front-End Operators in Decentralized Finance*. https://legalnodes.com/article/front-end-operator-legal-liability

[52] Financial Action Task Force. (2021). *Updated Guidance for a Risk-Based Approach to Virtual Assets and Virtual Asset Service Providers*. FATF. https://www.fatf-gafi.org/content/dam/fatf/documents/recommendations/Updated-Guidance-VA-VASP.pdf

[53] Aquilina, M., Frost, J., & Schrimpf, A. (2024). Decentralised dealers? examining liquidity provision in DEXs. *Bank for International Settlements Working Papers*, No. 1227.

[54] OAK Research. (2025). *Hyperliquid and the JELLY attack: Context, vulnerability and team solution*. https://oakresearch.io/en/analyses/investigations/hyperliquid-jelly-attack-context-vulnerability-team-solution

[55] CoinDesk. (2025). *HyperLiquid Delists JELLY After Vault Squeezed in $13M Tussle*. https://www.coindesk.com/markets/2025/03/26/hyperliquid-delists-jellyjelly-after-vault-squeezed-in-usd13m-tussle

[56] Halborn. (2025). *Explained: The Hyperliquid Hack (March 2025)*. https://www.halborn.com/blog/post/explained-the-hyperliquid-hack-march-2025

[57] Wright, A., & De Filippi, P. (2015). Decentralized blockchain technology and the rise of lex cryptographia. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2580664

[58] Presto Research. (2024). State of the Japanese Crypto Market. https://www.prestolabs.io/research/state-of-the-japanese-crypto-market

[59] Chainalysis. (2020). Japan's Cryptocurrency Regulation Updates: What You Need to Know. https://www.chainalysis.com/blog/japan-cryptocurrency-regulation-updates-2020/

[60] Finance Magnates. (2026). Has Europe Killed Crypto Perps Even before It Started? https://www.financemagnates.com/forex/10x-down-to-2x-has-europe-killed-crypto-perps-even-before-it-started/

[61] Financial Services Commission (Korea). (2025). *Guidelines on Virtual Asset Lending Services*. (Issued September 2025; prohibits leveraged borrowing in excess of collateral value, with a 20% cap on the lending rate.)

[62] Kaiko Research. (2024). Isolationary regulation, a double-edged sword. https://research.kaiko.com/insights/isolationary-regulation-a-double-edged-sword

[63] Feinstein, B. D., & Werbach, K. (2021). The Impact of Cryptocurrency Regulation on Trading Markets. *Journal of Financial Regulation*, *7*(1), 48–99. https://doi.org/10.1093/jfr/fjab003

[64] Saggu, A., Ante, L., & Kopiec, K. (2025). Uncertain Regulations, Definite Impacts: The U.S. SEC's Regulatory Interventions and Their Effects on Crypto Assets. *Oxford Business Law Blog*. https://blogs.law.ox.ac.uk/oblb/blog-post/2025/01/unintended-consequences-sec-crypto-enforcement-actions

[65] Stigler, G. J. (1971). The theory of economic regulation. *The Bell Journal of Economics and Management Science*, *2*(1), 3-21.

[66] Peltzman, S. (1976). Toward a more general theory of regulation. *Journal of Law and Economics*, *19*(2), 211-240. https://doi.org/10.3386/w0133

[67] Giorno, T., & Sayki, I. (2023). Cryptocurrency industry lobbying and political contributions skyrocketed in 2022. *OpenSecrets News*. https://www.opensecrets.org/news/2023/03/cryptocurrency-industry-lobbying-and-political-contributions-skyrocketed-in-2022

[68] Reklaitis, V. (2024). Crypto industry's lobbying hit a record in 2023 — and is off to a strong start in 2024. *MarketWatch*. https://www.marketwatch.com/story/crypto-industrys-lobbying-hit-a-record-in-2023-and-is-off-to-a-strong-start-in-2024-fa740b43

[69] Yaffe-Bellany, D., & Flitter, M. (2023). FTX Chief Once Met With Powell. Now D.C. Crypto Lobbyists Are Lying Low. *The New York Times*. https://www.nytimes.com/2023/09/27/business/economy/crypto-lobbying-washington-ftx.html

[70] Yablon, C. M. (2024). Crypto: The Prospect of Regulatory Competition. *Cardozo Law Review*, *45*(4), 1435-1489.

[71] Huang, R. H., Yang, D., & Loo, F. F. Y. (2020). The development and regulation of cryptoassets: Hong Kong experiences and a comparative analysis. *European Business Organization Law Review*, *21*(2), 319-347.

[72] Burgess, T., & Liu, J. (2025). A tale of two jurisdictions: Contrasting cryptocurrency regulations in Hong Kong and the United Kingdom. *Journal of Economic Criminology*, *10*, 100026. https://doi.org/10.1016/j.jeconc.2025.100150

[73] Posner, E., & Weyl, E. G. (2013). Benefit-cost analysis for financial regulation. *American Economic Review*, *103*(3), 393–397. https://doi.org/10.1257/aer.103.3.393

[74] Chainlink. (2026). *Payments compliance in the blockchain era*. https://chain.link/article/payments-compliance-blockchain

A Thought Experiment on Perpetual Futures Reshaping Global Asset Pricing

177 years ago, the birth of the standardized delivery futures contract shifted pricing power over agricultural commodities from a fragmented network of spot markets to a centralized exchange. 177 years later, perpetual futures — synthetic derivatives with no expiration date — have become the highest-volume financial product in crypto markets, with average daily notional turnover reaching the scale of hundreds of billions of dollars. For the major crypto assets, perpetual futures are already the dominant force in price discovery. Now consider a counterfactual history: what if, in 1848, the Chicago Board of Trade had launched not delivery futures but perpetual futures on grain? How would history have differed? With no expiration date, farmers and buyers would no longer be troubled by the roll costs of quarterly contracts; with around-the-clock trading, participants across the world's time zones could trade at any moment; with a funding rate mechanism, the perpetual price would be automatically anchored to spot without the complex logistics of physical delivery; and with global access, anyone in the world with a communications connection could take part. In 1848 there was no internet, no blockchain, and no global communications network, so perpetual futures were technically impossible. Today, however, all of these technological preconditions are in place [1].

The previous two chapters diagnosed the current state of the perpetual futures market. Chapter 25 used a five-dimensional model of market quality to characterize its asymmetric profile — strong on offense but weak on defense — and Chapter 26 analyzed the extent to which regulation can repair these defensive deficiencies. The central question of this chapter is the following: if we set these constraints aside for the moment, and if the defensive deficiencies are progressively repaired while the offensive advantages are fully released, where does the ultimate potential of perpetual futures as a financial instrument lie? Can it transcend crypto-native assets to become a general-purpose global pricing infrastructure? When perpetual futures on any asset can be traded by anyone, anywhere, at any moment, what happens to global asset pricing? The core mechanism of perpetual futures, funding rate anchoring, does not depend on any special properties of the underlying asset; it is a general-purpose mechanism for generating synthetic exposure [2]. In his theory of macro markets, Robert Shiller argued that derivatives markets should be created for all important economic variables so that risk can be shared more

efficiently [3]. Perpetual futures may be precisely the suitable mechanism for realizing this vision, because they are a low-friction instrument for synthetic exposure. When a faster, more continuous, and more accessible pricing venue emerges, the pricing power of incumbent venues comes under structural pressure to migrate. The speed and limits of this migration depend on the new venue's ability to close the trust gap and on the progress of institutional recognition.

This chapter explores the possibilities and boundaries of perpetual futures as a general-purpose pricing layer through three progressively more ambitious thought experiments. The first examines the world's largest market, foreign exchange, and considers how foreign exchange perpetual futures could address structural inefficiencies such as fragmented trading sessions, settlement delays, and access barriers, while confronting the ultimate obstacles of central bank monetary policy transmission and anti-money-laundering compliance. The second turns to equity indices, analyzing the potential of index perpetual futures as a permissionless alternative to exchange-traded funds, and building a scenario-analysis framework to examine the price-transmission thresholds at which single-stock perpetual futures move from insignificance to systemic influence. The third examines the technical feasibility and institutional obstacles of turning commodities and emerging assets into perpetual futures — from crude oil and gold to carbon credits and real estate indices, and ultimately to prediction markets as the limiting form. Through these thought experiments, we sketch four stages in the migration of pricing power: from the establishment of pricing power over crypto-native assets, to shadow pricing of traditional assets, to institutional recognition, and finally to the integration of infrastructure.

After reading this chapter, the reader will understand that funding rate anchoring, as a general-purpose mechanism for generating synthetic exposure, applies in principle to any underlying asset; will grasp the "faster, cheaper, more accessible" logic of pricing-power migration; and will see clearly the feasibility, advantages, and obstacles of an Everything Perpetual, along with the real-world boundaries drawn by its threefold technical, institutional, and fundamental constraints. This provides the analytical foundation for the subsequent parts of the book, which take up decentralized engineering and autonomous finance.

27.1 The universality of the perpetual futures mechanism

Before developing the analysis, a methodological note about this chapter's framework is in order. There are two parallel paths for extending perpetual futures to traditional assets. The first is the centralized exchange (CEX) path: regulated centralized order-book exchanges extend the range of underlyings for perpetual futures from crypto assets to traditional assets. Examples include the tokenized equity spot products (xStocks) listed on Kraken and the tokenized equity perpetual futures that some platforms are now launching. The second is the decentralized exchange (DEX) path: permissionless on-chain protocols offer perpetual futures trading on synthetic assets through automated market makers or on-chain order books, as Hyperliquid does. The two paths share the same core principles of mechanism design — funding rate anchoring, margin management, and liquidation logic — but their execution-layer trade-offs differ sharply.

The CEX path sacrifices open access in exchange for regulatory compliance, high execution speed, and deep liquidity; the DEX path sacrifices execution efficiency in exchange for permissionless global access and censorship resistance. This chapter uses the CEX order book as its primary analytical framework, because the overwhelming majority of current perpetual futures volume occurs on centralized exchanges (for the specific share, see Section 27.1.5), and because the CEX path is the more likely near-term form for turning traditional assets into perpetual futures. The DEX path is discussed throughout as a point of comparison and complement, especially in the analysis of access barriers and regulatory constraints.

In the evolution of financial infrastructure, a tool built for one narrow purpose can end up reshaping the structure of an entire market. Perpetual futures are just such an instrument, one with systemic potential. This section steps outside the specific context of crypto assets and returns to first principles to examine the core mechanism of perpetual futures: funding rate anchoring. Through this abstraction, we can see that the funding rate is essentially a general-purpose mechanism for generating synthetic exposure, one that does not depend on any special properties of the underlying asset. In principle, so long as a reliable price oracle for the underlying asset exists, this mechanism can create perpetual futures for any asset, thereby building a general-purpose pricing infrastructure that spans everything.

27.1.1 The funding rate anchoring mechanism

When we return to first principles to examine the funding rate mechanism, we must set aside for the moment the specific context of crypto assets such as Bitcoin. The core function of the funding rate is to anchor the contract price firmly to some reference price in the absence of an expiration date and without physical delivery. This anchoring is achieved through periodic funding payments between longs and shorts: when the contract price exceeds the reference price, longs pay shorts; in the opposite case, shorts pay longs. This arbitrage incentive forces the contract price back toward the reference price [1].

In practice, the funding rate typically consists of two components. The standard formula can be expressed as $F = \text{clamp}(\text{Premium}, -d, d) + \text{clamp}(\text{Interest Rate}, -d, d)$, where the interest rate component reflects the lending spread between the quote currency and the base currency, the premium component measures the deviation between the perpetual futures' mark price and the spot index price, and the clamp function restricts the rate to within upper and lower bounds of $\pm d$ to prevent extreme swings. Centralized exchanges differ in how they implement this framework. Binance uses a volume-weighted average price method to compute the index price and settles the funding rate every 8 hours; Bybit uses a fair mark price method, smoothing the mark price through a moving-average basis; and some exchanges have raised the settlement frequency to every hour or even to continuous settlement. The theoretical analysis of Ackerer, Hugonnier, and Jermann (2025) shows that, under no-arbitrage conditions, the perpetual futures price equals the risk-neutral expectation of the spot price sampled at a random time that reflects the strength of funding rate anchoring; as the anchoring strength approaches infinity, the perpetual price converges monotonically to the spot

price, and under a particular funding rate specification the two are identical [1].

This reference price is not limited to the spot price of a crypto asset. It can be any price indicator that is observable and verifiable in public markets — the euro-dollar exchange rate, the level of the S&P 500 index, or the spot price of West Texas Intermediate crude oil, for example. So long as this reference price can be reliably obtained and transmitted to the trading system, the funding rate mechanism can operate effectively. The physical properties, trading conventions, and storage costs of the underlying asset have no material bearing on the intrinsic logic of the mechanism.

The funding rate is to perpetual futures what the interest rate is to a currency swap. In both cases it is a general-purpose tool for anchoring prices, and its effectiveness does not depend on the specific object being anchored. Just as a currency swap can be applied to any two liquid fiat currencies, the funding rate mechanism of perpetual futures can be applied to any underlying asset with reliable price discovery. The simplicity and mathematical generality of this mechanism give it the potential for universality across asset classes.

27.1.2 Applicability across asset classes

If the funding rate mechanism is so broadly applicable, why are perpetual futures today confined almost entirely to crypto-asset markets? The answer lies not in any technical limitation of the mechanism itself but in a combination of historical path dependence and the prevailing regulatory environment.

Historically, the perpetual futures instrument was launched for Bitcoin by the crypto derivatives platform BitMEX in May 2016 and rapidly scaled into widespread use [2]. The crypto community was the first to adopt this convenient, roll-free instrument, and over the following years developed it into the highest-volume product in the ecosystem. This history has created a strong cognitive inertia, so that market participants often equate perpetual futures with crypto-native products.

From the standpoint of the regulatory environment, traditional asset classes such as equities, foreign exchange, and commodities all sit within highly mature and stringent regulatory frameworks. Introducing a new kind of derivative — one with no expiration date, high leverage, and no guarantee from a central counterparty — into these traditional markets faces high compliance barriers. Crypto markets, by contrast, occupied a regulatory gray zone in their early years, and this relatively permissive environment supplied the conditions necessary for the rapid iteration and development of perpetual futures.

From a pure mechanism-design perspective, however, no economic or engineering reason restricts perpetual futures to serving crypto assets alone. Indeed, some cutting-edge decentralized exchanges have begun to offer perpetual futures on foreign exchange, equities, and commodities. Although trading in these non-crypto perpetual futures remains small in scale, their existence already demonstrates the technical feasibility of extending the funding rate mechanism to traditional asset classes. This indicates that, once the regulatory environment evolves or the technical architecture can satisfy compliance requirements, no fundamental technical obstacle stands in the way of perpetual futures

expanding across asset classes.

27.1.3 The economics of synthetic exposure

Another key dimension for understanding the universality of perpetual futures is the economics of synthetic exposure. In financial markets, capturing the return from an asset's price movements does not always require holding the asset physically. Owning the asset and owning a contract — that is, obtaining synthetic exposure — differ markedly in their economic consequences and in their trading friction.

Holding a physical asset typically requires purchase, custody, and safekeeping, a complex chain that necessarily involves settlement delays, delivery logistics, and possible restrictions on cross-border capital flows. Holding a contract, by contrast, requires the trader only to post a certain proportion of margin to obtain price exposure identical to that of holding the underlying asset. The advantages of this synthetic exposure are systemic: it achieves zero settlement delay and zero delivery cost, removes the geographic barriers to cross-border transactions, builds in leverage, and makes shorting as simple and direct as going long.

Among all instruments that provide synthetic exposure, perpetual futures represent the lowest-friction form. Traditional futures also provide synthetic exposure, but their inherent expiration date requires traders to roll their contracts periodically, and this roll cost constitutes a persistent trading friction. Exchange-traded funds (ETFs) likewise provide exposure, but they rely on an in-kind or cash creation-and-redemption mechanism, involve a complex system of authorized participants, and are confined to specific trading sessions. Perpetual futures eliminate these frictions: no expiration date means no roll cost, no delivery means no logistical friction, and the automatic anchoring of the funding rate ensures effective price tracking.

Within the spectrum of synthetic-exposure instruments, the contract for difference (CFD) closely resembles perpetual futures in functional terms: both have no expiration date, and both support leveraged, two-way trading. The two differ in one important respect at the level of market microstructure, however. A CFD is made by a broker acting as the sole counterparty, so that a trader's gain is the broker's loss; this concentrates counterparty risk in a single entity and gives the broker a structural incentive to manipulate quotes. Perpetual futures, by contrast, are traded through order-book matching on a centralized exchange or through an automated market maker on a decentralized exchange, so that counterparty risk is dispersed across the liquidation engine and the insurance fund system. In regulatory terms, CFDs are constrained by the retail-derivatives rules of particular jurisdictions — the United States prohibits retail CFD trading outright, and the European Union imposes strict leverage caps through MiFID II — whereas crypto perpetual futures have in practice achieved global access through on-chain protocols. The marginal innovation of perpetual futures relative to CFDs lies in two dimensions: the funding rate mechanism provides transparent, verifiable price anchoring (a CFD's quotes are determined entirely within the broker), and the decentralized implementation path removes the concentration of credit risk in a single counterparty.

In a 1993 book, the economist Robert Shiller set out a grand vision of macro markets.

He argued that liquid derivatives markets should be created for all important economic variables — gross domestic product, real estate prices, and national income among them — so that these major risks bearing on human welfare could be diversified and managed more effectively [3]. Given the financial technology of the time, this vision faced enormous implementation costs. Yet perpetual futures, a synthetic-exposure instrument superior to traditional alternatives in both flexibility and friction cost, fit Shiller's macro-markets vision more closely than any other existing synthetic-exposure mechanism.

27.1.4 The concept of an Everything Perpetual

Building on the preceding analysis of the universality of the funding rate mechanism and the economics of synthetic exposure, we can construct a thought experiment about a general-purpose pricing layer. At the heart of this thought experiment is a construct that we call the "Everything Perpetual" (a term coined by the author).

Imagine a unified global trading infrastructure on which perpetual futures on Bitcoin against the dollar, the euro against the dollar, the S&P 500 index, gold, and even carbon-emission metrics all trade simultaneously. All of these contracts, spanning different asset classes, share the same stablecoin margin system, the same unified clearing mechanism, and the same governance framework. On a single seamless interface, a trader can switch in an instant from long gold to short euro to long an equity index. All of this trading proceeds continuously, around the clock, is open without distinction to every participant worldwide, and settles instantly.

This construct qualifies as a pricing infrastructure, rather than merely an efficient trading platform, because once enough participants trade on the system, the prices it generates are positioned to become one of the important reference prices for global markets. Just as the futures prices of the Chicago Mercantile Exchange have long been regarded as benchmark prices for global commodities, this Everything Perpetual platform would establish pricing power across asset classes.

There is a fundamental difference from a traditional commodity exchange, however: this perpetual-futures-based general-purpose pricing layer is permissionless. Anyone, anywhere, with an internet connection and a digital wallet can take part in this global pricing process. This means that the billions of unbanked people excluded from the traditional financial system would, for the first time, gain a channel to participate in the pricing of core global assets and to manage their own economic risks. This is the central proposition of the chapter's thought experiment, and it concerns not only gains in financial efficiency but also the reconstruction of financial power.

27.1.5 The order-book microstructure of centralized exchanges

In practice, the principal vehicle for perpetual futures is the order-book matching system of centralized exchanges. Understanding this microstructure is essential for assessing the pricing efficiency of perpetual futures.

At the level of fee structure, mainstream centralized exchanges generally adopt a maker-taker fee model. The maker, who provides liquidity to the order book, typically enjoys

a lower fee or even a negative rate as an incentive, while the taker, who consumes liquidity, bears a higher fee. The economic logic of this asymmetric pricing structure is to use differentiated incentives to elicit a continuous supply of limit orders, thereby maintaining the depth of the order book and the tightness of the spread.

At the level of risk management, centralized exchanges build a multi-layered defense system to preserve the solvency of the system. The insurance fund, accumulated from liquidation gains and platform revenue, absorbs the shortfall when a liquidated position cannot be filled at the bankruptcy price or better. When the insurance fund is insufficient to cover the losses, the auto-deleveraging (ADL) mechanism ranks counterparties by profit and forcibly closes out the most profitable counterparty positions against the bankrupt position. Some exchanges also retain a socialized-loss fallback, allocating the residual loss pro rata across all profitable accounts. Together, these mechanisms form the solvency-assurance chain for perpetual futures on centralized exchanges.

Centralized exchanges also maintain a multi-tiered price system to balance different risk-management needs. The last traded price reflects the actual price of the most recent match in the order book. The mark price, typically computed as a weighted combination of the spot index price and a decaying basis, is used for profit-and-loss calculation and for triggering liquidations; its purpose is to filter out the price spikes caused by short-term manipulation. The index price aggregates a weighted average across multiple spot exchanges and serves as the anchoring benchmark for the funding rate calculation. Binance uses a volume-weighted average price method to compute the index price, whereas Bybit uses a fair mark price method, and the two methods perform differently under extreme market conditions. These design choices give the perpetual futures of centralized exchanges a marked competitive advantage in liquidity depth, execution speed, and risk-control maturity, making them the dominant form in the current perpetual futures market. As of 2025, centralized exchanges accounted for more than 90% of total perpetual futures volume (by notional turnover) [4].

27.2 The historical logic of pricing-power migration

The migration of pricing power is not a theoretical hypothesis but an objective phenomenon that recurs throughout financial history. In every major shift in market structure, pricing power moves from the old trading venue to the new one. This migration typically follows a similar logical pattern, and understanding this historical regularity is important for assessing whether perpetual futures have the potential to take over global asset pricing.

27.2.1 Three migrations of pricing power

Looking back over the development of modern financial markets, we can clearly identify three consequential migrations of pricing power.

The first migration occurred in the mid-to-late nineteenth century and marked the movement of pricing power from fragmented spot markets to a centralized futures exchange. The Chicago Board of Trade was founded in 1848 and initially traded forward (or “to-arrive”) contracts; its truly landmark innovation came around 1865, when it

transformed forward contracts into standardized delivery futures — unifying quantity, quality, delivery location, and delivery date, and introducing margin and formal settlement rules [5]. Before this, pricing power over agricultural commodities was dispersed across countless local spot-trading networks, and information asymmetry produced violent price swings. The advent of the Chicago Board of Trade allowed farmers and buyers to lock in future buying and selling prices in advance, giving them certainty. This standardized and centralized trading model sharply reduced information costs and price uncertainty, so that the exchange's futures prices quickly became the reference prices for the global grain market, completing the first major migration of pricing power.

The second migration took place between the 1990s and the 2010s and was characterized by the movement of pricing power from traditional open-outcry floor trading to electronic trading systems. As computing technology and communications networks spread, electronic matching platforms — exemplified by the Chicago Mercantile Exchange's Globex system and Nasdaq — gradually displaced the traditional trading floor [6]. This migration greatly increased trading speed, compressing execution time from seconds to milliseconds; it also sharply reduced the high costs of manual matching and broke the constraints of physical space, allowing investors worldwide to participate more easily in market pricing.

The third migration occurred mainly from the early 2000s into the 2020s and manifested as the movement of pricing power from traditional delivery futures to exchange-traded funds. For certain assets, particularly equity indices and gold, the trading volume of exchange-traded funds gradually surpassed that of the corresponding futures contracts (such as the Chicago Mercantile Exchange's E-mini futures on the S&P 500), making ETFs the primary tool through which retail investors and some institutional investors obtain market exposure [7][8]. The success of exchange-traded funds rests on their minimalist user experience: investors need not manage complex futures accounts and margin, and can participate simply through an ordinary securities account. Moreover, the total holding cost of an exchange-traded fund is typically lower than the roll cost of a futures contract, which further accelerated this migration.

27.2.2 The drivers of migration

The three migrations of pricing power follow a common driving logic that can be distilled into four core elements: faster (lower trading latency and higher-frequency information processing), cheaper (lower total transaction and access costs), more accessible (a broader range of participants and a lower barrier to entry), and more trustworthy (regulatory protection, depth of institutional participation, and legal certainty). In all three historical migrations, the new venue was no weaker than the old on the trustworthiness dimension, so that dimension did not become a constraint; yet it is precisely the principal weakness of perpetual futures relative to traditional markets. From standardized futures lowering search costs, to electronic matching compressing execution time to milliseconds, to exchange-traded funds turning complex exposure into a product accessible through an ordinary securities account, each migration was the result of the new venue accumulating sufficient advantage along these three dimensions. Throughout this

process, resistance from incumbents delayed but could not reverse the migration; once the efficiency advantage of the new venue is sufficiently pronounced, the migration of pricing power tends ultimately to occur. Section 27.7 further analyzes how this driving logic acts specifically on the extension of perpetual futures to traditional assets.

27.2.3 The candidacy of perpetual futures

Assessed against this framework, perpetual futures hold a structural advantage on two dimensions: faster (around-the-clock continuous trading, which eliminates the price vacuum of non-trading hours) and more accessible (permissionless global access with a very low barrier to entry). On the cheaper dimension, they hold an advantage in some asset classes because they eliminate roll costs and delivery logistics. Their principal weakness lies in the trustworthy dimension: the institutional accumulation of traditional markets in regulatory protection, depth of institutional participation, and legal certainty constitutes a trust gap that perpetual futures must close. The speed at which this gap narrows will directly determine the course of pricing-power migration, and Section 27.7 develops a detailed analysis through a four-stage model.

27.3 Thought experiment I: foreign exchange perpetual futures

Foreign exchange is the largest financial market in the world, and its depth of liquidity and breadth of trading form a cornerstone of the modern international financial system. According to the Bank for International Settlements (BIS), average daily turnover in the global foreign exchange market reached about $9.6 trillion in April 2025, roughly 28% higher than in the same period of 2022 [9]. Despite its vast scale, however, this market still exhibits pronounced structural inefficiencies. This section takes foreign exchange perpetual futures as the starting point for a thought experiment, examining which problems the perpetual futures mechanism can solve when applied to the world's largest financial market, and which insurmountable institutional obstacles it would face.

27.3.1 Structural inefficiencies of the foreign exchange market

The foreign exchange market is nominally a market that operates around the clock, yet its internal structure is markedly fragmented. The first issue is the uneven distribution of liquidity across trading sessions. Liquidity is heavily concentrated in the overlapping London and New York sessions, while the Asian session is relatively thin. More critical still, during the non-trading hours of the weekend the market loses its price-discovery function entirely. When a geopolitical event or major macroeconomic data release occurs over the weekend, participants must wait for the "open" on Sunday evening or Monday morning before they can trade. This "weekend gap" constitutes a serious information-transmission delay and a structural inefficiency.

The second issue is the systemic risk created by settlement delay. The standard settlement cycle for traditional foreign exchange transactions is typically 2 days, which means that after reaching an agreement the two parties must bear 2 days of settlement

risk exposure. In a market with daily volume in the trillions of dollars, this delay accumulates enormous systemic risk. Although some traditional financial markets have recently begun to shorten their settlement cycles — the U.S. equity market, for example, has shortened its cycle from 2 days to 1 day [10] — the complex network of the global foreign exchange market makes universal instant settlement a formidable challenge.

The final issue is access barriers and execution costs. For retail investors and small and medium-sized enterprises, the barrier to direct participation in the core foreign exchange market is extremely high. They can generally trade only through brokers or commercial banks, and in this "last mile" they face wide bid-ask spreads and suboptimal execution quality, bearing a disproportionate share of transaction costs.

27.3.2 The design of foreign exchange perpetual futures

If we introduce the perpetual futures mechanism into the foreign exchange market — designing, say, a euro-dollar perpetual futures contract — its core architecture would display characteristics quite different from those of traditional foreign exchange derivatives. The contract could be denominated and settled in a stablecoin, with its price anchored to the spot euro-dollar exchange rate through the funding rate mechanism. An oracle network would obtain and transmit the spot price in real time, while the funding rate could settle at an hourly or even continuous frequency.

Compared with the traditional foreign exchange futures offered by the Chicago Mercantile Exchange, foreign exchange perpetual futures eliminate the constraint of an expiration date and thus dispense with the roll operations traders must perform at expiry and the costs that accompany them. More importantly, they achieve genuine around-the-clock continuous trading and eliminate the weekend gap. At the same time, a stablecoin settlement mechanism built on a blockchain network achieves instant delivery, replacing the 2-day settlement delay of traditional markets.

From the standpoint of pricing theory, the pricing of traditional foreign exchange forwards rests on covered interest rate parity. Within the perpetual futures framework, the funding rate should, in long-run equilibrium, converge toward the risk-free interest rate differential between the two underlying currencies. This prediction is empirically testable, and it shows that the funding rate mechanism does more than anchor prices: it intrinsically reflects differences in macroeconomic variables.

27.3.3 Potential advantages

The core structural advantage of foreign exchange perpetual futures lies in their capacity for around-the-clock continuous pricing. In today's highly interconnected world, geopolitical risks and sudden events are not confined to the business hours of a traditional exchange. If foreign exchange perpetual futures achieved around-the-clock trading, a sudden event on a Friday evening could be absorbed by the market and reflected in prices within minutes, rather than waiting for an open more than 40 hours later. Such a mechanism would produce a structural leap in informational efficiency.

The instant settlement mechanism likewise has significant practical value. Instant settlement via stablecoins eliminates the 2-day settlement risk exposure, which matters

greatly for the risk management of multinational corporations and large financial institutions. It reduces counterparty risk and substantially improves capital efficiency.

Moreover, undifferentiated global access gives foreign exchange perpetual futures an important dimension of financial inclusion. Around the world, billions of people live in countries or regions with strict foreign exchange controls, and they cannot freely convert into strong currencies through legal channels to protect their assets. Foreign exchange perpetual futures running on a decentralized network can provide these populations with a tool for obtaining synthetic foreign exchange exposure. The core value of such a tool lies not in speculation but in giving ordinary people a way to hedge against depreciation of their home currency, which carries considerable social value.

27.3.4 Core obstacles

Although foreign exchange perpetual futures hold clear advantages in mechanism design, their broad adoption faces powerful resistance from sovereign states and the existing financial system. The foremost obstacle is the monetary policy transmission mechanism of central banks. Central banks rely on their intervention in, and influence over, the foreign exchange market to implement monetary policy and safeguard domestic economic stability [11]. A globally mobile perpetual futures market controlled by no single jurisdiction could weaken the transmission of central bank monetary policy. Sovereign states will not lightly relinquish control over the pricing power of their currencies or over the foreign exchange market.

Anti-money-laundering compliance is another formidable obstacle. The foreign exchange market has always been a high-risk arena for cross-border capital flows and money laundering. A permissionless, decentralized foreign exchange perpetual futures platform faces enormous technical and institutional challenges in implementing adequate know-your-customer (KYC) and anti-money-laundering (AML) procedures. Regulators take a cautious stance toward financial infrastructure that cannot trace the source of funds or the identity of traders [12].

The settlement needs of the real economy also mean that perpetual futures cannot fully replace the spot market. Multinational enterprises making international trade payments require genuine transfers of currency, whereas perpetual futures can offer only a synthetic derivative of price exposure and cannot meet the need for physical delivery. The contract market and the spot market are therefore bound to coexist over the long run, with the former relying on the latter for its price anchor.

Finally, the reliability of the oracle constitutes a single point of risk at the technical level. Trading volume in the foreign exchange market is extremely large, and even a tiny deviation in an oracle's quote could be exploited by high-frequency traders, producing large arbitrage losses. Whether current oracle infrastructure possesses the stability and manipulation resistance required to support a foreign exchange derivatives market on the order of trillions of dollars remains an open question in urgent need of verification.

27.3.5 Feasibility assessment

Taking the preceding analysis together, foreign exchange perpetual futures are more likely in the near term to serve as a complement to the existing traditional foreign exchange market than as its replacement. They can play a distinctive role in filling the weekend pricing vacuum and lowering the barrier to entry for retail investors. For ordinary people in emerging markets, they offer a practical tool for hedging the risk of home-currency depreciation.

Over the longer term, if oracle technology achieves a fundamental breakthrough and the regulatory framework for decentralized finance becomes progressively clearer and widely accepted, foreign exchange perpetual futures could divert a portion of the speculative trading volume of traditional markets. Yet because the central position of central banks and sovereign governments in the monetary system is a cornerstone of the modern state, foreign exchange perpetual futures cannot fully replace the traditional foreign exchange market. The migration of pricing power will be a gradual and constrained process, finding its ultimate equilibrium in the interplay between technological innovation and institutional constraint.

27.4 Thought experiment II: equity index perpetual futures

The success of exchange-traded funds demonstrates substantial latent demand for simplified ways of obtaining asset exposure. Along this trajectory, equity index perpetual futures can reasonably be understood as the ETF's next evolutionary form. They build an around-the-clock, global, and permissionless trading environment. Inevitably, however, this innovative mechanism must confront the institutional advantages that ETFs have accumulated over time, as well as the challenge of winning regulators' acceptance.

27.4.1 The success and limits of the ETF

Since the first ETF tracking the S&P 500 index (SPY) appeared in 1993, ETFs have become one of the most successful products in the democratization of finance. By the end of 2025, global ETF assets under management approached $20 trillion (of which U.S.-listed ETFs accounted for about $13 trillion), and U.S.-listed ETFs attracted about $1.5 trillion of net inflows in 2025 alone, a record high [7]. This product structure allows ordinary investors to obtain broad market exposure at very low cost.

Despite this enormous success, the ETF mechanism retains inherent structural limits. The first is the rigidity of the trading session. SPY, for example, trades mainly during New York's regular session, from 9:30 a.m. to 4:00 p.m. Eastern Time. Participants in other time zones must passively adapt to this window, and Asian investors cannot directly participate in the pricing and trading of SPY during their own regular working hours [13]. The second is geographic and institutional barriers. Non-U.S. residents who wish to buy U.S.-listed ETFs generally must open a securities account across borders, facing burdensome compliance review and obstacles to moving funds internationally.

A more fundamental limit lies in the ETF's liquidity-creation mechanism. The arbitrage anchoring of an ETF depends heavily on the creation-and-redemption mechanism

operated by authorized participants. In the primary market, these large financial institutions swap a basket of underlying securities for ETF shares, keeping the ETF's market price closely aligned with its net asset value [14]. During extreme market volatility, however, authorized participants may withdraw from market-making because of evaporating liquidity or risk considerations, causing the ETF to trade at a large premium or discount — a phenomenon analyzed in detail in Chapter 18 of this book.

27.4.2 The design of index perpetual futures

To break through these limits, on-chain equity index perpetual futures offer a wholly new mechanism design. Consider the SPX perpetual futures tracking the S&P 500 index: their core architecture typically uses a stablecoin (such as USDC) as the unit of denomination and settlement. The contract involves no actual delivery of the underlying stocks; instead, an oracle system continuously obtains the latest level of the S&P 500 index as the reference benchmark. On this basis, the funding rate mechanism substitutes for the arbitrage behavior of authorized participants, automatically anchoring the perpetual futures' trading price to the spot index price through funding exchanges between longs and shorts [15].

To display the structural differences of this design more clearly, Table 27-1 systematically compares the traditional SPY ETF with an idealized SPX perpetual futures contract across 10 core dimensions. This multidimensional comparison helps identify the dimensions on which perpetual futures hold a structural advantage and those on which they carry an inherent institutional weakness.

Comparison dimension	SPY ETF (traditional finance)	SPX perpetual futures (on-chain finance)
Trading session	New York regular session (6.5 hours per day)	Around the clock, uninterrupted (24/7)
Global access	Geographic barriers; requires a U.S. securities account	Permissionless; requires only a crypto wallet (decentralized version)
Access to leverage	Depends on broker financing; constrained by regulatory margin requirements	Built into the protocol; typically 1x to 20x or higher
Shorting mechanism	Requires borrowing shares, with high costs and limited supply	Open a short directly, with no additional friction cost
Settlement cycle	Delayed settlement (T+1)	Instant settlement via smart contract
Holding cost	Fixed annual management fee (about 0.09%)	Dynamic funding rate (potentially much higher than the management fee)
Price anchoring	In-kind creation and redemption by authorized participants	Algorithmically driven funding rate arbitrage
Regulatory protection	Strictly regulated by the U.S. Securities and Exchange Commission (SEC)	Regulation absent or in a gray zone
Oracle dependence	No external price feed required	Core mechanism; highly dependent on oracle accuracy
Ownership attributes	Includes voting and dividend rights of the constituent stocks	Pure price exposure, with no corporate governance rights

Table 27-1. Comparison of the core mechanisms of the SPY ETF and SPX perpetual futures (Data source: CoinDesk reporting on Kraken's tokenized equity perpetuals [15]; Schwab Asset Management report on the ETF creation-and-redemption mechanism [14]). Note: the comparison of 6.5 hours with 24/7 is a coarse-grained contrast along the trading-session dimension; assessing the actual execution cost of arbitrage across U.S. cash equities, futures, and ETFs would also require incorporating the pre-market and after-hours sessions, the CME's near-around-the-clock trading, and the authorized-participant creation-and-redemption mechanism discussed in Chapter 18

As shown in Table 27-1, SPX perpetual futures hold a structural advantage over the SPY ETF on five dimensions: trading session, global access, access to leverage, ease of shorting, and settlement speed. These advantages are not marginal improvements but stem from a fundamental difference in the underlying architecture: the synthetic-exposure mechanism eliminates the frictions of physical delivery and intermediary institutions. Yet the table also clearly reveals the institutional weaknesses of perpetual futures: the absence of regulatory protection means that investors lack legal recourse in a dispute; the loss of ownership attributes means that they cannot replace the ETF's function as a tool for long-term value investing; and the uncertainty of holding cost (a dynamic funding rate can, under extreme market conditions, far exceed an ETF's fixed management fee) adds further unpredictability to long-term holding strategies. This asymmetric profile of "strong on offense, weak on defense" runs throughout the comparison between perpetual futures and traditional financial products.

27.4.3 Potential advantages

The clearest near-term value of index perpetual futures is filling the Asian-session pricing vacuum. In the current market structure, after the U.S. equity market closes, for more than 10 hours global investors have no effective venue to price their views on changes in the U.S. macroeconomy or in corporate fundamentals. If a major geopolitical event or an emergency statement by the Federal Reserve (Fed) occurs during this window, traditional markets must wait for the New York open to release the accumulated pricing pressure. The around-the-clock nature of SPX perpetual futures can effectively fill this vacuum, so that information enters the price-discovery process the instant it arises [16].

Moreover, this mechanism greatly extends the boundaries of financial inclusion. Billions of people worldwide are constrained by local financial infrastructure and cannot easily buy U.S. ETFs. Through the SPX perpetual futures of a decentralized exchange, these groups need only an internet connection and a digital wallet to gain exposure to the world's most important equity market, which democratizes global capital allocation to a degree.

At the level of market microstructure, perpetual futures markedly reduce the friction cost of expressing a bearish view. In the traditional equity market, shorting requires locating a supply of shares and paying high borrowing fees, and this asymmetric transaction cost means that negative information often fails to be reflected promptly and fully in asset prices. The symmetric design of perpetual futures makes shorting as convenient as going long, which helps improve the price-discovery efficiency of the market as a whole.

27.4.4 Core obstacles

For all the ingenuity of their mechanism design, index perpetual futures still face insurmountable obstacles in real-world adoption. The first is the regulatory challenge. The core duties of securities regulators are investor protection and the maintenance of orderly markets, and they are highly unlikely to embrace, in the near term, an S&P 500 derivative that sits outside the traditional compliance framework. Although a decentralized exchange can, through its technical architecture, evade some direct jurisdiction, doing so also means transferring all compliance risk and counterparty risk entirely onto ordinary users [15].

The reliability of the oracle system constitutes another key technical bottleneck. During the regular U.S. equity session, perpetual futures can easily obtain real-time index quotes from data sources such as Bloomberg or Reuters. But during the long non-trading hours, the spot market ceases to operate, and the oracle can rely only on the relatively thin after-hours session or on E-mini futures prices as a reference [17]. This fragility of the data source makes perpetual futures highly susceptible to malicious manipulation during non-trading hours, and a tiny oracle deviation, amplified by leverage, could trigger a catastrophic liquidation cascade.

Another core tension is that index perpetual futures are, in essence, decoupled to some degree from the real economy. They provide pure synthetic price exposure and confer no

ownership whatsoever in the underlying listed companies. This means that a holder of perpetual futures can neither exercise shareholder voting rights nor participate directly in corporate dividends. In terms of financial function, they are not fully equivalent to an ETF that represents ownership of real assets, and this limits their potential to become the preferred tool for long-term value investing.

27.4.5 Near-term value

Weighing their advantages and obstacles together, we can reach a relatively conservative judgment: even if index perpetual futures cannot fully replace the traditional ETF in the foreseeable future, the single function of filling the Asian-session pricing vacuum is enough to establish their distinctive value in the modern financial ecosystem. They can be positioned precisely as an extension of, and complement to, the traditional ETF in trading session and geographic reach.

This convergence and complementarity has in fact already begun to take place in the market. In early 2026, leading platforms including Kraken began launching regulated tokenized equity perpetual futures, supporting 24/7 trading of major indices including the S&P 500 [15]. At the same time, oracle infrastructure such as Chainlink launched U.S. equity data streams covering the full trading day on weekdays (24/5, including the pre-market, after-hours, and overnight sessions), providing continuous pricing support for on-chain derivatives — though continuous pricing over the weekend remains an unresolved gap [17]. Although overall trading volume still appears negligible compared with traditional markets, the completion of this proof of concept marks a key step in the penetration of crypto-native mechanisms into the pricing power of traditional financial assets.

27.5 Scenario analysis: price transmission in TSLA perpetual futures

In the evolution of financial infrastructure, the process by which a new market moves from the periphery to the core is not a linear, gradual progression but often exhibits pronounced phase-transition characteristics. When a new derivatives market attempts to contest pricing power over an asset, it must cross a series of thresholds in liquidity and informational efficiency. This section builds a conceptual scenario-analysis framework, taking Tesla, Inc. (ticker: TSLA) as the benchmark asset, to examine how a hypothetical "TSLA perpetual futures" market might, under varying conditions, progressively acquire and even come to dominate global pricing power over the asset. This thought experiment is not only a test of the theoretical framework developed above but also an attempt to answer a central question: what scale must permissionless on-chain perpetual futures reach before they materially affect the price-discovery process of traditional financial markets, forcing institutional traders to incorporate on-chain price signals into their decision-making framework?

27.5.1 The traditional pricing ecosystem of TSLA

The subjects of the previous two thought experiments — the foreign exchange market and equity indices — are macro-level asset classes whose pricing is dominated by the aggregation of many constituents. This section shifts the analytical granularity down from the index level to the level of a single stock. We choose an individual stock rather than an index because individual stocks have higher volatility, more pronounced information asymmetry, and fiercer contests between institutions and retail investors, which makes for a more extreme test of pricing-power migration. If perpetual futures can capture a material contribution to price discovery in the demanding case of a single stock, their prospects at the index level become more certain.

The choice of Tesla as the subject of the scenario analysis is not random. In contemporary financial markets, TSLA has a set of distinctive microstructure characteristics that make it a suitable subject for testing the pricing-power-migration hypothesis. As the world's most valuable automaker, TSLA has a highly complex and active pricing ecosystem. As of March 2026, Tesla's market capitalization stood at around $1.48 trillion, and its stock exhibited very high liquidity, with average daily volume steady at around 65 million shares and average daily turnover exceeding $25 billion [18]. Such a vast liquidity pool means that any new market seeking to challenge its pricing power must possess extremely deep market thickness.

More critically, in traditional markets TSLA displays a dual character of high volatility and high information sensitivity. Its 5-year monthly beta is as high as 1.93, and its implied volatility has long held at around 45% [18]. In the options market, TSLA is one of the most actively traded single-stock options in the world, with open interest in call options alone remaining above 3.9 million contracts year-round [18]. This microstructure, driven jointly by a vast retail base and high-frequency quantitative institutions, makes TSLA's price-discovery process highly dependent on the capacity to process information instantly.

Yet although TSLA enjoys near-perfect pricing efficiency during the regular session (9:30 a.m. to 4:00 p.m. Eastern Time), the after-hours and weekend periods expose the inherent flaws of the traditional market structure. During the after-hours session, TSLA's volume typically plunges to around 1.5 million shares, less than 3% of the regular session [19]. This evaporation of liquidity causes the after-hours spread to widen markedly, and the price's response to information becomes extremely fragile and prone to overshooting. When a major event — such as an Elon Musk tweet or a sudden geopolitical event — occurs during non-trading hours, traditional markets can absorb the information only through the extremely thin after-hours session or by waiting for the open on the next trading day, producing a large "time gap" and arbitrage opportunity. It is precisely this session-based pricing vacuum that offers around-the-clock (24/7) perpetual futures a structural opening into the global pricing system.

Before constructing the TSLA scenario analysis, empirical data from the Bitcoin perpetual futures market provide a useful reference anchor. Using the information-share method of Hasbrouck (1995), the empirical study of Alexander and Heck (2020) shows that unregulated crypto derivatives markets (perpetual futures and futures) already command about half of the information share in Bitcoin price discovery and domi-

nate the price-discovery process [20][21]; the study by De Blasis and Webb (2022) of arbitrage and spillover effects between quarterly and perpetual contracts on crypto exchanges likewise confirms the dominant role of derivatives markets in pricing [22]. As of 2025, according to CoinGlass data, the average daily volume of Bitcoin perpetual futures accounted for more than 60% of Bitcoin's total market volume (spot and derivatives combined), making perpetual futures the dominant venue for pricing crypto assets [4]. This "already-realized first stage" provides an empirical reference for the TSLA scenario analysis that follows: the trajectory by which Bitcoin perpetual futures grew from a marginal tool into the dominant force in price discovery can serve as a lower-bound estimate of the potential development path for perpetual futures on traditional assets. Bitcoin, however, as a digitally native asset, owes the success of its perpetual futures to special conditions — no need for physical delivery, and little competition from traditional alternatives — so extrapolating directly to traditional equities such as TSLA still calls for caution.

27.5.2 The price-transmission model

To assess systematically the potential influence of TSLA perpetual futures on the spot market, we construct a conceptual price-transmission framework. A caveat is essential here: the functional forms, thresholds, and PTI values given in the following model and in the subsequent figures are all illustrative estimates intended to organize thinking and to display orders of magnitude and nonlinear shapes; they are neither a directly runnable quantitative model nor to be read as precise forecasts or statistical estimates. In market microstructure theory, the contribution to price discovery of two parallel markets is usually measured through the information-share model proposed by Hasbrouck (1995) or the Gonzalo-Granger component-share model [21]. Drawing on this methodological framework, we define the "price transmission intensity" (PTI) as the weight of the perpetual futures market in the total price-discovery process, ranging from zero (completely passive acceptance of the spot price) to one (complete dominance of the spot price); intuitively, PTI measures the "share of voice" the new venue holds in price movements.

Price transmission intensity is not determined by a single variable but is a nonlinear function of the interaction of several market-microstructure parameters. In our model, the core variables that determine PTI span four dimensions. The first is relative liquidity depth — the ratio of the bid-ask depth of the perpetual futures market to that of the traditional spot and options markets. Only when the depth of the perpetual market is sufficient to absorb institutional-sized orders without producing a destructive price impact will market makers in traditional markets have an incentive to arbitrage across markets. The second is the difference in information-processing speed, which reflects the time lag between the two markets in completing a price update after receiving new information. The block-confirmation time and oracle latency of on-chain perpetual futures must be low enough to gain an edge in the competition over information.

The third determinant is the sensitivity of the funding rate. The funding rate is the sole mechanism by which perpetual futures anchor to the spot price, and its speed of response to price deviations and the severity of its penalty directly determine the

direction in which arbitrage capital flows. If the funding rate adjusts too sluggishly, the perpetual futures will devolve into a freestanding speculative casino; if it adjusts too violently, it will trigger frequent liquidation cascades. The last determinant is cross-market arbitrage friction, which includes trading fees, delays in transferring funds, compliance costs, and the friction of on-chain and off-chain fiat channels. This friction coefficient constitutes the physical resistance to price transmission, and only when the price difference between the perpetual futures and spot exceeds this friction threshold is the price-transmission mechanism truly activated.

By feeding these four variables into a nonlinear dynamical system, we can simulate how, as the perpetual futures market grows in scale, its pricing power penetrates the traditional TSLA spot market. This process is not a smooth linear rise but undergoes abrupt shifts at particular parameter thresholds, displaying clear stage-wise leaps.

27.5.3 A five-stage scenario simulation

Building on the model framework above, we simulate the development trajectory of TSLA perpetual futures as five progressive stages, each representing a paradigm shift in the transmission of pricing power.

The first stage is the "shadow market" stage. Here, TSLA perpetual futures have a daily volume of less than 1% of the spot market (about $250 million). The perpetual market is at this point entirely a dependency of the traditional market, its price passively supplied by an oracle. Because liquidity is thin, any medium-sized order pushes the perpetual price sharply away from the spot price, after which a high funding rate forcibly pulls it back. At this stage, price transmission intensity is close to zero; the market lies outside the monitoring field of traditional financial institutions and has not yet drawn the attention of institutional traders. It exists only as a marginal tool for high-leverage speculation by crypto-native users.

The second stage is the "extreme-conditions buffer" stage. When the daily volume of perpetual futures reaches 5% to 10% of spot (about $1.2 billion to $2.5 billion), the market structure begins to change subtly. During the regular session, the perpetual market remains a price taker. When the market experiences extreme volatility, however — say, an earnings shock that drives the spot price limit-down — some of the panicked selling that cannot be closed out in the traditional market spills over into the perpetual market. The depth of the perpetual futures is not yet sufficient to reverse the spot trend, but the additional liquidity pool it provides begins to serve as a buffer. Market makers start to run algorithms for rough statistical arbitrage between the two markets, and under extreme conditions PTI leaps briefly to around 10%.

The third stage is the "non-trading-hours dominator" stage, and it marks the first critical threshold in pricing-power migration. As perpetual futures volume reaches 20% of spot (about $5 billion), their around-the-clock trading advantage begins to show. If a major sudden event involving TSLA occurs over the weekend or after hours, the traditional market cannot respond because it is closed, while the perpetual market immediately begins price discovery. Because the perpetual market now has considerable depth, the price it forms is no longer noise but embodies a genuine market consensus. When Nasdaq opens on Monday, the spot price gaps up or down directly to meet the

price the perpetual market formed over the weekend. At this stage, the perpetual market's PTI during non-trading hours is as high as 80% or more, though during the regular session it still takes second place.

The fourth stage is the "arbitrage engine" stage. When the perpetual market reaches 50% of spot (about $12.5 billion), it has grown too large for the traditional market to ignore. Large quantitative funds and high-frequency trading firms begin to make it part of their core strategies. At this stage, the funding rate of the perpetual futures is no longer merely an anchoring mechanism but becomes a leading indicator for predicting the short-term direction of spot. If the perpetual price stays persistently above spot, large arbitrage capital buys TSLA in the spot market while shorting in the perpetual market, and this arbitrage transmits the perpetual market's long sentiment directly to the spot market, pushing the spot price up. At this point, PTI during the regular session holds steady at 30% to 40%, and the perpetual market has become an indispensable engine in TSLA's global pricing system.

The fifth stage is the "pricing-power reversal" stage. In the limiting scenario of the simulation, TSLA perpetual futures volume exceeds the combined total of the spot and traditional options markets (what follows is an extrapolation of the model's limiting scenario, not a forecast about TSLA). This typically occurs after compliance channels are fully opened and traditional institutional capital connects seamlessly to on-chain infrastructure. At this stage, the perpetual market — by virtue of zero settlement delay, no physical-delivery friction, and undifferentiated global access — becomes the venue with the deepest liquidity and the fastest response to information. The traditional spot market is instead reduced to a "derivative," its price passively following the movements of the perpetual market. Market makers set their spot-market quotes primarily by reference to the real-time order book of the on-chain perpetual futures. At this point, PTI exceeds 60%, and pricing power reverses in the sense of the model, just as delivery futures once wrested pricing power over agricultural commodities from fragmented spot markets.

Figure 27-1 plots the PTI trajectory of these five stages as an S-shaped curve against the perpetual futures' share of trading volume. Its core message is that pricing-power migration is not a linear, quantitative process but exhibits phase-transition behavior at two critical thresholds, 20% and 50%: before a threshold, the marginal contribution of incremental liquidity is extremely limited, but once it is crossed, a positive feedback loop drives PTI to leap rapidly toward saturation.

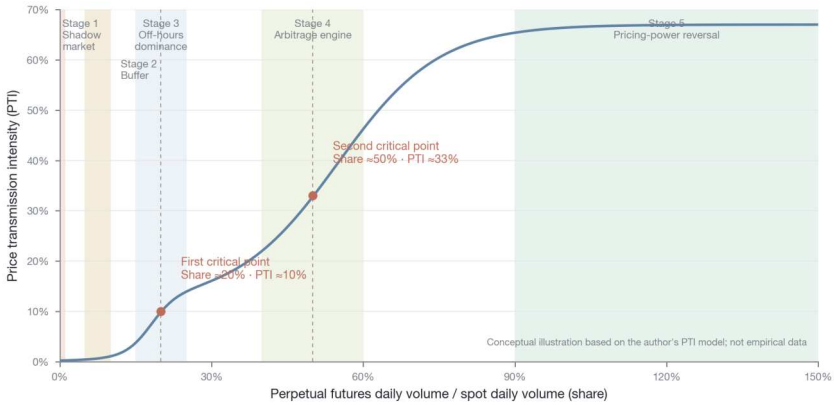


Figure 27-1. The five-stage evolution of PTI as a function of the perpetual futures' share of trading volume (conceptual illustration, not a forecast and not empirical: the curve is an extrapolation of the author's PTI model, using the double-logistic function $\mathrm{PTI}(x) = 0.12{\cdot}\sigma(40(x - 0.18)) + 0.55{\cdot}\sigma(10(x - 0.55))$, where σ is the logistic function; the horizontal axis is the model's independent variable, namely the perpetual-to-spot daily volume ratio, not a time axis; the two critical points at 20% and 50% and the individual PTI values are illustrative parameter choices, not statistical estimates; the PTI concept draws on the information-share method of Hasbrouck 1995 [21], which provides the measurement framework but is not the calibration source for this figure's thresholds)

27.5.4 Sensitivity analysis

In the evolution across these five stages, the weight of each variable is not constant. A sensitivity analysis reveals the key bottlenecks that determine whether a threshold can be crossed.

The first is sensitivity to oracle latency. The model shows that when the perpetual futures' oracle update frequency is slower than one second, its PTI ceiling during the regular session is locked at 15%, because on Nasdaq the reaction time of high-frequency trading firms is measured in microseconds. If an on-chain oracle needs several seconds to reflect a change in the spot price, the perpetual market will remain at a disadvantage in passive arbitrage and become a venue from which high-frequency traders systematically extract value [23]. Only when oracle latency falls to the hundred-millisecond level, with very high manipulation resistance, does the perpetual market gain the physical basis to contest pricing power during the regular session.

The second is the nonlinear effect of the friction coefficient. The simulation finds a clear convex relationship between cross-market arbitrage friction and price transmission intensity. When the all-in cost of moving funds across the on-chain and off-chain boundary — including slippage, gas fees, and losses on fiat on- and off-ramps — exceeds 0.5%, price transmission is significantly suppressed. When this friction coefficient falls below 0.1% through improvements in infrastructure (such as the spread of stablecoin settlement and the establishment of institutional-grade clearing networks), however, PTI grows rapidly. This shows that pricing-power migration is not a linear quantitative change but depends on a threshold effect that takes hold once infrastructural resistance falls below a particular level.

The last is sensitivity to the funding rate cycle. The traditional funding rate mechanism, which settles once every 8 hours, proves too sluggish for a high-volatility asset. The model indicates that shortening the funding rate settlement cycle of TSLA perpetual futures to hourly or even continuous settlement would improve its price-anchoring efficiency by more than 40%. Tighter anchoring means smaller price deviations, which attracts more slippage-sensitive institutional capital and forms a virtuous circle of rising liquidity and improving pricing efficiency.

Figure 27-2 presents, in two panels, the sensitivity of PTI to two bottleneck variables: (a) before oracle latency falls to the hundred-millisecond level, the regular-session PTI ceiling is locked at 15%; and (b) once cross-market friction falls below 0.1%, PTI grows rapidly. Together, the two panels reveal the same regularity: the bottleneck in pricing-power migration is not liquidity itself but the quality of the infrastructure connecting the two markets.

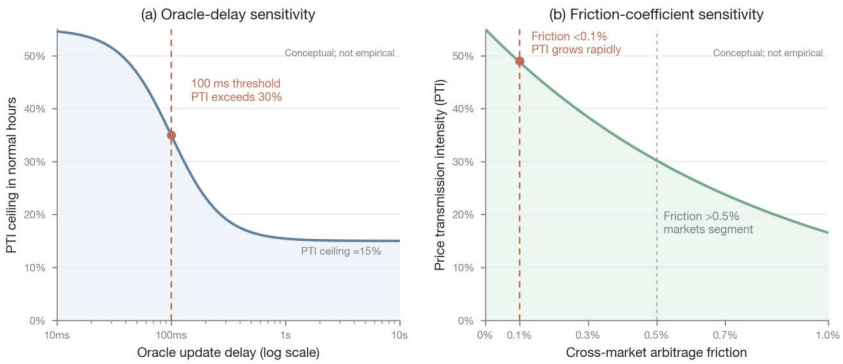


Figure 27-2. Sensitivity curves of PTI to oracle latency and the friction coefficient (conceptual and parameter-sensitivity modeling illustration, not empirical: all are extrapolations of the author's model; (a) oracle-latency model: PTI = $0.15 + 0.40 \cdot \sigma(4.5(2 - \log_{10}(\text{latency})))$; (b) friction-coefficient model: PTI = $0.55 \cdot \exp(-120 \cdot \text{friction coefficient})$, with the friction coefficient entered as a decimal and calibrated to the thresholds in the text as 0.1% → 0.49, 0.3% → 0.38, 0.5% → 0.30, 1% → 0.16)

27.5.5 Summary of the thresholds

Combining the scenario simulation with the sensitivity analysis, we can clearly delineate the quantitative thresholds at which TSLA perpetual futures acquire pricing power. The first threshold is the "non-trading-hours dominance threshold": the perpetual market's daily volume reaches 20% of spot (about $5 billion), and cross-market friction is below 0.3%. Crossing this threshold means that perpetual futures have formally filled a structural gap in the traditional financial market and begin to demonstrate irreplaceable informational value.

The second threshold is the "regular-session competition threshold," which requires the perpetual market's daily volume to reach 50% of spot and oracle latency to fall below 100 milliseconds. Crossing this threshold means that perpetual futures have gained the ability to compete with Nasdaq in price discovery during the regular session and have become a central hub for global arbitrage capital.

The third threshold is the "pricing-power reversal threshold," which requires not only

that perpetual volume surpass spot but also that a widely accepted institutional trust be established, so that traditional asset managers can lawfully use the on-chain perpetual price as a benchmark reference.

These quantitative thresholds reveal an important economic implication: for permissionless financial infrastructure to reshape global asset pricing, it cannot rely on the sophistication of its ideas alone but must meet extremely high engineering standards on the hard metrics of microstructure — liquidity thickness, latency, and friction cost. The migration of pricing power is a strict competition on efficiency, and only when the new system is superior to the old on every dimension can it complete the final substitution.

27.5.6 Case simulation: price transmission of an after-hours event

To make the abstract model above more concrete, we simulate a specific scenario. At 9:00 p.m. on a Friday (Nasdaq has closed, and the after-hours session has ended as well), Elon Musk posts on social media a major tweet announcing a breakthrough in Tesla's autonomous-driving technology. Empirical research shows that Musk's tweets tend to produce measurable abnormal returns in TSLA's stock price within minutes to hours of posting, often accompanied by increased trading activity and volatility [24][25].

Under the traditional market structure, the weekend closure means this major positive news cannot be translated into price immediately. Millions of retail and institutional investors can only crowd in together at the Monday morning open, causing the opening price to show a violent upward gap. This pricing vacuum not only reduces the market's informational efficiency but also forces investors to bear a full 2 days of overnight risk.

Now introduce a TSLA perpetual futures market in the "third stage" (non-trading-hours dominator). After the tweet is posted, algorithmic trading bots instantly detect the shift in sentiment, and a flood of buy orders pours within milliseconds into the around-the-clock on-chain perpetual futures order book. Because the spot market is closed and the oracle price is frozen at Friday's close, the perpetual price rapidly unmoors to the upside, standing about 5% above the spot close within the first 10 minutes. The large basis forces longs to pay shorts an extremely high funding rate, while shorts, tempted by that funding rate, supply liquidity. Over the roughly 60 hours of the weekend that follow, the perpetual market, through ample turnover, independently completes the pricing of this information without a spot anchor, ultimately reaching a new equilibrium at a level 8% above the original close.

When Nasdaq opens at 9:30 a.m. on Monday, market makers in the traditional market no longer need to guess blindly at the opening price. They refer directly to the equilibrium price the on-chain perpetual futures formed over the weekend and use it as the benchmark for setting the spot open. In this specific event, TSLA perpetual futures took on the function of price discovery during non-trading hours; they did not replace the spot market but performed the spot market's work of absorbing information during the closure. This price-transmission path clearly demonstrates how, when the mechanistic advantages of perpetual futures are fully released, they can seamlessly embed in and optimize the pricing network for global assets.

Figure 27-3 uses three liquidity conditions — shadow market, non-trading-hours dominator, and pricing-power reversal — to simulate the evolution of the TSLA perpetual futures price from Friday's close to Monday's open (a 72-hour window) in the scenario above, along with the corresponding implied funding rate dynamics. Its core regularity is that the deeper the liquidity, the faster weekend price discovery converges: the first stage barely responds to the event, whereas the fifth stage converges to a new equilibrium within hours of the event.

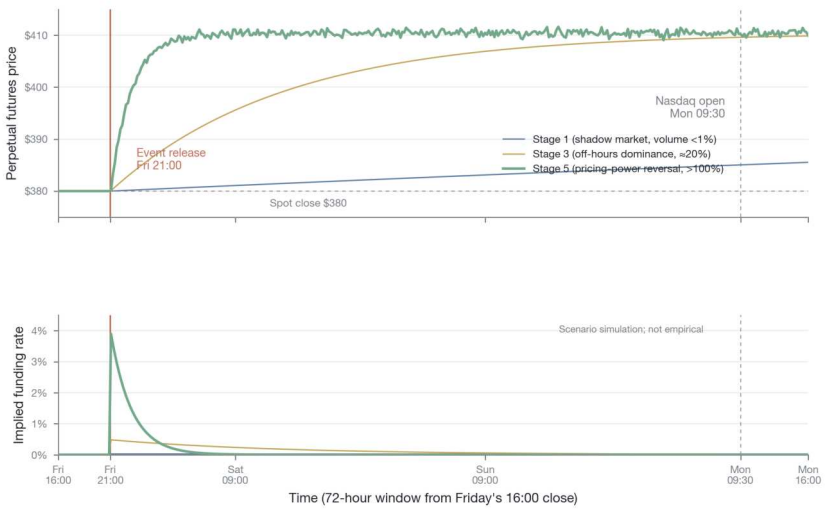


Figure 27-3. TSLA weekend-event simulation: the price-transmission trajectories and funding rate dynamics of perpetual futures at different stages (scenario simulation, not empirical: built by the author on the exponential-convergence model $P(t) = P_0 + \Delta P \cdot (1 - \exp(-\lambda \cdot t))$, with price convergence timed from the event at 9:00 p.m. on Friday; the total shock magnitude of +8% is a scenario shock assumption set by the author, not a point estimate from any study; Dam 2023 [24] supports only the directional conclusion that Musk's Tesla-related tweets produce measurable abnormal returns, not the +8% magnitude; the calibration parameters are TSLA's average daily volume of about $25 billion, a representative magnitude as of March 2026 [18], and an after-hours volume share of about 3% [19])

Figure 27-4 uses a stacked bar chart to quantify the information-incorporation efficiency of the five stages: the green bars represent the price gain already absorbed over the weekend, and the red bars represent the residual gap remaining at Monday's open. Efficiency rises from 17% in the first stage to 97% in the third, exhibiting the same nonlinear leap as Figure 27-1. This shows that when perpetual futures volume reaches 20% of spot (the third-stage threshold), the problem of the weekend pricing vacuum has, in substance, been solved.

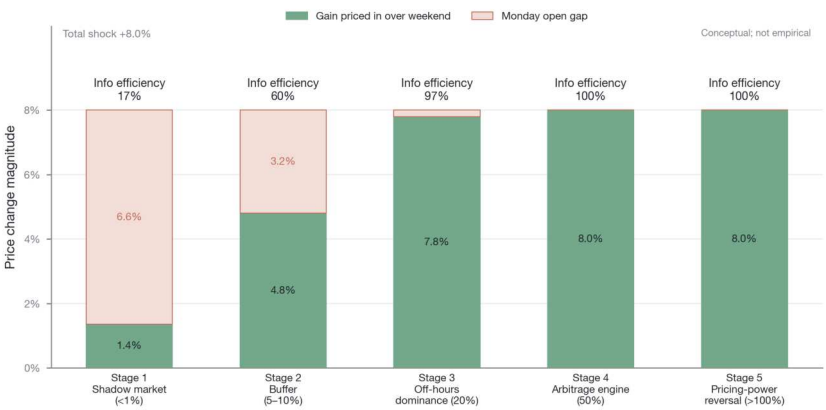


Figure 27-4. The five-stage information-incorporation efficiency of TSLA perpetual futures and the Monday-open gap (conceptual illustration, not empirical: built by the author on the exponential-convergence model $\eta = 1 - \exp(-\lambda \cdot T)$, with five-stage information efficiency of 17% → 60% → 97% → 100% → 100%, derived from the share of the total shock already incorporated at each stage and internally consistent, with an effective convergence window of T = 60.5 hours timed from the event at 9:00 p.m. on Friday to the Monday open; the total shock magnitude of +8% is the author's scenario assumption; the λ values are illustrative choices reflecting the liquidity depth of each stage; Dam 2023 [24] supports only the direction and not the 8% magnitude, and Hasbrouck 1995 [21] is the conceptual anchor for the information-share method — neither is a source of point estimates)

27.6 Thought experiment III: commodities and emerging assets

The pricing mechanism of perpetual futures shows potential not only in crypto assets, foreign exchange, and equity indices; its properties of no expiration date and synthetic exposure also offer a wholly new pricing perspective for commodities and emerging asset classes. This section extends the boundaries of the thought experiment further, examining how perpetual futures could reshape the pricing logic of non-traditional underlying assets — from crude oil and gold to carbon credits, real estate indices, and even prediction markets. The traditional derivatives markets in these areas often suffer from fragmented liquidity, an excessively high barrier to entry, or the outright absence of any effective pricing mechanism, and the introduction of perpetual futures may be not only a complement to existing markets but also a creator of entirely new liquidity.

27.6.1 Crude oil, gold, and agricultural commodities

In the traditional financial system, commodity futures markets are the central hub of price discovery. Take crude oil: the West Texas Intermediate crude oil futures and options of the Chicago Mercantile Exchange together have average daily volume exceeding 1 million contracts and combined open interest of about 4 million contracts, forming the benchmark for global energy pricing [26]. The gold market is likewise vast: open interest in the gold futures of the New York Mercantile Exchange has long held at roughly 400,000 to 550,000 contracts (with a peak of about 900,000 contracts

on a combined-options basis and an all-time record of about 800,000 contracts) [27]. For these highly mature markets, the incremental value of perpetual futures is not to replace the existing liquidity pool but to eliminate the friction costs inherent in traditional delivery futures.

The principal drawback of traditional commodity futures is the roll cost. Because a futures contract has a fixed expiration date, an investor who does not wish to take physical delivery — a commodity index fund or a retail speculator, for instance — must close the position before expiry and buy a forward contract. This periodic roll not only incurs high trading fees but also exposes the investor to the implicit losses of the term structure (such as contango or backwardation). Perpetual futures anchor automatically to the spot price through the funding rate mechanism, eliminating the explicit roll operation and its trading friction (the holding cost is instead expressed as the funding rate). A trader can hold synthetic exposure to crude oil or gold indefinitely, without attending to a complex delivery calendar. Decentralized derivatives platforms such as Hyperliquid have recently reported record-high open interest in their crude oil and gold perpetual futures, providing preliminary confirmation of genuine demand for roll-free commodity exposure [28].

Commodity perpetual futures nonetheless face an insurmountable physical constraint: the need for physical delivery (this is one of the four ultimate constraints on the "Everything Perpetual," discussed in full in Section 27.8.2). Although synthetic exposure holds advantages for price speculation and risk hedging, it can never replace the movement of supply chains in the physical world. In crude oil and gold, therefore, perpetual futures are more likely to evolve into a parallel, "purely financialized" pricing layer that coexists over the long run with the traditional futures markets that bear the function of physical delivery.

Unlike crude oil and gold, the traditional futures markets for agricultural commodities and metals are often characterized by insufficient liquidity. Research shows significant differences in liquidity cost across agricultural futures contracts, with trading volume and the delivery period the key factors affecting liquidity [29]. For many farmers in developing countries, participating in the agricultural futures markets of Chicago or Dalian entails extremely high barriers to opening an account, moving funds across borders, and managing margin. Here, the global access and very low barrier of perpetual futures reveal enormous potential for financial inclusion. Through a smartphone and a crypto wallet, a coffee farmer in South America or a cocoa grower in Africa could use agricultural perpetual futures directly to lock in a future selling price and hedge the risk of price fluctuations. This is the closest realization of the vision Robert Shiller depicted in Macro Markets: enabling groups once excluded from the modern financial system to effectively manage the economic risks they face.

27.6.2 Carbon credits and real estate indices

As the thought experiment extends outward, we enter areas where traditional derivatives markets are weaker still. The carbon-credit market is an important global mechanism for addressing climate change, but the liquidity of the current voluntary carbon market is highly fragmented. Although trading volume in compliance carbon markets

continues to grow, the voluntary carbon market still faces problems such as opaque pricing and high transaction costs, with significant price differences between different types of carbon credits (such as forestry carbon sinks and industrial emissions reduction) [30]. The introduction of carbon-credit perpetual futures could provide this fragmented market with a unified global pricing venue. By indexing carbon credits of a particular standard and issuing perpetual futures on that index as the underlying asset, global capital could participate seamlessly in the pricing of carbon emissions, substantially improving the liquidity and price-discovery efficiency of the carbon market.

Real estate is another quintessential "non-traditional" underlying asset. Housing is not only the principal store of wealth for most households but also one of the most broadly influential variables in the macroeconomy. Real estate, however, is extremely illiquid and difficult to short. The Chicago Mercantile Exchange once launched futures based on the S&P CoreLogic Case-Shiller Home Price Index, but because of the high barrier to entry and insufficient willingness among market makers, its volume has always remained extremely low [31]. Although the Real Estate Index of the Shenzhen Stock Exchange in China reflects the performance of the relevant listed companies, it cannot directly hedge the risk of fluctuations in physical home prices [32].

Real estate index perpetual futures could break this deadlock. Because they require no complex traditional brokerage account and support very fine-grained trade sizes, home-price perpetual futures could attract large numbers of ordinary residents seeking to hedge the risk of home purchases, as well as global capital seeking diversified investment. Once the participant base crosses a critical point, liquidity would reinforce itself, ultimately forming a continuous pricing curve that reflects expectations of future home prices. The common feature of these non-traditional assets is that the liquidity of their traditional derivatives markets is weak or even nonexistent. Perpetual futures need not compete with entrenched incumbents; instead, by lowering the barrier to entry, they create an entirely new liquidity pool "from scratch."

27.6.3 Prediction markets

If the logic of the funding rate anchoring mechanism is pushed to its limit, the underlying asset of perpetual futures need no longer be confined to the price of a financial asset but can extend to any quantifiable variable or discrete event. Prediction markets are the direct application of this generalized logic. In a prediction market, the value of a contract depends on whether a particular event — an election result, an economic data release, or a geopolitical conflict, for example — occurs.

Decentralized prediction markets, exemplified by Polymarket, have grown rapidly. During the 2024 U.S. presidential election, Polymarket's trading volume far surpassed that of the traditional regulated prediction platform Kalshi, and it maintained strong momentum thereafter. Recent data show that Polymarket's single-day notional volume once exceeded $478 million (March 2026) [33]. Academic research on prediction markets indicates that they are highly efficient in price discovery and that in some settings their accuracy can match or even exceed that of traditional opinion polls [34].

From the standpoint of mechanism design, an event contract in a prediction market is essentially a binary perpetual futures contract. Traders continuously price "whether an

event will occur," and the price fluctuates between zero and one, reflecting the market's consensus on the probability of the event. So long as a reliable oracle can adjudicate the event's outcome without dispute, any uncertainty in the real world can be turned into a tradable financial instrument. This trend of financializing temperature changes, macroeconomic indicators, and even the course of geopolitics has provoked a profound controversy over "the financialization of everything." Supporters hold that it greatly extends the boundaries of risk sharing, drawing more latent information into the price system, while critics worry that it may lead to excessive speculation, degrading serious social issues into a global gambling game. This tension constitutes the core challenge of permissionless finance at the level of social ethics.

27.6.4 Technical feasibility and institutional obstacles

In sum, as a general-purpose pricing infrastructure, the trajectory by which perpetual futures extend their coverage from crypto assets to everything displays a clear gradient of technical feasibility and an increasing set of institutional obstacles. Table 27-2 systematically assesses the prospects for turning six categories of underlying asset into perpetual futures across four dimensions: oracle reliability, the strength of traditional-market substitutes, the core incremental value of perpetual futures, and near-term feasibility. This assessment framework will help the reader understand why the speed at which perpetual futures penetrate different asset classes varies so markedly.

Underlying asset class	Oracle reliability	Traditional-market substitute	Core incremental value of perpetual futures	Near-term feasibility assessment
Crypto-native assets	Very high (on-chain native data)	None (perpetual futures already dominate)	Mechanism fully validated and implemented	Fully realized
Major foreign exchange pairs	High (deep liquidity, stable data sources)	Interbank spot market and CME foreign exchange futures	Around-the-clock (24/7) continuous pricing and inclusive global access	Relatively high
Equity indices (such as the S&P 500)	Moderate (limited by the data vacuum during non-trading hours)	ETFs and CME equity index futures	Filling the Asian-session pricing vacuum and providing permissionless global access	Moderate
Commodities (gold/crude oil)	Moderate (dependent on traditional exchange quotes)	Mature commodity futures such as those of the CME and LME	Eliminating roll costs and lowering the barrier to entry	Moderate
Non-traditional assets (carbon credits/home prices)	Relatively low (low data-collection frequency, difficult to standardize)	Market liquidity weak or nonexistent	Creating an entirely new, unified global pricing market for illiquid assets	Relatively low
Prediction markets (events of every kind)	Very low (heavily dependent on manual adjudication and complex fact-checking)	Traditional betting, or no substitute	Realizing the highest-potential risk sharing, while facing the greatest oracle challenge	Very low

Table 27-2. Feasibility assessment of turning different underlying assets into perpetual futures (Data source: constructed by the author on the basis of the three preceding thought experiments)

As shown in Table 27-2, from crypto-native assets to prediction markets, the feasibility of turning assets into perpetual futures displays a clear declining gradient. This gradient is determined mainly by two variables. The first is oracle reliability: as the underlying asset extends from highly standardized financial instruments toward non-standardized physical entities and real-world events, oracle reliability declines exponentially. An oracle must transmit off-chain information to an on-chain contract accurately, with low latency, and tamper-resistantly. For crude oil or an equity index, the oracle need only relay quotes from a traditional exchange; but for a home-price index or a complex political event, the design of the oracle faces enormous game-theoretic challenges. Moreover, when perpetual futures attempt to price carbon credits or macroeconomic indicators, they inevitably touch the regulatory boundaries of sovereign states. Technology can build a permissionless trading layer, but it cannot eliminate the physical world's need for delivery, nor can it bypass the deeply rooted trust structures and institutional constraints of human society. Together, these ultimate constraints draw the real-world

boundary of the "Everything Perpetual" thought experiment.

27.7 The dynamics of pricing-power migration

The migration of pricing power requires several preconditions: liquidity reaching critical mass, institutional recognition, and, for non-crypto-native assets, oracle reliability (the preconditions and timeline of each stage are developed one by one in Section 27.7.5). The speed of migration depends on the overall performance of perpetual futures across several dimensions: efficiency, cost, access, resilience, and fairness.

Given these preconditions, the migration of pricing power is not a sudden event in the financial system but an evolutionary process that advances gradually along liquidity, informational efficiency, and institutional acceptance. If the perpetual futures mechanism does have the potential to take over global asset pricing, this process will not be completed overnight but will follow a clear four-stage dynamic model. Each stage marks a structural reorganization of the relationship between on-chain pricing infrastructure and the traditional financial system, and crossing each stage requires a particular level of technical maturity and specific institutional preconditions as support.

27.7.1 Stage one: crypto-native assets

The first stage of pricing-power migration is not a theoretical extrapolation but an empirical fact that has already occurred. In this stage, perpetual futures successfully established their status as the core pricing venue for a particular asset class. For crypto-native assets such as Bitcoin and Ethereum, the perpetual futures market has already surpassed both the spot market and the traditional delivery futures market to become the dominant venue for price discovery.

According to data on the crypto derivatives market in 2025, total volume in this space reached about $85.7 trillion, with average daily volume exceeding $260 billion, and perpetual futures occupied the dominant position [4]. In the quantitative analysis of price discovery, as cited in Section 27.5.1, crypto derivatives markets already command about half of the information share in Bitcoin price discovery [20][21][22]. Given that perpetual futures surpass quarterly futures in both volume and continuity, their share in price discovery may be higher still; but there is at present no direct point estimate of the information share of perpetual futures specifically, so this judgment is a qualitative inference based on volume and continuity. The microfoundation of this phenomenon is that the built-in leverage, seamless shorting mechanism, and around-the-clock continuous trading of perpetual futures make informed traders more inclined to express their informational advantage in such venues.

The completion of the first stage proves the effectiveness of the funding rate anchoring mechanism. It shows that, absent the constraint of physical delivery, pure synthetic exposure can still bear enormous sums and sustain an efficient price-discovery function. Yet the success of this stage owes much to the characteristics of crypto assets themselves: they are digitally native, have no hard binding to physical-world delivery, and, in their early development, faced little competition from strong traditional-finance substitutes.

27.7.2 Stage two: shadow pricing

When the underlying assets of perpetual futures extend from crypto-native assets to traditional financial assets — foreign exchange, equity indices, or commodities — pricing-power migration enters its second stage. In this stage, on-chain perpetual futures begin to provide price signals for traditional assets, but these prices have not yet won the institutional recognition of the traditional financial system. They play the role of a "shadow price" in the system, analogous to the shadow-pricing role that offshore currency markets play relative to the onshore official exchange rate.

The core value of shadow pricing lies in filling the structural gaps of traditional markets, especially gaps in time coverage. Consider the Asian-session pricing vacuum of SPY, the core ETF product on the S&P 500 index (its causes are discussed in Section 27.4) [35]: if a major macroeconomic event or geopolitical shock occurs during this period, the traditional market can offer no venue for immediate price discovery, and the around-the-clock on-chain perpetual futures become the only active pricing source.

Although in this stage the trading volume of on-chain perpetual futures may account for only a very small proportion of total traditional-market volume — 0.1% to 1%, say — its monopoly on information gives it a disproportionate influence over pricing. When the traditional market reopens after a closure, market makers and algorithmic trading systems refer to the perpetual futures' price movements during the closure to set the opening quote. This information-transmission mechanism allows the shadow price to materially influence the price formation of the traditional market, even before it is officially recognized as a benchmark price.

27.7.3 Stage three: institutional recognition

The third stage of pricing-power migration marks the initial narrowing of the trust gap between the on-chain pricing system and the traditional financial system. In this stage, the liquidity depth and oracle reliability of the perpetual futures market reach critical mass, prompting regulators, asset managers, and traditional market makers to begin formally citing or relying on on-chain price data, thereby granting it institutional recognition.

A precedent for this recognition has already appeared in the 2024 approval of the Bitcoin ETF. The fund was the first institutional bridge between traditional finance and crypto finance, and its pricing mechanism relies indirectly on the liquidity and information flow of the perpetual futures market (the microstructure of this triangular-arbitrage price transmission, intermediated by the Chicago Mercantile Exchange futures market, is detailed in Section 27.9.2).

For non-crypto assets, reaching the stage of institutional recognition requires overcoming more complex obstacles. First, traditional market makers must participate in on-chain market-making across the boundary at scale, to ensure that prices in the two markets are tightly coupled. Second, oracle infrastructure must demonstrate its absolute reliability under extreme market conditions, particularly how it provides a manipulation-resistant reference price while the traditional market is closed. When these conditions are met, the information share of perpetual futures will rise signifi-

cantly, possibly reaching the 20% to 35% range, making them a major pricing venue on a par with traditional exchanges.

27.7.4 Stage four: infrastructure integration

The ultimate form of the four-stage model is the deep integration of infrastructure. In this stage, the binary opposition between "traditional finance" and "on-chain finance" is broken down, and the two merge into different execution layers of a single unified global pricing system. Perpetual futures are no longer a challenger to, or a complement of, the traditional market but become a core component of the underlying infrastructure.

Under this integrated architecture, traditional exchanges might directly operate a clearing layer based on distributed ledger technology (DLT), or introduce the perpetual futures mechanism into their core product lines. A trader could switch seamlessly on a single interface among traditional spot, regulated futures, and permissionless perpetual futures, with all risk exposures net-settled within a single unified global liquidity pool. Such a convergence would not only resolve the settlement-delay risk that has long existed in the traditional foreign exchange market but also greatly widen the access boundary for global capital [36].

Realizing the infrastructure-integration stage would mean a fundamental reconstruction of the underlying logic of financial markets. It would require the regulatory framework to shift from territorial oversight based on institutional entities to global coordination based on code function and protocol logic. In this stage, the vision of perpetual futures as a general-purpose pricing layer would be fully realized, and any asset could be turned, through the funding rate mechanism, into synthetic exposure that is globally accessible, tradable around the clock, and instantly settled.

Figure 27-5 visualizes the four-stage model with the qualitative stages of evolution (from early to distant) on the horizontal axis: stage one, "crypto-native assets," is essentially complete (the information share for BTC and ETH has reached about 50%); stage two, "shadow pricing," is now taking place; and stage three, "institutional recognition," and stage four, "infrastructure integration," are anticipated extrapolations, with the key preconditions for crossing each stage marked at the junctions. The core lesson of this evolutionary path is that the speed of pricing-power migration is determined not by technological progress alone but by the slower of technical maturity and institutional acceptance.

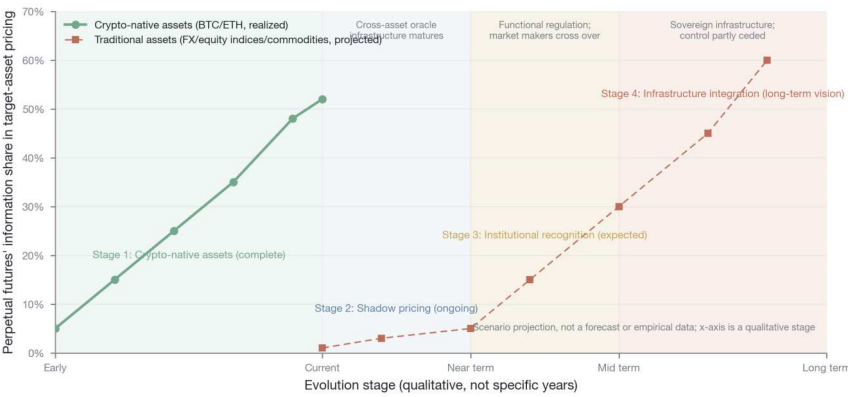


Figure 27-5. The four-stage model of pricing-power migration: stages of evolution and key preconditions (scenario extrapolation, not a forecast and not empirical: built by the author on the four-stage model, with the horizontal axis showing qualitative stages of evolution from early to distant rather than specific years; the stage-one figure "BTC/ETH information share ≈ 50%" draws on Alexander & Heck 2020 [20], which uses the Hasbrouck 1995 information-share method [21], and the figure that "perpetuals account for about 60% of BTC's total market volume" draws on the CoinGlass crypto derivatives report [4]; De Blasis & Webb 2022 [22] is corroborating evidence of derivatives dominance and spillover effects, does not itself provide an information-share estimate, and is not on its own a source for the 50% information share; stages two through four are framework-based scenario projections)

27.7.5 Preconditions and timeline

Each stage of pricing-power migration is subject to particular technical and institutional constraints, and these constraints determine the speed and boundaries of the evolution. The first stage (crypto-native assets) is complete, having taken nearly a decade (2016–2025). The second stage (shadow pricing of traditional assets) is now taking place: several decentralized exchanges have launched perpetual futures on foreign exchange, equity indices, and commodities, completing the proof of concept and the early accumulation of liquidity.

The leap from the second stage to the third (institutional recognition) is the most difficult transition in the entire dynamic model, because it goes beyond engineering and enters the reconstruction of social trust and institutional design. The preconditions for this leap include building a decentralized oracle network capable of withstanding state-level attacks, forming a functional regulatory framework interoperable with traditional finance, and establishing deep trust in smart contract code among institutional investors. Based on the historical pace at which financial infrastructure evolves, realizing this stage may take 5 to 10 years.

As for the fourth stage (infrastructure integration), its timeline is more distant and fraught with uncertainty, and it may take 10 to 20 years to unfold. It depends not only on the full maturation of the technical and regulatory conditions described above but also on a fundamental cession of sovereign states' control over financial markets. The rigid demands of sovereign states regarding capital flows, monetary policy transmission, and anti-money-laundering compliance constitute the fundamental constraint on this

ultimate picture. The complete migration of pricing power may therefore forever be an asymptote — infinitely approached but never fully reached — and it is along this asymptote that perpetual futures will continue to reshape the boundaries of global asset pricing.

27.8 The ultimate constraints

A thought experiment must confront its insurmountable boundaries honestly. In examining the potential of perpetual futures as a general-purpose pricing layer, we must distinguish which constraints are technical, which are institutional, and which are fundamental. Technical constraints can be solved through engineering innovation, and institutional constraints may loosen as regulation evolves, but fundamental constraints reflect the deep structure of human society and the objective laws of the physical world; they will not vanish with technological progress. This section defines the ultimate picture of this thought experiment and analyzes in depth the four ultimate constraints that limit its unbounded expansion.

27.8.1 The ultimate picture

In the limiting case where all technical and institutional constraints are perfectly resolved, perpetual futures could constitute a fully comprehensive global pricing layer. In this picture, anyone, anywhere, at any moment can trade and price any asset. All financial exposure is turned into synthetic products that settle instantly, are globally accessible, and operate continuously around the clock. This eliminates not only the geographic and temporal fragmentation of traditional financial markets but also the access barriers between different asset classes.

The value of this ultimate picture lies not in offering a definite forecast but in revealing the theoretical limit of financial-market evolution. It shows how high the efficiency of price discovery could rise if friction costs fell to zero. In this ideal state, the funding rate mechanism anchors synthetic assets to their underlying perfectly, the oracle network transmits information with no latency and absolute reliability, and smart contracts execute all clearing and settlement logic with a zero failure rate. Yet this perfect state is merely a theoretical asymptote. The physical properties, political structures, and human psychology of the real world constitute insurmountable boundaries, so that this ultimate picture can only be approached infinitely and never fully realized.

27.8.2 The settlement needs of the physical world

There is an essential economic difference between synthetic exposure and physical delivery, and this is the first ultimate constraint that prevents perpetual futures from fully replacing traditional markets. Synthetic products can provide perfect price exposure, but they cannot meet the physical needs of the real economy. A refinery needs real crude oil to keep running, a flour mill needs real wheat to make bread, and a multinational enterprise needs real foreign currency to pay overseas suppliers. However deep the liquidity of a crude oil perpetual futures contract, and however efficient its price discovery, it cannot be refined into gasoline.

In traditional futures markets, although the overwhelming majority of contracts are ultimately closed through cash settlement rather than physical delivery, the possibility of delivery is the ultimate guarantee that the futures price anchors to the spot price. This delivery mechanism builds a solid bridge connecting financial derivatives to physical reality [37]. Perpetual futures attempt to substitute the funding rate mechanism for this anchoring function of physical delivery. For crypto-native assets, the substitution is successful, because crypto assets are themselves digital and have no logistical friction of physical delivery. But when we extend perpetual futures to commodities or agricultural products, whether the funding rate can fully substitute for the anchoring force of physical delivery under extreme market stress remains an unverified assumption.

Perpetual futures may therefore come to dominate pricing power over many assets in the future and become the primary venue for price discovery, but the physical need for settlement and delivery means that physical markets and traditional delivery futures markets must coexist with them. The two would form a division of labor: perpetual futures provide the most efficient information processing and price discovery, while traditional markets handle the logistics and settlement of the physical world. This dual structure is the inherent constraint that the physical world imposes on purely digital finance.

27.8.3 The financial power of sovereign states

The second ultimate constraint arises from the political-economy foundations of the modern nation-state system. The power to issue currency, to control foreign exchange, and to regulate financial markets constitutes the core power of a sovereign state. A fully permissionless global financial market that crosses national borders is, in essence, a direct challenge to a state's financial sovereignty. History shows that sovereign states will not lightly surrender these powers, nor will they remain indifferent to a technology that could weaken the transmission of their monetary policy [38].

The stability of money depends not only on technical architecture but also, and more importantly, on credible public support and the enforcement of par value. In successive financial crises, stability was restored not by the public's access to a particular payment technology but by the ability of the central bank's balance sheet to absorb systemic risk [39]. A permissionless financial network, for all the efficiency it displays in day-to-day operation, lacks a lender of last resort capable of creating liquidity without limit when faced with a systemic collapse.

Faced with the expansion of permissionless finance, the response of sovereign states need not be a simple prohibition; it is more likely to be adaptation and discipline. A state might upgrade its own monetary infrastructure by introducing a central bank digital currency (CBDC), or exert control by bringing decentralized protocols into a new regulatory framework. This means that the globalizing vision of perpetual futures will always operate within the space constrained by national sovereignty. Different jurisdictions will inevitably differ in the degree to which they accept these financial instruments and in the regulatory requirements they impose, leading to a degree of fragmentation in global liquidity. The utopia of fully eliminating national borders and regulatory arbitrage is unattainable so long as the system of sovereign states exists.

27.8.4 The oracle problem

At the technical level, the oracle problem is the foremost obstacle to extending perpetual futures to non-crypto assets. The oracle is the key infrastructure connecting isolated blockchain networks to data from the external world [40]. For crypto-native assets, an oracle can obtain the price directly from the on-chain transactions of a decentralized exchange, and the data source is itself trustless. But for traditional assets such as equities, foreign exchange, and commodities, the oracle must rely on off-chain data sources, such as the market feeds of Bloomberg, Reuters, or a traditional exchange.

This gives rise to a fundamental trust paradox: the goal of the entire decentralized finance system is to build a trustless architecture, yet its acquisition of the underlying asset's price requires absolute trust in these centralized off-chain data sources. The degree to which a system is trustless can never exceed the degree to which its oracle is trustworthy. Even if a decentralized oracle network is used to reduce single-point-of-failure risk through multi-node aggregation, that decentralization is merely superficial if all nodes ultimately depend on the same data interface from Nasdaq [41].

Moreover, during the non-trading hours when the traditional market is closed, the oracle faces the predicament of having no real-time price to report. It can then rely only on alternative data sources or historical data, which greatly weakens the pricing quality and anchoring efficiency of perpetual futures during these periods. Although cryptography and mechanism design can mitigate these problems, so long as the external world's data-generation process is not decentralized, the oracle problem will forever be a technical and trust bottleneck that cannot be entirely eliminated.

27.8.5 Human trust structures

The final ultimate constraint is rooted in the very nature of human cognition and social structure. Technology can shift the object of trust and reduce dependence on particular intermediaries, but it can never fully eliminate the need for trust itself. In traditional finance, participants trust banks, exchanges, and the legal and regulatory systems behind them. In a permissionless financial system, this trust is displaced onto the code logic of smart contracts, the security of cryptographic primitives, the honesty of the oracle network, and the governance mechanisms of DAOs [42].

Although this displacement of trust increases the transparency and verifiability of the system, it does not achieve genuine "trustlessness." For the overwhelming majority of market participants, verifying complex smart contract code or assessing the game-theoretic security of an oracle network far exceeds their cognitive capacity. They are in fact trusting the reputation of code-audit firms, the expertise of core developers, or simply the system's past record of stable operation.

The ultimate constraints of the thought experiment are therefore not only physical, political, and technical but also sociological and psychological. A fully trustless global financial market is a theoretical ideal that points the way toward lower friction and higher efficiency, but the inherent complexity of human society and its psychological need for certainty mean that some form of institutional trust will always be an indispensable foundation of the financial system.

27.9 Anchoring in reality

The value of a thought experiment lies not only in exploring the boundaries of theory but also in providing an analytical framework for observing reality. The foregoing extrapolation of perpetual futures as a universal global pricing infrastructure is not pure logical fantasy; its core elements have already begun to appear in the real world. This section anchors the theoretical extrapolation of the thought experiment to the current reality of financial markets, showing which transformations have already occurred, which infrastructure is being built, and which obstacles the path to the future vision must cross.

27.9.1 Early practice of the Everything Perpetual

Extending the underlying logic of perpetual futures from crypto-native assets to a broader range of financial instruments has already been validated on decentralized exchanges. The decentralized exchange Hyperliquid, for example, offers preliminary empirical evidence for the concept of the Everything Perpetual through its all-asset strategy. The platform not only lists perpetual futures on crypto assets but has also introduced exposure to traditional financial assets through its permissionless builder-deployed market framework (HIP-3).

In implementation, the platform has successfully launched perpetual futures products covering foreign exchange, equity indices, and commodities. Using the same margin system in a fully permissionless environment, a trader can obtain synthetic exposure to the stocks of traditional technology companies such as Nvidia and Tesla, as well as to macro commodities such as gold and crude oil. This aggregation of liquidity across asset classes proves the universality of the funding rate anchoring mechanism: it does not, in fact, depend on the specific physical properties of the underlying asset but requires only a reliable price feed from an oracle.

The evolution of market data further confirms that this cross-boundary demand genuinely exists. Measured by share of average daily volume, from the fourth quarter of 2025 to the first quarter of 2026 the share of traditional assets in Hyperliquid's total volume underwent a marked structural leap, climbing from less than 5% to more than 30% (about 31.6% by the end of January 2026, according to Delphi Digital data) [43]. The structural advantage of the around-the-clock trading of perpetual futures is especially evident in the face of sudden geopolitical events. In late February 2026, for instance, during tensions in the Middle East, the traditional commodity markets were closed for the weekend, and capital seeking to hedge risk poured into the decentralized perpetual futures market, driving a surge in the trading volume of synthetic crude oil and gold assets [44]. This episode bears out the expectation about the advantage of continuous pricing set out in the thought experiment above.

27.9.2 The bridging role of the Bitcoin ETF

Even as the permissionless on-chain financial system extends into the domain of traditional assets, the traditional financial system is actively building channels toward crypto-native assets. The approval of the spot Bitcoin ETF in 2024 marked the es-

tablishment of the first key institutional bridge between traditional finance and crypto finance. The significance of this event lies not only in providing traditional capital with a compliant investment channel but also in connecting two very different pricing ecosystems at the level of microstructure.

From the perspective of pricing-power migration, the introduction of the ETF significantly changed the market's information-transmission mechanism. The influx of traditional institutional capital caused the assets under management of the spot ETF to swell rapidly, and more than 2,000 U.S. investment advisory firms began incorporating such assets into their allocation portfolios [45]. Yet the ETF's own pricing mechanism depends heavily on the arbitrage activity of market makers and authorized participants. In executing their arbitrage strategies, these institutional participants inevitably need to hedge risk using the standardized futures contracts of the Chicago Mercantile Exchange and the more deeply liquid offshore perpetual futures market.

This cross-market arbitrage forms a hidden but efficient chain of price transmission. When large-scale inflows or outflows occur in the traditional market during the regular session, this information is reflected first in the ETF's premium or discount, then transmitted through market makers' hedging operations to the futures market of the Chicago Mercantile Exchange, and finally resonates in price with the perpetual futures market. Although this connection remains indirect at the current stage and is constrained by the trading hours of the traditional market, it establishes an ever-deepening interconnection of liquidity between the two worlds, laying the institutional foundation for fuller infrastructural integration in the future.

27.9.3 The evolutionary roadmap

A significant gap separates today's nascent all-asset perpetual futures from the ultimate vision of genuinely taking over pricing power for core global assets. Closing this gap requires not only technical breakthroughs at the engineering level but also a systematic evolution of institutional and trust structures.

At the level of technical infrastructure, oracle reliability remains the core bottleneck constraining the scaling of perpetual futures on non-crypto assets. When the traded underlying shifts from crypto assets with native on-chain prices to traditional assets that depend on external data sources, the oracle network must be able to provide accurate, latency-free, manipulation-resistant price feeds under extreme market volatility. Moreover, to bear the enormous transaction throughput of traditional financial markets, the clearing efficiency and concurrent processing capacity of the underlying blockchain must still improve by orders of magnitude.

At the level of institutions and social consensus, establishing a functional regulatory framework is unavoidable. The bottom-line demands of sovereign states regarding monetary policy transmission and anti-money-laundering compliance mean that a shadow-pricing system entirely detached from regulation can hardly win broad adoption by mainstream financial institutions. The more likely evolutionary path is a gradual convergence: preserving the core advantages of perpetual futures — continuous trading and efficient settlement — while achieving selective compliance verification at the application layer through cryptographic techniques such as zero-knowledge proofs. The

timeline of this process is fraught with uncertainty, but the direction of technological evolution and the market's intrinsic demand for a more efficient pricing mechanism have already pointed the way for this potentially transformative change to financial infrastructure.

27.10 Chapter summary

The universality of the funding rate anchoring mechanism is the logical starting point of this chapter's thought experiment. This mechanism does not depend on the physical or financial properties of the underlying asset and can, in principle, turn any continuously priceable asset into perpetual futures that trade around the clock, are globally accessible, and settle instantly [1]. The "faster, cheaper, more accessible" driving logic of financial history was confirmed repeatedly across three migrations — from spot to futures, from the trading floor to the electronic screen, and from delivery futures to ETFs [46]. Perpetual futures hold a structural advantage on the first two of these dimensions but still face a significant gap on the "more trustworthy" dimension.

The three thought experiments in this chapter reveal a consistent pattern: perpetual futures offer distinctive structural incremental value in each asset class — around-the-clock pricing in foreign exchange, Asian-session coverage for equity indices, and zero roll cost for commodities — but each also faces the hard constraints of oracle reliability, regulatory compliance, and the need for physical delivery. The ultimate picture of the "Everything Perpetual" is bounded by four insurmountable limits: the settlement needs of the physical world, the financial power of sovereign states, the trust paradox of the oracle, and the trust structures inherent in human society [47].

Returning to the counterfactual with which the chapter opened: if the Chicago Board of Trade had launched perpetual futures rather than delivery futures in 1848, the globalization of pricing power might have arrived earlier, but the constraints of physical delivery and sovereign control would not have vanished as a result. What perpetual futures can do is push the efficiency frontier of financial markets as far as possible toward its limit, within these irremovable constraints. The next part of the book, Part 10, "Trustless Engineering," examines how, through decentralized architectural design, one might build an on-chain financial system that preserves the offensive advantages of perpetual futures while repairing their defensive deficiencies.

References

[1] Ackerer, D., Hugonnier, J., & Jermann, U. (2025). Perpetual futures pricing. *Mathematical Finance*. https://doi.org/10.1111/mafi.70018

[2] Chainalysis. (2026). *An introduction to perpetual futures*. https://www.chainalysis.com/blog/perpetual-futures/

[3] Shiller, R. J. (1993). *Macro markets: Creating institutions for managing society's largest economic risks*. Oxford University Press.

[4] CoinGlass. (2025). *2025 Crypto Derivatives Market Annual Report*. https://www.coinglass.com/learn/2025-annual-report-en

[5] CME Group. (n.d.). *History of the Chicago Board of Trade*. https://www.cmegroup.com/company/history/

[6] CME Group. (n.d.). *CME Globex overview*. https://www.cmegroup.com/globex.html

[7] BlackRock / iShares. (2025). *2025 ETF & ETP Market Trends: Flow and Tell Year in Review*. https://www.ishares.com (global and U.S. ETF industry totals additionally per ETFGI 2025 annual data)

[8] CME Group. (2026). *E-mini S&P 500 futures overview*. https://www.cmegroup.com

[9] Bank for International Settlements. (2025). *Triennial central bank survey of foreign exchange and OTC derivatives markets*. https://www.bis.org/statistics/rpfx25.htm

[10] U.S. Securities and Exchange Commission. (2023). *Shortening the Securities Transaction Settlement Cycle* (Final Rule, Release No. 34-96930; T+1 effective May 28, 2024). https://www.sec.gov/rules/final/2023/34-96930.pdf

[11] Mohanty, M. S. (n.d.). Central bank views on foreign exchange intervention. *BIS Papers*. https://www.bis.org/publ/bppdf/bispap73e_rh.pdf

[12] U.S. Department of the Treasury. (2023). *Illicit finance risk assessment of decentralized finance*. https://home.treasury.gov/system/files/136/DeFi-Risk-Full-Review.pdf

[13] MacroMicro. (2026). *US - SPDR S&P 500 ETF Trust trading volume*. https://en.macromicro.me

[14] Schwab Asset Management. (2024). *Understanding the ETF creation and redemption mechanism*. https://www.schwabassetmanagement.com

[15] CoinDesk. (2026). Kraken rolls out crypto-style, 24/7 perpetuals trading for tokenized U.S. stocks. https://www.coindesk.com

[16] RootData. (2026). *The next trillion-dollar battlefield for on-chain derivatives*. http://www.rootdata.com

[17] Chainlink Blog. (2026). Chainlink launches 24/5 U.S. equities streams. https://blog.chain.link

[18] Yahoo Finance. (2026). Tesla, Inc. (TSLA) stock price, news, quote & history. Retrieved March 7, 2026, from https://finance.yahoo.com/quote/TSLA/

[19] CNBC. (2026). TSLA: Tesla Inc - stock price, quote and news. Retrieved March 7, 2026, from https://www.cnbc.com/quotes/TSLA

[20] Alexander, C., & Heck, D. F. (2020). Price discovery in Bitcoin: The impact of unregulated markets. *Journal of Financial Stability*, 50, 100776. https://doi.org/10.2139/ssrn.3583843

[21] Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *The Journal of Finance*, 50(4), 1175-1199.

[22] De Blasis, R., & Webb, A. (2022). Arbitrage, contract design, and market structure in bitcoin futures markets. *Journal of Futures Markets*, *42*(3), 492-524. https://doi.or

g/10.1002/fut.22305

[23] Biais, B., Foucault, T., & Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial Economics*, *116*(2), 292-313. https://doi.org/10.1016/j.jfineco.2015.03.004

[24] Dam, J. G. (2023). *CEO's tweets and firm stock returns: A case study of Elon Musk and Tesla* (Honors College Theses No. 850). Georgia Southern University.

[25] VeriDelisi. (2025). How Elon's tweets affect Tesla stock price. https://veridelisi.substack.com/p/elon-talks-tesla-falls-the-more-elon

[26] CME Group. (2025). *WTI crude oil futures volume*. https://www.cmegroup.com/markets/energy/crude-oil/light-sweet-crude.volume.html

[27] MacroMicro. (2025). *Global gold futures open interest by exchanges*. https://sc.macromicro.me/charts/110916/global-gold-futures-open-interest-by-exchanges-stacked-bar-chart

[28] LiquidityFinder. (2026). Hyperliquid achieves record open interest in perpetual contracts for real-world assets. https://liquidityfinder.com/news/hyperliquid-achieves-record-open-interest-in-perpetual-contracts-for-real-world-assets-6a09e

[29] Wang, H. H., & Ke, B. (2005). Efficiency tests of agricultural commodity futures markets in China. *Australian Journal of Agricultural and Resource Economics*, 49(2), 125-141. https://doi.org/10.1111/j.1467-8489.2005.00283.x

[30] KAPSARC. (2025). *Voluntary carbon market (VCM) total value, volume, price*. https://datasource.kapsarc.org/explore/assets/voluntary-carbon-market-transaction-value-volume-and-price/

[31] CME Group. (n.d.). *S&P CoreLogic Case-Shiller home price index futures*. https://www.cmegroup.com/markets/real-estate/residential/SandP-702702case-shiller-702702home-702702price-index.html

[32] Investing.com. (2025). *SZSE Real Estate Index (SZREI)*. https://www.investing.com/indices/szse-real-estate

[33] Yahoo Finance. (2026). Polymarket breaks $478 million record as Kalshi Khamenei market faces backlash. https://finance.yahoo.com/news/polymarket-breaks-478-million-record-193853484.html

[34] Bossaerts, P., et al. (2024). Price formation in field prediction markets: The wisdom in the crowd. *Journal of Financial Markets*, *68*. https://www.sciencedirect.com/science/article/pii/S1386418123000794 https://doi.org/10.1016/j.finmar.2023.100881

[35] Market Chameleon. (2024). *SPY best hour/worst hour State Street SPDR S&P 500 ETF Trust*. https://marketchameleon.com/Overview/SPY/Stock-Price-Action/Best-and-Worst-Hour-of-Trading

[36] Bank for International Settlements. (2023). *Supervisory guidance for managing settlement risk in foreign exchange transactions*. https://www.bis.org/publ/bcbsc1310.pdf

[37] Loader, D. (2005). *Clearing and settlement of derivatives*. Butterworth-Heinemann.

[38] Buchan, J., & Ungor, M. (2026). Monetary sovereignty in the digital age: The

role of central bank digital currencies. *International Journal of Political Economy*. https://doi.org/10.1080/08911916.2026.2624900

[39] Reichlin, L. (2026). Central bank digital currency and monetary sovereignty. *VoxEU, CEPR*. https://cepr.org/voxeu/columns/central-bank-digital-currency-and-monetary-sovereignty

[40] Chainlink. (2025). *The blockchain oracle problem*. https://chain.link/education-hub/oracle-problem

[41] Caldarelli, G. (2020). Understanding the blockchain oracle problem: A call for action. *Information*, 11(11), 509.

[42] Bangui, H., & Buhnova, P. (2024). When trustless meets trust: Blockchain consensus review and reconsideration. *Procedia Computer Science*.

[43] Yahoo Finance. (2026). How Hyperliquid's TradFi edge could lift HYPE price 90%. https://finance.yahoo.com/news/hyperliquid-tradfi-edge-could-lift-200000961.html

[44] Bloomberg. (2026). Crypto traders hedge Iran war risks with 24/7 oil, gold perpetuals. https://www.bloomberg.com/news/articles/2026-02-28/oil-gold-rally-on-hyperliquid-as-24-7-crypto-prices-iran-risk

[45] TRM Labs. (2026). The rise of crypto ETPs: How a fringe idea became a pillar of institutional adoption. https://www.trmlabs.com/resources/blog/the-rise-of-crypto-etps-how-a-fringe-idea-became-a-pillar-of-institutional-adoption

[46] Working, H. (1953). Futures trading and hedging. *The American Economic Review*, 43(3), 314-343.

[47] Cong, L. W., & He, Z. (2019). Blockchain disruption and smart contracts. *The Review of Financial Studies*, 32(5), 1754-1797. https://doi.org/10.1093/rfs/hhz007

From the Monolithic Black Box to Layered Architecture

In November 2022, FTX—one of the top three cryptocurrency exchanges by global trading volume—collapsed within days, revealing a gap of more than $8 billion between customer assets and corporate liabilities [1]. The bankruptcy investigation exposed a central engineering fact: the platform had never technically segregated customer assets from corporate assets, and all funds were commingled in a wallet system that lacked cryptographic access controls [2][3]. The affiliated trading firm Alameda Research received a secret exemption from the automated liquidation system, and before the system collapsed, no publicly available method allowed outside observers to detect these violations [4]. This event was not an isolated case; rather, it laid bare a structural flaw in the architectural design of centralized exchanges.

This flaw can be defined as the monolithic black box architecture. In a traditional centralized exchange, the core functions of the trading system—including the matching engine, the clearing and settlement system, asset custody, the market-making system, and the risk-control system—are all integrated and run within a single opaque server cluster controlled by one entity: the exchange itself. The data that users see in the front-end interface and through the API are merely the selective output of this black box. Inside the box, users have no way whatsoever to verify whether assets are being kept safe, whether orders are matched honestly according to price-time priority, or whether liquidation parameters are being secretly altered to trigger improper liquidations.

When we apply the lens of systems engineering to this architecture, the lesson of FTX is no longer merely an isolated case of *one management team's moral failure*; it is a textbook instance of systemic architectural failure. The monolithic black box architecture concentrates all of a system's operating trust assumptions in a single entity, thereby creating an enormous single point of trust failure across the entire financial ecosystem. In distributed-systems theory, the failures of different components are typically independent of one another. In a monolithic trust architecture, however, because all functions share the same root of trust, once that root becomes corrupted or collapses, the propagation of failure is total and instantaneous. The system moves directly from a state of seemingly flawless operation into one of complete collapse, with no intervening buffer or graceful degradation.

This brings us to the central question of this chapter: starting from the first principles of engineering, can we design a fundamentally different architecture for a trading sys-

tem? Such an architecture must preserve the performance advantages of centralized exchanges (low latency and deep liquidity) while decomposing the *monolithic black box* into multiple independent and verifiable modules, minimizing the trust dependency at each one. Its core objective is to build a highly resilient system in which, even if the operating entity of one module suffers a moral hazard or a technical failure, the core functions of the whole system and the safety of user assets remain protected by cryptography and consensus mechanisms.

Solving this problem is not a simple binary choice between a centralized exchange (CEX) and a decentralized exchange (DEX). The analyses in the preceding nine parts have repeatedly exposed a central tension: centralized exchanges deliver the best trading performance but require users to bear extremely high trust risk, whereas early decentralized exchanges eliminated trust risk through fully on-chain execution yet were constrained by the physical limits of blockchains and could not meet the microsecond latency and high throughput that professional trading demands. We therefore need an engineering solution that strikes an optimal balance between *trust minimization* and *performance maximization*. This search is not without precedent. Over several centuries, and through repeated crisis-driven reform, traditional finance gradually built a layered trust architecture that separates matching, clearing, custody, and regulation. The establishment of institutions such as the central counterparty (CCP) and the DTCC was, in essence, a decomposition of the early monolithic financial black box [5]. The cryptocurrency market, however, is attempting to achieve this goal through technical code rather than regulatory agencies. A new generation of infrastructure, exemplified by application-specific blockchains (app-chains) such as Hyperliquid, is demonstrating the feasibility of this migration from monolithic to layered design [6]. Through customized consensus mechanisms and specially optimized state machines, these systems are attempting to find a new and efficient frontier between performance and trust.

This chapter serves as the overall blueprint for building a trustless financial system (Figure 28-1), guiding the reader from the perspective of an analyst to that of a systems architect. It first uses the trust stack to dissect the implicit trust embedded in the life cycle of a single trade; it then traces, along the *trust-performance frontier*, the migration path from the monolithic black box to a layered, transparent architecture; it proceeds to conduct an architectural audit of representative protocols such as Hyperliquid; and it finally distills these into an engineering framework for judging *sufficient decentralization*.

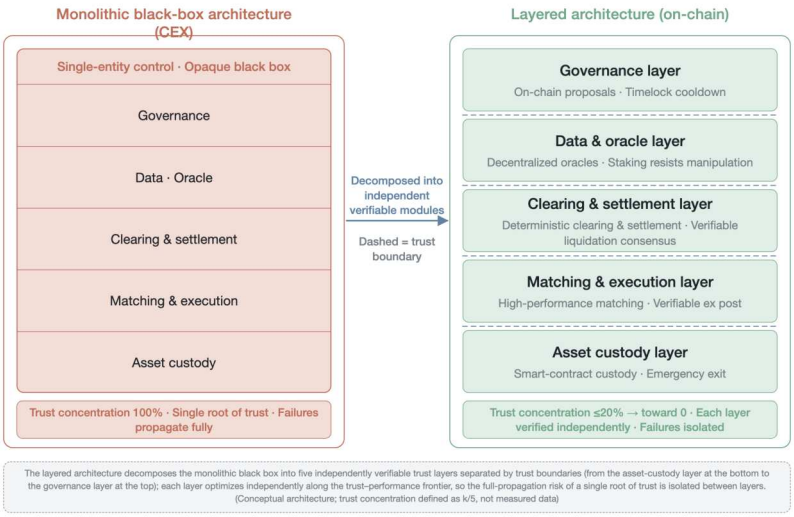


Figure 28-1. A comparison of the monolithic black box architecture and the layered architecture (conceptual schematic: on the left, a centralized exchange fuses five functions—asset custody, matching and execution, clearing and settlement, data oracles, and governance—into an opaque black box controlled by a single entity, with trust concentration reaching 100%, so that once the single root of trust is corrupted, failure propagates completely and instantaneously; on the right, the layered architecture decomposes the same five functions into mutually independent, separately verifiable trust layers, running from the asset-custody layer at the bottom to the governance layer at the top, with each layer optimized independently along the trust-performance frontier and isolated from the others by trust boundaries, lowering trust concentration to 20% or even close to 0)

28.1 A trust anatomy of the CEX monolithic black box

Before discussing how to build a trustless financial system, we must first identify precisely what the current system requires users to trust. When we say that centralized exchanges carry structural risk, that risk is not an abstract notion; it is embedded deep in their architectural design. In using a centralized exchange, a user makes at least five implicit trust assumptions over the life cycle of every trade. Most users are entirely unaware of them—until a crisis erupts and these implicit assumptions are systematically violated.

28.1.1 The five trust assumptions in a trade's life cycle

When a user opens a long position in Bitcoin perpetual futures on a centralized exchange, that user traverses a long and fragile chain of trust. Each step in this chain is an implicit trust assumption. This example is not a hypothetical thought experiment but a routine operation that millions of traders repeat every day. Every click of the trade button, every entry of an order quantity, means that the user is entrusting the safety of their funds and the fairness of their trades to a centralized, opaque system.

The first is trust in the asset-custody layer. The user deposits dollar stablecoins into the exchange, which marks the start of the entire trade life cycle. At that instant, the user must trust that the exchange will safeguard these funds. Such trust means believing

that the exchange will not misappropriate customer assets for high-risk investments, will not lose the private keys through internal mismanagement, and will not invoke various pretexts to refuse a withdrawal when the user requests one. This assumption is the foundation for all subsequent trading, yet it rests entirely on the exchange's reputation and the moral commitments of its management.

The second is trust in the matching and execution layer. When a user places a limit buy order, they must trust that the exchange's matching engine will execute the order honestly according to price priority and time priority. Although users can obtain some order-book data through the API, they cannot verify the completeness or authenticity of that data, and the exchange is fully capable of hiding internal orders or fabricating liquidity in the API output. Users must therefore trust that the exchange will not let its proprietary market-making desk exploit an informational advantage to jump the queue, will not deliberately delay their orders to capture front-running profits, and will not, under extreme market conditions, prevent them from closing positions by severing their connection.

The third is trust in the clearing and settlement layer. Once each trade is executed, account state is taken over by this layer, which performs two kinds of continuously recurring work: clearing (reckoning the books) and settlement (posting the transfers). Clearing translates a *trade event* into a *change in account state*: it recomputes long and short positions, realized and unrealized profit and loss (P&L), and margin utilization, and it nets the multiple trades of a single account. Settlement is the actual transfer of value, including continuous mark-to-market against the mark price, the periodic funding payments unique to perpetual futures, and the booking of P&L at position closure. In both of these processes, the user must trust the exchange to keep the books and move the funds honestly—neither tampering with the account records nor distorting the mark-to-market and the funding rate. The fairness and transparency of the mark price are the shared benchmark of this layer: it drives both the routine mark-to-market settlement and the determination of whether margin is sufficient. Users must therefore be especially confident that the exchange will not manipulate the mark price in order to trigger liquidations that should not occur, will not abuse the auto-deleveraging (ADL) mechanism when a bankruptcy event arises, and that its declared insurance fund genuinely exists and is properly managed. Liquidation and auto-deleveraging are merely this layer's default-handling tools when an account becomes insolvent, not the whole of its function (the strict distinction among clearing, settlement, and liquidation is developed systematically in Section 28.4.3).

The fourth is trust in data authenticity. The prices, depth, trading volumes, and other data that users see on the exchange's front end are their sole basis for making trading decisions. Users must trust that these data are genuine rather than fabricated. This means believing that trading volume has not been artificially inflated by wash trading, that the order book is not flooded with ghost liquidity, and that market depth reflects genuine buying and selling intent rather than numbers invented by the exchange to manufacture an illusion of prosperity.

The last is trust in governance fairness. The exchange holds the unilateral power to change the rules. If the exchange revises its fee formula, adjusts liquidation parameters,

or alters its market-maker agreements, users must trust that these changes serve the overall interest of the market rather than the exchange's private gain. At the same time, users must trust that such rule changes will be disclosed to all market participants promptly and publicly, rather than made in secret.

These five trust assumptions form the foundation on which a centralized exchange operates. The core engineering insight, however, is this: the violation of any one of the five can inflict irrecoverable losses on users. More critically, in most cases users have no way at all to verify whether these trusts are being honored. Before a collapse, no publicly available method can detect that assets have been misappropriated. From the standpoint of behavioral finance, the way users form trust in a centralized platform is often shaped by cognitive biases: the availability heuristic leads users to substitute a platform's past record of normal operation for a rational assessment of its architectural security, while the herding effect leads large numbers of users to abandon independent judgment in response to the social signal that *everyone else is using it*. These cognitive limitations imply that the value of a layered architecture lies not only in providing verifiability for rational users, but also in protecting—through the architecture itself—those users who are cognitively unable to assess trust risk correctly. This unverifiable trust is precisely the root of systemic risk.

28.1.2 The trust stack model

To analyze the trust assumptions in a trading system more systematically, we propose the *trust stack* method of analysis. This method decomposes the trust assumptions of a trading system into five layers, from asset custody at the very bottom to the governance mechanism at the top. This layered model applies not only to centralized exchanges but equally to decentralized exchanges and to various hybrid architectures.

At the very bottom of the trust stack is the asset-custody layer, whose core question is whether assets are safe; the consequences of its trust failure are the most severe—assets are directly misappropriated or lost—and in a centralized exchange it depends entirely on the exchange's internal management and moral restraint. The second layer up is the matching and execution layer, which concerns whether orders are matched honestly; its trust failures manifest as front-running, queue-jumping execution, and price manipulation. Because the matching engine runs on private servers, outsiders cannot verify the fairness of order matching and can only accept the final execution results the exchange provides. The third layer is the clearing and settlement layer, responsible for post-trade clearing and settlement and for using liquidation as default handling when an account is insolvent (the distinction among the three is detailed in Section 28.1.1). When its trust is broken, the consequences include both the covert distortion of the books and settlement and unfair liquidations and abuse of auto-deleveraging; yet the clearing-and-settlement conventions and liquidation parameters are often treated as trade secrets that users cannot review. The fourth layer is the data and oracle layer, responsible for the authenticity of prices and market data; when its trust fails, the market is flooded with fake volume and price manipulation, and the exchange, as the sole data provider, holds a complete informational monopoly. The top layer is the governance layer, which determines how rules are set and changed; its trust failures manifest as covertly chang-

ing the rules for profit. In a centralized exchange, this governance is usually implicit, with the power to interpret and amend the rules held entirely by management.

Table 28-1 summarizes the core question, the consequences of trust failure, and the verification difficulty at each of the five layers. The *verification difficulty* column reveals a consistent pattern: from the bottom layer to the top, verification difficulty is generally high, which means that under the current centralized architecture users lack effective external means of verification at nearly every layer.

Layer	Core question	Consequences of trust failure	Verification difficulty
Layer 5: governance layer	How are rules set and changed?	Covertly changing rules for profit	Very high (usually black-box decisions)
Layer 4: data layer	Are prices and data genuine?	Fake volume / price manipulation	High (requires complex data forensics)
Layer 3: clearing and settlement layer	Are clearing (reckoning) and settlement (posting) executed per the rules?	Distorted books / unfair liquidation / ADL abuse	High (internal algorithms undisclosed)
Layer 2: matching layer	Are orders matched honestly?	Front-running / queue-jumping / order delay	Very high (full order flow unobtainable)
Layer 1: custody layer	Are assets safe?	Assets misappropriated or lost	Very high (no on-chain verifiable reserves)

Table 28-1. The five-layer trust stack model with its consequences of trust failure and verification difficulty (Data source: compiled by the author)

The analytical value of the trust stack is that it lets us ask three key questions about any trading system at each layer. First, who is being trusted—a particular entity, a piece of code, or a validator set? Second, can this trust be verified—does the user have any means of checking whether the trust is being honored? Third, can this trust be replaced—is there a technical mechanism that can eliminate the dependence on this particular entity? Through this structured questioning, we no longer reduce the problem of trust to a binary opposition between centralized and decentralized; instead, we turn it into a precise problem of engineering design.

The core strength of the trust stack model is that it offers a unified template for comparative analysis across different trading systems. The trust stack of a traditional financial centralized exchange, for example, is entirely different from that of a generalized automated market maker (AMM). Likewise, the trust stack of a perpetual futures exchange on an app-chain differs from that of a centralized exchange. This layered perspective lets us move beyond a simple binary account and into the details of system architecture.

The layers of the trust stack fail in different ways. The asset-custody layer is the most fundamental, and its failure is irreparable because the assets cannot be recovered. A failure in the matching and execution layer may amount only to a poorly executed trade and causes no permanent loss of assets. A failure in the clearing and settlement layer may distort the books or lead to unfair liquidations, but as long as the clearing, settlement, and liquidation rules are transparent, users can at least appeal. A failure

in the data layer may lead to poor decisions, yet once the data are published on chain they remain externally verifiable. A failure in the governance layer is the most insidious: rule changes can be hidden temporarily, until a crisis brings them to light.

In addition, the layered design provides a framework for understanding why different exchanges choose different trust-minimization strategies: centralized exchanges and app-chain exchanges often make opposite trust trade-offs at the custody layer and the matching and execution layer (see the trust-stack audit in Section 28.5), and system designers choose the optimal architecture accordingly, adapting to their specific circumstances.

28.1.3 FTX's extreme failure mode

The true value of the trust stack model is that it is not merely a theoretical framework but an analytical tool that helps us understand the real risks that exist in practice. When we revisit an actual historical event, the model's usefulness becomes especially clear.

When FTX collapsed in November 2022, a gap of more than $8 billion opened between customer assets and corporate liabilities, and at the moment of collapse the firm could not honor redemptions on demand [1]. The subsequent bankruptcy investigation and court filings revealed a grave engineering fact: the platform had never technically segregated customer assets from corporate assets [7]. All assets were commingled in a single wallet system, and no code-level access control or multi-signature mechanism existed to prevent management from drawing on customer funds at will [8]. The only safeguard was management's verbal promise and the users' one-sided trust.

This is precisely the extreme instance of the monolithic black box architecture defined at the opening of this chapter: all functional modules are concentrated in an opaque cluster controlled by a single entity, and the account balance a user sees on the front end is merely a number in a database, backed by no cryptographic proof. Whether assets were transferred to an affiliated market maker, whether orders were matched honestly, and whether liquidation parameters were secretly altered to protect internal accounts—users could neither know nor, still less, verify.

The core defect of the monolithic black box architecture lies in its correlation trap. In a well-designed distributed system, the failures of different components are expected by design to be independent of one another. Yet even in a carefully designed distributed system, common-mode failure—the simultaneous failure of multiple components triggered by a shared software dependency, infrastructure provider, or operational process—remains a systemic risk that is difficult to eliminate entirely. In a monolithic architecture, this problem is pushed to its extreme: all functional modules share the same root of trust. This means that once that root becomes corrupted or fails, the system faces not a partial degradation but the simultaneous collapse of all trust layers.

Examining this collapse through the trust stack model, we can see clearly the catastrophic consequences of the monolithic black box. At the custody layer, customer funds were directly misappropriated to cover the trading losses of an affiliated firm. At the matching layer and the clearing and settlement layer, internal accounts were granted

special privileges that allowed them to bypass the ordinary risk controls and liquidation mechanisms. At the governance layer, the setting and enforcement of rules served the interests of the controlling party entirely, rather than fairness to market participants. Not only were all five trust layers concentrated in the same entity, but that entity used its complete control to violate its implicit promises to users at every layer.

It was precisely to address this single-point-of-failure risk that the traditional financial system gradually developed an architecture of functional separation (see Section 28.3). In modern traditional finance, the exchange handles only matching, the clearinghouse handles risk management and settlement, the custodian bank handles asset safekeeping, and the regulator oversees the rules. Each function is performed by an independent entity, which greatly reduces systemic risk. By recombining all of these functions into a single opaque entity, the cryptocurrency centralized exchange represents, in architectural terms, a severe regression. Academic research shows that this excessive concentration of trust forces centralized exchanges to hold capital reserves far larger than those of traditional financial institutions in order to withstand potential run risk—yet even so, it cannot fundamentally eliminate the possibility of insider malfeasance [9].

FTX's collapse was not an isolated event. In the years that followed, the industry found that other centralized exchanges in the cryptocurrency market suffered from similar architectural defects. Commingled assets, the absence of on-chain verifiable proof of reserves, and inadequate technical barriers against insider malfeasance are not the individual misfortunes of one platform but inherent defects of the monolithic black box architecture of centralized exchanges.

28.1.4 Trust concentration and systemic risk

To assess quantitatively the risk level of different trading-system architectures, we introduce trust concentration as an analytical metric. Trust concentration measures what fraction of the trust layers in a trading system are controlled by the same entity. Its basic definition is as follows: for the five layers of the trust stack, if k of them are controlled by the same entity, that entity's trust concentration is $k/5$. When one entity controls all five layers, trust concentration reaches 100%; when the five layers are controlled by independent entities, the maximum trust concentration of any single entity is only 20%. This definition uses equal weighting—that is, it assumes the five layers carry the same risk weight. In practical assessment, one can introduce differentiated weights according to the severity of the consequences of each layer's trust failure; for example, because a custody-layer failure causes irreversible asset loss, its weight should exceed that of a data-layer failure, which merely biases trading decisions. A weighted trust concentration reflects the actual distribution of a system's fragility more precisely. One illustrative weighting scheme is as follows: custody layer 0.35 (asset loss is irreversible), clearing layer 0.25 (directly affects solvency), matching layer 0.20 (affects the fairness of trading), data layer 0.12 (indirectly affects decision quality), and governance layer 0.08 (affects the direction of rule evolution). Under this weighting, if an entity controls all five layers, the weighted trust concentration is still 100%; but the partial de-trusting of different layers produces differentiated reductions in concentration.

In a typical centralized exchange, all five trust layers are controlled by the same operating entity. Its trust concentration therefore reaches 100%. Such extremely high trust concentration implies extremely high systemic risk. Should the controlling entity's internal controls fail, or should its management act in bad faith, the entire system fails rapidly. No external check exists that could sound a warning or intervene before the crisis strikes.

By contrast, the separated architecture of the traditional financial system substantially lowers trust concentration through institutional design. If the five trust layers are the responsibility of an independent exchange, a central counterparty clearinghouse, a custodian bank, a data provider, and a regulator, respectively, then each entity's trust concentration is only 20%. Although users must still trust these centralized institutions, the trust is dispersed. The failure of one entity does not necessarily lead to the collapse of the others, and the entities have an incentive to monitor one another.

An ideal decentralized architecture pursues a trust concentration close to 0%. In such an architecture, no layer depends on the moral commitment of any single entity; each is instead controlled by open-source code, a consensus algorithm, or a decentralized validator set. Assets are held in custody by smart contracts, the matching and clearing logic executes publicly on the blockchain and is verified by the whole network, data are supplied by a decentralized oracle network, and governance is carried out through on-chain voting by token holders.

Trust concentration is not only a security metric; it also bears directly on the fairness and resilience of the market. The higher the concentration, the greater the room for the controlling party to engage in rent-seeking by exploiting its informational privilege and rule-making power, and the worse the market's fairness. At the same time, high concentration means the system carries a severe single-point-of-failure risk and lacks resilience in the face of internal corruption or external attack. Lowering trust concentration—decomposing the monolithic black box into a layered, transparent architecture—is therefore an engineering necessity for building a resilient on-chain financial system.

Measuring trust concentration provides an objective framework for assessing the relative margin of safety of different trading systems. If an emerging trading platform, by placing its custody layer in an on-chain smart contract and delegating its clearing layer to an independent app-chain, achieves a trust concentration markedly lower than that of a competitor that concentrates all functions in a single entity, then that platform holds a structural advantage along the systemic-risk dimension. This improvement in the margin of safety does not translate directly into an improvement in user experience (which also depends on factors such as latency, liquidity depth, and interface design), but it lowers the probability that users suffer catastrophic losses under extreme market conditions.

Trust concentration also explains the relationship between regulatory pressure and architectural choice. An exchange with higher trust concentration, lacking any architectural technical constraint on insider malfeasance, must compensate for its deficient architectural trust through stricter external regulatory scrutiny and more frequent proof of reserves. This is, in essence, a substitution mechanism that offsets an architectural defect with institutional cost.

28.2 The trust-performance frontier

Given that the centralized architecture carries such a severe single-point-of-failure risk, why not simply migrate every stage of trading onto the blockchain? If smart contracts could execute predetermined logic flawlessly and cryptographic proofs could eliminate the dependence on a single entity, building a *fully on-chain*, trustless trading system would seem to be the only ultimate solution. Reality, however, is otherwise: in the engineering practice of decentralized finance, a fundamental physical constraint stands between trust minimization and performance maximization. This trade-off does not stem from a temporary limit of current engineering capability; it is dictated by the inherent constraints of distributed-systems theory. Understanding this frontier between trust and performance is a prerequisite for designing any robust on-chain financial architecture.

28.2.1 The performance cost of a fully on-chain architecture

Executing the matching, clearing, and settlement of every order on a blockchain network means that every trade's change of state must be confirmed by the consensus of the whole network's nodes or its validator set. Although such global verification achieves an extremely high degree of trust minimization, it exacts a significant performance cost, first evident in a sharp increase in system latency. The internal processing latency of a top centralized exchange's matching engine is typically in the microsecond range, but this figure is meaningful only for colocated users; for ordinary users connecting through a remote API, end-to-end latency typically ranges from 10 to 100 milliseconds, encompassing network transmission, authentication, and risk-control checks. Binance, for example, has continuously increased its matching engine's order-processing capacity from the 1.4 million orders per second published in its 2017 white paper [10], and the peak capacity of leading exchanges now reaches the order of millions per second. By contrast, even an app-chain designed specifically for high-performance trading tends to remain at the millisecond or sub-second level: through its customized HyperBFT consensus, Hyperliquid achieves a median end-to-end latency of about 200 milliseconds and a processing capacity of roughly 200,000 orders per second [11], still three orders of magnitude short of a centralized exchange's microsecond response. Since the Merge, the block time of the Ethereum mainnet has been fixed at about 12 seconds (proof of stake), with a throughput of only 15–30 transactions per second [12], wholly unable to meet the speed requirements of high-frequency trading.

Beyond the hard limits on latency and throughput, a fully on-chain architecture also entails significant economic costs. In a centralized exchange, matching against the internal order book is almost costless, and users pay only a very small percentage fee upon final execution. In a fully on-chain model, however, every placement, cancellation, or modification of an order consumes network resources and incurs a corresponding transaction fee. This mechanism not only greatly increases traders' friction costs but also severely constrains liquidity providers. In the discussion of the structural arbitrage residual in Chapter 18, the efficiency gap between centralized and decentralized exchanges is, in essence, the quantifiable manifestation of this *cost of trustlessness*.

More seriously, the complete transparency of a fully on-chain architecture, while it

solves the trust problem, also introduces a new market-making dilemma. In a fully transparent on-chain environment, a market maker's order flow and capital movements are openly visible to all network participants, which makes them highly vulnerable to attacks based on maximal extractable value (MEV) [13]. To hedge this additional risk created by on-chain transparency, market makers are forced to adopt a series of defensive behaviors: widening the bid-ask spread, reducing quoted depth, increasing the frequency of quote flickering, and, in extreme cases, withdrawing liquidity entirely. In the end, this cost of trustlessness is passed on to all ordinary traders through multiple channels. Attempting to force every stage of trading onto the blockchain therefore not only struggles technically to meet the performance requirements of a professional market but also carries economic frictions that are difficult to overcome.

28.2.2 The systemic risk of a fully off-chain architecture

Given that a fully on-chain architecture faces such formidable performance and cost challenges, should we retreat to a fully off-chain, centralized model? The answer is no. The current centralized exchange is the very archetype of the fully off-chain model, and its extremely high performance rests entirely on an extreme concentration of trust: while users enjoy microsecond latency and cost-free cancellation, they are forced to entrust the full power over asset safekeeping, order matching, clearing execution, and rule-making to a single opaque entity.

The core defect of this fully off-chain black box architecture is that users have no technical means of verifying what actually happens inside the box. Even if a centralized exchange voluntarily chooses to improve transparency—by publishing proof of reserves periodically or disclosing some wallet addresses, for instance—such transparency still rests on a voluntary basis. The exchange can change its disclosure rules at any time, or even stop publishing key data on the eve of a crisis. Such non-architectural guarantees often prove worthless under extreme market conditions. The FTX collapse analyzed in Section 28.1.3 amply demonstrates that when one entity holds both control over assets and the power to interpret the rules, no internal risk-control mechanism or management promise can withstand the systemic default induced by the temptation of enormous gain.

Another latent hazard of the fully off-chain architecture is the high correlation of its failures: as analyzed in Section 28.1.3, all functional modules of the monolithic black box share the same root of trust, and once that root fails—through a hacking attack, internal misappropriation, or a regulatory freeze—the collapse is total and instantaneous, with no buffer between flawless operation and a complete standstill. To retreat to the fully off-chain black box merely because a fully on-chain architecture is not performant enough is therefore to trade a short-term expedient for long-term risk.

28.2.3 Characterizing the trust-performance frontier

Through the comparative analysis of fully on-chain and fully off-chain architectures, we can see clearly a fundamental trade-off in system design: trust minimization and performance maximization cannot both be achieved to the fullest along the same dimension. This trade-off can be described as a *trust-performance frontier*. If we place

the degree of trust minimization (verification depth) on the horizontal axis and system performance (the reciprocal of latency, or throughput) on the vertical axis, then, under given technical conditions, all feasible architectural designs form a set whose Pareto frontier is the trust-performance frontier. Any architectural choice can find its footing only on or inside this frontier; it cannot surpass the frontier itself. In most engineering settings, this frontier curve is convex: when the number of validators is used as a proxy for verification depth, the communication overhead of Byzantine fault-tolerant (BFT) consensus tends to grow superlinearly as the validator set expands, so that the performance cost of deepening verification further from a high-trust region increases at the margin. Verification depth and the number of validators are not the same dimension: an app-chain with only about twenty validators still achieves greater trust minimization than a centralized exchange controlled by a single entity. Convexity is therefore a reasonable characterization of the frontier's shape rather than a strict theorem. This convexity is consistent with the layered strategy: provided that the trust requirements of the layers are heterogeneous and separable, rather than pushing the whole system to one extreme of the frontier, it is better to select a different position on the frontier for each layer, thereby obtaining a superior overall combination of trust and performance. This concept shares a common theoretical root with the widely discussed *scalability–decentralization–security* trilemma of the blockchain field; both arise from the computation and communication overhead inherent in global verification within distributed systems. The trust-performance frontier differs from the trilemma in that the former focuses on the two-dimensional trade-off between trust assumptions and execution efficiency in a trading-system architecture, whereas the latter is a more general constraint on blockchain design as a whole.

Locating three representative categories of trading-system architecture on this frontier makes their respective trade-offs along the trust and performance dimensions immediately visible. In Table 28-2, from the traditional centralized exchange to the Ethereum-mainnet DEX, the degree of trust minimization and the level of performance exhibit a nearly perfect negative correlation, whereas the app-chain architecture attempts to open a new feasible region between the two extremes.

Architecture type	Degree of trust minimization	Performance (latency and throughput)	Position on the frontier
Traditional centralized exchange	Very low (relies entirely on a single entity)	Very high (microsecond latency, throughput in the millions)	Upper-left end of the frontier
Ethereum-mainnet decentralized exchange	Very high (whole-network consensus verification)	Very low (latency on the order of ten seconds, low throughput)	Lower-right end of the frontier
High-performance app-chain (such as Hyperliquid)	Relatively high (customized consensus, decentralized verification)	Moderate (sub-second latency, throughput on the order of 200,000)	Middle of the frontier, toward the right

Table 28-2. The positions of three trading-system architectures on the trust-performance frontier

(Data source: compiled by the author)

On this frontier curve, the traditional centralized exchange sits at the upper-left extreme, trading away trust minimization entirely for the highest performance. The decentralized exchange built on a general-purpose layer-1 network such as Ethereum sits at the lower-right extreme, achieving the greatest degree of trustlessness but forced to endure lower performance. The app-chain architectures that have since emerged, such as Hyperliquid, attempt to find a more balanced equilibrium point on the frontier. Through customized consensus mechanisms and specially optimized network architecture, they raise performance to a level sufficient to support professional derivatives trading while maintaining a relatively high degree of trust minimization.

As the underlying technology advances—through more efficient consensus algorithms, more sophisticated zero-knowledge proofs, and the adoption of parallel execution environments—this trust-performance frontier is indeed shifting toward the upper right. This means that, in the future, systems will be able to obtain higher performance while maintaining the same degree of trust minimization. Yet however the technology evolves, the additional computation and communication overhead of trustless verification is an objective reality, and under any known consensus model that overhead will not be eliminated entirely. Lower trust assumptions generally come with some form of performance compromise.

28.2.4 The logic of layered architecture

The trade-off between trust and performance is unavoidable, and neither the fully on-chain nor the fully off-chain approach is an ideal final solution. The way to resolve this tension lies in breaking the fixed assumption that *the system must, as a single whole, choose one position on the frontier*. A complete trading system is not an indivisible monolith but a trust stack composed of multiple functional layers. The layers differ in the severity of the consequences of trust failure and in their sensitivity to performance. The optimal architectural design should therefore evaluate the requirements of each layer independently and choose the best position on the trust-performance frontier for each. This is the underlying logic of layered architecture.

Three representative layers make this differentiated trade-off clear. A trust failure in the asset-custody layer is irreversible, while deposits and withdrawals are low-frequency operations that are insensitive to latency; this layer should therefore pursue trust minimization to the fullest (custody optimizes for safety). The matching and execution layer directly determines the fluidity of trading and the quality of liquidity, with stringent requirements for latency and throughput; as long as matching is verifiable after the fact, it can tolerate a degree of trust assumption in exchange for performance (matching optimizes for performance). The clearing and settlement layer sits between the two: computationally intensive in normal times yet insensitive to the latency of any single transaction, it must, during violent volatility, balance sufficient performance with an extremely high degree of trust minimization to guarantee the fairness of the rules (clearing sits in between). The optimal positions of these three layers, together with the specialized requirements of the data and oracle layer and the governance layer, are established layer by layer as design principles in Section 28.4.

The core idea of layered architecture is to reject the *one-size-fits-all* compromise. By decomposing the vast monolithic black box into independent modules and implementing a customized trust-performance trade-off within each, layered architecture can preserve the performance advantages of a centralized exchange at the critical stages while achieving the security guarantees of a decentralized network at the critical risk points. This engineering method of recombining the layers is the core design framework for building the next generation of robust on-chain financial systems.

28.3 The layered trust architecture of traditional finance

The trust problems facing the perpetual futures market—counterparty risk, asset safety, and the fairness of the rules—are not new problems of the cryptocurrency era. Over centuries of crisis and reform, traditional finance developed a mature trust infrastructure that includes central counterparty clearinghouses, central securities depositories, and regulators. These institutions are, in essence, the traditional-finance version of the trust stack, each corresponding to one or more layers of the stack. Only by understanding the operating logic, strengths, and limitations of these known solutions can we judge in which respects an on-chain architecture is an improvement, in which a regression, and in which merely a different trade-off.

28.3.1 The historical evolution of functional separation

Traditional finance did not begin with a well-developed trust infrastructure. The nineteenth-century exchange was likewise a monolithic black box, in which trade matching, clearing and settlement, and asset safekeeping all took place within the same exchange or broker. Under this architecture, problems such as counterparty default, brokers' misappropriation of customer funds, and price manipulation occurred frequently—strikingly similar to the problems that today's cryptocurrency centralized exchanges face.

The financial history of the past 200 years is a history of crisis-driven reform. Every major crisis pushed some layer of the trust stack toward independence and institutionalization. The Great Crash of 1929 directly gave rise to the U.S. Securities and Exchange Commission (SEC) in 1934, making the governance layer independent. The Wall Street paperwork crisis of 1973—the breakdown of the settlement system for paper stock certificates—prompted the creation of the Depository Trust Company (DTC, later the depository subsidiary of the Depository Trust & Clearing Corporation, DTCC), marking the independence of the custody layer [14]. Black Monday in 1987 drove the strengthening of clearing mechanisms, and the G20 Pittsburgh consensus after the 2008 global financial crisis directly gave rise to the reform of mandatory central counterparty clearing for over-the-counter (OTC) derivatives—embodied in the Dodd-Frank Act and the European Market Infrastructure Regulation (EMIR)—establishing the independent status of the clearing layer [15].

Traditional finance took 200 years to decompose the trading system from a monolithic black box into five independent layers, each managed by a different institution. At the governance layer are independent regulators such as the Commodity Futures Trading

Commission (CFTC) and the SEC; at the data layer, regulated data providers such as Bloomberg and Reuters; at the clearing layer, independent central counterparties such as CME Clearing and LCH; at the matching layer, regulated matching institutions such as the New York Stock Exchange and Nasdaq; and at the custody layer, independent central depositories such as the DTCC and custodian banks. Each layer is operated by a different specialized entity, with institutional checks and balances among them. In the span of a mere decade, cryptocurrency centralized exchanges recombined these five layers into a single entity—an architectural regression whose extreme consequence was precisely the collapse of FTX.

28.3.2 The CCP and clearing-layer trust

The central counterparty is the core mechanism by which traditional finance addresses clearing-layer trust. Its most critical function is novation. Once a buyer and a seller reach a trade, the CCP steps in, and the original trade is replaced by a trade between the buyer and the CCP and another between the CCP and the seller. The CCP becomes the central counterparty to all trades, so the buyer need not trust the seller—only the CCP [16]. In the language of the trust stack, the CCP concentrates clearing-layer trust from *trust your counterparty* into *trust one strictly regulated, specialized clearinghouse*. This differs fundamentally from the crypto market. In crypto perpetual futures there is no central counterparty, and a trader's counterparty is either a market maker or another trader. Counterparty risk is fully exposed, or borne implicitly by the exchange—yet the exchange may be unable to bear it, as the FTX episode plainly showed. Some leading crypto exchanges have established insurance funds and auto-deleveraging mechanisms that functionally assume part of the CCP's risk-mitigation role. The key difference between these mechanisms and an institutionalized CCP, however, is this: the insurance fund is limited in size and lacks any cross-platform mutualization layer, there is no mutual-aid mechanism analogous to the *loss-sharing by non-defaulting clearing members* in a default waterfall, and there is certainly no central bank acting as lender of last resort to backstop liquidity. They are therefore only an incomplete replica of the CCP's functions, not a genuine institutional substitute.

The CCP absorbs default losses layer by layer through the default waterfall (Chapter 1 has already discussed the basic mechanism of the central counterparty and the default waterfall; the point is restated here briefly in the context of clearing-layer trust): in sequence, the defaulter's initial and variation margin, the defaulter's contribution to the default fund, the CCP's own capital (its *skin in the game*), the mutualized sharing of losses from the default funds of non-defaulting clearing members, and finally the recovery and resolution procedures [17].

Compared with the crypto market, the CCP holds significant advantages. First, trust is concentrated but strictly regulated: the user need only trust one specialized institution supervised by a central bank or the CFTC. Second, the default waterfall has depth—not one major CCP defaulted even during the 2008 financial crisis, which proved the mechanism's effectiveness [18]. Yet the CCPs' effective performance in 2008 rested on an important implicit premise: central banks around the world provided a backstop for the entire financial system through emergency liquidity support, and this sovereign-

credit support is an institutional advantage that a decentralized system cannot replicate. Third, because the CCP's own capital ranks ahead of the clearing members in the default waterfall, the CCP has a strong incentive to maintain the quality of its risk controls.

The CCP also has limitations, however, and these limitations are precisely the directions in which an on-chain system might improve. The CCP is itself a *too big to fail* single point: if it runs into trouble, the entire market seizes up—which has been identified as a new form of systemic risk [19]. The London Metal Exchange (LME) nickel contract crisis of March 2022 laid this limitation bare. When the nickel price rose about 250% over three trading days (climbing to roughly 3.5 times its level), the LME (the decision was made at the exchange level, not by its clearinghouse, LME Clear) chose to suspend trading and retroactively cancel executed trades. This discretionary governance decision provoked serious doubts among market participants about the fairness of the exchange and its clearing infrastructure, and it led to multiple lawsuits [20]. The episode shows that a CCP's discretionary power under extreme market conditions can conflict with its positioning as neutral infrastructure. In addition, the CCP has a high barrier to entry: only large financial institutions can become clearing members, and small participants must go through intermediaries, making it a professional club rather than a permissionless system. The CCP's operating hours are also limited, and it does not support around-the-clock trading. Its transparency, moreover, is insufficient: its risk models and stress tests are confidential, so clearing members and end users cannot fully verify the quality of its risk controls—a form of regulated opacity.

28.3.3 The DTCC and custody-layer trust

The DTCC is the core institution that addresses custody-layer trust. Its core functions include central securities depository services and net settlement. All securities traded in the U.S. market are ultimately held in custody by the DTCC and its subsidiaries; buying and selling are merely transfers of title on its books, and physical certificates no longer move. Trust is concentrated from *every broker safely holds my shares* into *the DTCC safely holds everyone's shares* [21]. Through net settlement, of the trillions of dollars in daily trades, the funds or securities that actually need to be transferred amount to only a small fraction of total trade value, which greatly reduces settlement risk and funding requirements. According to reported figures, in 2024 the DTCC processed securities transactions worth about $3.79 quadrillion and provided custody and asset services for roughly $99 trillion in assets [22].

Within the trust stack, the DTCC solves the core problem of asset custody: assets are held in a strictly regulated, specialized institution independent of the exchange and the broker, so that even if a broker fails, customers' securities remain safe. Compared with the crypto market, the DTCC guarantees, through law and institutional design, the separation of customer assets from the broker's own assets, whereas most cryptocurrency centralized exchanges commingle customer assets with their own. As for settlement delay, the DTCC currently operates on T+1 settlement, whereas an on-chain system can achieve near-instant settlement. Yet, while instant settlement eliminates settlement risk, it also sacrifices the capital efficiency of net settlement: traditional T+1 settle-

ment offsets large volumes of opposing trades within the settlement window, so that the funds actually transferred amount to only a small fraction of total trade value. As for transparency, the DTCC is subject to regulatory audit but is not fully transparent to external participants, whereas in an on-chain system all balances are publicly auditable.

The DTCC's success demonstrates the institutional value of separating custody from the trading and market-making functions. Its depository subsidiary, the DTC (in operation since 1973 and folded into the DTCC as its holding company in 1999), has never lost customer assets in its entire operating history [23]. This approach, however, is a form of centralized institutional trust that depends on a combination of law, regulation, and specialized institutions. The crypto market can achieve the same goal through decentralized architectural trust. In principle, the asset segregation of a smart contract is stronger than the DTCC's, because the code executes deterministically according to preset logic, excluding the possibility of selective default by management. Yet the code itself carries the risk of bugs, and once a flaw appears its consequences are equally deterministic, since an attacker can exploit the vulnerability systematically. The DTC custody system, with more than 50 years of operating record and well-developed legal recourse mechanisms, offers a proven alternative safeguard [23].

28.3.4 The CFTC/SEC and governance-layer trust

Regulators play an irreplaceable role in the trust stack. The CFTC regulates the U.S. futures and derivatives markets, setting margin standards, position limits, and market-maker obligations, prosecuting manipulation and fraud, and overseeing the operation of central counterparties. The SEC regulates the securities markets, performing similar functions [24]. Within the trust stack, these agencies address trust at the governance layer and, in part, at other layers, ensuring that the rules are fair, that participants have channels of appeal, and that rule changes follow due process.

Regulation as a trust mechanism has notable advantages. The first is enforcement power: regulators have the legal authority to penalize violators or shut down offending institutions—a hard power that on-chain governance currently lacks. The second is professional judgment: regulators command experts in finance, law, and technology and can make complex discretionary decisions, and in extreme situations human judgment may be superior to the rigid execution of code. Finally, being regulated is itself a trust signal that lowers participants' cost of gathering information.

But regulation also has limitations. Regulatory processes are often measured in months or years, whereas market changes are measured in days or hours, so regulation always lags innovation. A regulator can govern only the markets within its jurisdiction, whereas the crypto market is global. There is also the risk of regulatory capture, in which the regulated can influence the regulator and skew the rules toward industry interests. Regulators can usually detect problems only after the fact and can hardly guarantee in advance that all is well; FTX, for instance, was never flagged by the CFTC before its collapse, because FTX exploited the regulatory fragmentation among multiple jurisdictions and, through a complex corporate structure (incorporated in the Bahamas, operating in Hong Kong, serving users worldwide), effectively evaded the

effective jurisdiction of any single regulator.

Compared with on-chain governance, regulators' rule-making is expert-led and legally binding, whereas on-chain governance relies on token-holder voting and can be manipulated by large holders. Regulators' human enforcement is subject to delay but carries coercive force, whereas on-chain code executes automatically without delay but is rigid. Regulators' rule-change procedures are transparent but slow, whereas on-chain governance may be fast or slow. Regulators offer legal remedies such as courts and arbitration, whereas the verdict of code in an on-chain system is usually final.

28.3.5 An overall assessment of the traditional trust architecture

The traditional trust architecture—the combination of the CCP, the DTCC, and regulators—has successfully solved many core problems. The CCP resolves counterparty default excellently, and the DTCC resolves asset misappropriation excellently; regulators' containment of price manipulation is moderate, and their guarantee of rule fairness is relatively good (the specific mechanisms of each institution are given in Sections 28.3.2–28.3.4).

The traditional trust architecture also has unsolved problems, however, and these are precisely the opportunities for on-chain systems. The traditional architecture serves only regulated local markets and excludes the world's large unbanked population, whereas an on-chain system is permissionless and open to anyone worldwide. Traditional institutions operate on business days and during trading hours, whereas an on-chain system natively supports around-the-clock operation. The traditional architecture's T+1 settlement still carries a full day of settlement risk, whereas on-chain settlement completes within block confirmation. The traditional architecture's risk models and position data are not fully transparent, whereas in an on-chain system all rules, trades, and positions are visible to everyone. In the traditional architecture, new products require a lengthy regulatory approval process, whereas an on-chain system is permissionless and innovation can be deployed instantly. Moreover, the core institutions of the traditional architecture themselves become systemic single points that require a government backstop, whereas the decentralization of an on-chain system in theory eliminates this single-entity risk.

From this, we can distill three paradigms for realizing trust. The first is institutional trust, exemplified by traditional finance, in which the source of trust is law, regulation, and specialized institutions, and the core mechanisms include the default waterfall, independent custody, and enforcement. Its advantages are 200 years of validation, coercive force, and professional judgment; its disadvantages are a high barrier to entry, geographic limits, slowness, and opacity. The second is platform trust, exemplified by the cryptocurrency centralized exchange, in which the source of trust is trust in the exchange or platform, and the core mechanisms are internal systems, brand reputation, and voluntary audits. Its advantages are the highest performance and the best user experience; its disadvantage is complete dependence on a single entity, with an extreme risk of collapse. The third is architectural trust, exemplified by on-chain systems, in which the source of trust is code, cryptography, and consensus mechanisms, and the core mechanisms are smart contracts, on-chain clearing, and on-chain governance. Its

advantages are transparency, global reach, around-the-clock operation, and permissionlessness; its disadvantages are limited performance, rigid rules, and the risk of code vulnerabilities.

The optimal solution may not be to choose one of the three, but to blend all three. At the asset-custody layer, architectural trust is the strongest—the equivalent of a code version of the traditional depository. At the clearing and settlement layer, architectural trust and institutional trust can be combined, taking the depth of the CCP's waterfall together with the transparent verifiability of the chain. At the governance layer, institutional, architectural, and platform trust can be blended. The design principle of layered architecture is precisely to find, at each layer, the optimal blend of these three trust paradigms.

28.3.6 Functional inheritance from the CCP to on-chain clearing

The narrative that on-chain finance is meant to replace traditional finance is mistaken; it pits the two against each other and ignores the institutional wisdom that traditional finance has accumulated over the long run. A better narrative is that on-chain finance is a technical upgrade of traditional finance's trust infrastructure. It uses code to achieve what the central counterparty already achieves through law, on-chain transparency to achieve what the depository already achieves through audit, and DAOs or multisignature schemes to achieve what regulators already achieve through administrative procedure.

This functional inheritance manifests in several respects. The CCP's novation is realized on chain by a smart contract acting as the impersonal counterparty to all trades, which eliminates the CCP's own default risk but loses the human judgment for handling extreme situations. The default waterfall is realized on chain as margin, the insurance fund, liquidity-provider takeover of liquidated positions, and auto-deleveraging, each step publicly auditable and automatically executed, eliminating the delay of human intervention—but its depth may fall short of a traditional CCP's, and there is no government backstop. The on-chain waterfall also typically lacks a layer corresponding to *skin in the game* (in which own capital bears losses ahead of members), and auto-deleveraging trims the positions of the profitable side, which is not equivalent, in either the population affected or the incentives, to the mutualized loss-sharing by nondefaulting clearing members in a traditional waterfall. Daily marking is upgraded on chain to real-time mark-price computation and continuous margin monitoring—from daily to per-second—reducing risk exposure, though the smoothing of the mark price may lag in a fast-moving market. Margin calls become instant liquidation on chain. This eliminates settlement risk but also denies the trader a second chance to top up margin, making the system more rigid. In a high-volatility environment, however, instant liquidation avoids the danger of risk exposure widening during the call period and may in fact be more prudent. This is all the more true under a multichain architecture, where cross-chain fund transfers often take several minutes or more—limited by source-chain confirmation and bridging delay—so that adding margin in practice becomes extremely difficult, which further reinforces instant liquidation as the appropriate default. Confidential stress tests become real-time risk monitoring based on public data,

which anyone can run independently, though there is as yet no standardized framework.

The design goal of an on-chain clearing and settlement system is not to replace the central counterparty but to realize its core functions through technical means while overcoming its limitations. This raises the key design questions: can an on-chain system achieve a default waterfall of equal depth? How can the optimal balance between risk control and user experience be found? How can the problem of incentive alignment be solved in an on-chain system? These questions constitute the core challenges of building a resilient on-chain financial system.

28.4 Design principles of layered architecture

The essence of layered architecture is not simple modular partitioning but a precise engineering strategy of trade-offs: it allows a system to select the optimal architectural position independently at each functional layer, according to that layer's specific requirements for trust minimization and performance maximization [25], thereby abandoning the crude mode of the monolithic black box entirely and building instead a transparent, verifiable, and highly resilient financial ecosystem.

Layered architecture comprises six core design principles: the first five correspond to the five layers of the trust stack, addressing each layer's distinctive trust challenge and establishing the corresponding standard for technical implementation; the sixth concerns the design of the interfaces between layers, ensuring secure coordination among the modules.

28.4.1 De-trusting the asset-custody layer

The asset-custody layer sits at the very bottom of the trust stack and is the foundation of the entire system's security. In the monolithic black box architecture, users' assets are concentrated in a wallet controlled by a single entity, a design that creates the most severe single-point-of-failure risk. Should the controlling entity succumb to moral hazard or suffer a security breach, all users' assets face irreversible loss. The design principle of the asset-custody layer is therefore clear: it must achieve maximal de-trusting and, even at the cost of sacrificing some performance, must guarantee the safety of assets.

In engineering terms, this means stripping away, to the greatest extent possible, a centralized entity's unilateral control over user assets. Modern on-chain financial systems typically use smart contracts as the vehicle for asset custody, employing immutable code logic to strictly constrain the flow of funds and the conditions for withdrawal. Under this architecture, the transfer of assets no longer depends on management's promise but is bound by cryptographic proof and predefined consensus rules. To strengthen security further, multi-party computation (MPC) and hardware security modules (HSMs) are widely applied in key management, ensuring that even if some nodes are compromised, an attacker cannot reassemble the complete signing authority [26].

In addition, the asset-custody layer must be able to operate independently of the other layers. Even if the matching engine crashes or the clearing and settlement system fails, users' assets must be able to remain safely in the custody smart contract, or users must

be allowed to withdraw their funds back to the underlying blockchain network through an emergency exit mechanism. In practice, however, the emergency exit mechanism faces operational risk: when an app-chain stops producing blocks entirely, the exit mechanism must rely on preset timeout-triggered logic, and in a mass-exit scenario the gas fees on the underlying network may spike sharply, making the exit cost uneconomical for small users. This design logic is of a piece with the traditional-finance principle that the custodian bank is independent of the exchange, but in the on-chain environment it is realized through mathematics and code rather than law and regulation. Smart-contract custody, however, carries an inherent legal vulnerability. In August 2022, the U.S. Treasury's Office of Foreign Assets Control (OFAC) added the immutable smart-contract addresses of Tornado Cash to its sanctions list, showing that a smart contract, too, can become the object of sovereign sanction; yet in November 2024, in Van Loon v. Department of the Treasury, the U.S. Court of Appeals for the Fifth Circuit ruled that an immutable smart contract does not constitute sanctionable "property" within the meaning of the International Emergency Economic Powers Act (IEEPA), and the Treasury accordingly delisted it in March 2025 [27]. This reversal shows that the legal boundaries of smart-contract sanctions are still evolving, but while the sanctions were in force, the legal risk faced by users interacting with the sanctioned contract was real. Moreover, the conflict between a judicial freeze order and the immutability of a smart contract has yet to be adequately resolved within the legal framework. Thus, while cryptographic custody effectively eliminates the operator's moral hazard, it does not eliminate legal risk or sovereign risk, and these two kinds of risk lie beyond the capacity of the technical architecture. At this layer, performance metrics such as transaction latency and throughput recede to secondary importance, because depositing and withdrawing assets are low-frequency operations, whereas security is the non-negotiable bottom line.

28.4.2 The verifiability of the matching and execution layer

Unlike the asset-custody layer, whose core demand is security, the matching and execution layer is the most performance-demanding stage of the entire trading system. Microsecond latency and throughput of tens of thousands of orders per second are necessary conditions for ensuring market liquidity and the trading experience. Subjecting the matching of every order to global consensus verification on the underlying blockchain would produce a performance loss that no active market could bear. The design principle of the matching and execution layer is therefore this: while meeting the demand for high performance, constrain the behavior of the matcher through after-the-fact verifiability mechanisms so as to ensure that orders are executed honestly.

In a layered architecture, the matching engine is usually designed as a high-performance independent module that can run on a customized app-chain or exist as an off-chain compute node. To prevent the matcher from exploiting its informational privilege to front-run or manipulate prices, the system must introduce strict fairness-guarantee mechanisms. The order-matching engine must strictly follow the price-time priority principle, and its execution results must be endorsable by cryptographic proof [28].

One engineering direction under active research and early exploration is the use of

batch proofs or zero-knowledge proof techniques. The matching engine processes vast numbers of orders at high speed off chain and then periodically submits state updates and execution proofs to the chain for verification. This approach preserves the efficiency of centralized matching while, through mathematical proof, stripping the matcher of the ability to act maliciously. Users need not verify the correctness of every trade in real time as it occurs, yet they can be confident that any matching that violates the established rules cannot pass the final on-chain verification. In this way, the matching and execution layer finds an effective balance point on the frontier between trust and performance: it tolerates local centralization in the execution process but, through global verifiability, eliminates one-sided trust in the matcher's honesty.

28.4.3 The clearing and settlement layer: from reckoning and posting to default handling

The clearing and settlement layer connects the results of matching and execution with the state of asset custody, and it is the layer of an on-chain system that corresponds most closely to the function of the traditional central counterparty. Yet it is precisely at this layer that the ambiguity of the Chinese terminology most easily obscures a key engineering distinction. In the context of financial infrastructure, clearing, settlement, and liquidation are three concepts that must not be used interchangeably. Clearing is *reckoning the books*—the state transition after a trade is executed but before funds actually move: replacing the counterparty with the system through novation, performing multilateral netting across multiple trades, recomputing positions and realized P&L, and updating margin utilization. Settlement is *posting*—the actual change of hands of value: mark-to-market against the mark price, the periodic funding transfers unique to perpetual futures, the formal booking of realized P&L at position closure and the release of margin, and, layered on top in the on-chain setting, finality confirmation. Liquidation is another matter entirely: it is the forced closure of a position triggered when account equity falls below the maintenance margin, the layer's handling mechanism in the event of default, not the whole of the layer's function. Many design discussions misjudge risk precisely because they blur all three together under the single word *clearing*. Accordingly, this section decomposes the clearing and settlement layer into three subfunctions—clearing, settlement, and liquidation—and establishes the design principle of each.

The overall design principle of this layer is that all rules must be executed consistently and transparently, so that no one—including the system's developers—can covertly alter parameters or interfere with the process. In traditional finance, this function is assumed by the central counterparty, which raises capital efficiency and isolates systemic risk through multilateral net settlement [29]. The goal of the on-chain architecture is not to abolish this function but to replace the dependence on a trusted clearinghouse with deterministic code and consensus verifiability.

Clearing: the deterministic recomputation of account state. The clearing subfunction answers the question of *what a given trade means at the account level*. In a traditional central counterparty, novation means the buyer and seller no longer expose each other to credit risk and instead face only the clearinghouse as their common

counterparty; in an on-chain system, this role is assumed by a smart contract or the state machine of an app-chain, so that after a trade both sides' positions are recorded according to uniform rules, and a default by either side does not propagate directly to its counterparty but is absorbed by the system. Multilateral netting consolidates a single account's multiple trades into a net change in position, and positions, realized P&L, and margin utilization are recomputed accordingly. The core requirement of the clearing sub-function is determinism: every margin determination and every update to the books must be independently recomputable by each validator, yielding the same result. This is precisely the on-chain embodiment of the principle that *the clearing authority must be a single point, while verification must be open to everyone*—the business-layer discretion over clearing is concentrated in a single deterministic engine, while the consensus layer guarantees that every step of that engine's output can be recomputed and checked by any node.

Settlement: the actual delivery of value. Unlike a traditional contract, which settles once at expiry, the settlement of perpetual futures occurs continuously. Mark-to-market settlement continuously revalues positions against the mark price, reflecting unrealized P&L in account equity in real time; funding settlement transfers funds between the long and short sides at fixed intervals, which is the core mechanism that lets perpetual futures anchor to the spot despite having no expiry; and closing settlement formally books realized P&L and releases margin when a position is closed out. In the on-chain environment, settlement adds a further layer of finality: a change of state becomes irreversible only after it is written into a block and confirmed by consensus. The design principle of the settlement sub-function is the transparency and immutability of posting—every funding transfer and every booking of P&L should be auditable on chain, foreclosing the covert transfers common in centralized exchanges.

Liquidation: the line of defense for handling default. When account equity falls below the maintenance margin, the liquidation mechanism steps in and forcibly closes the position to prevent it from going bankrupt. The complexity of liquidation requires this layer to possess a moderate degree of computational performance while maintaining an extremely high standard of trust minimization: if the liquidation rules could be changed at will, or if the liquidation engine could selectively ignore certain high-risk positions, the entire system's solvency would fail rapidly. The triggering conditions and execution flow of liquidation must therefore be hard-coded as public smart-contract or app-chain consensus logic. In implementation, the system needs a distributed liquidation orchestrator that continuously monitors the risk exposure of all accounts and, when the margin ratio falls below the maintenance threshold, triggers liquidations in an orderly fashion according to the urgency of the exposure. Modern liquidation engines generally adopt a partial-liquidation strategy—closing only the smallest fraction of a position needed to restore the account above the maintenance margin, rather than the traditional full liquidation—thereby reducing market impact while lowering the risk of a systemic reaction set off by a liquidation cascade.

To prevent an attacker from manufacturing spurious liquidations by manipulating oracle prices, the liquidation sub-function must establish a robust interface with the data layer and build a manipulation-resistant mark-price system. A mature mark-price construction typically includes the index price (the median of prices across multiple

exchanges), a basis moving-average term (which uses an exponential moving average to characterize the deviation between the contract mid-price and the index price and adds it to the index price additively), price-band protection (which limits the jump magnitude of the mark price when the deviation exceeds a threshold), and an outlier-exclusion mechanism (which automatically excludes any single data source exhibiting an extreme deviation). When liquidation is still insufficient to cover the losses, the insurance fund serves as the next line of defense to absorb bankruptcy losses; its operating state must likewise be fully on chain, and the injection and depletion of its funds should be bound by strict smart-contract constraints, ensuring that under extreme market conditions the system absorbs tail losses in a transparent and predictable manner rather than relying on an opaque internal bailout.

The relationship among the three sub-functions constitutes the internal logic of this layer: clearing comes first, keeping the books consistent before any further matching occurs; settlement then effects the actual transfer of value and confers on-chain finality; and liquidation is the last line of defense, maintaining the system's solvency when a default occurs. This distinction also runs through the chapters that follow—the next chapter's design of a decentralized clearing and settlement architecture rests precisely on treating clearing, settlement, and liquidation as each performing its own role, rather than conflating them.

28.4.4 Manipulation-resistant design of the data and oracle layer

The data and oracle layer is the sole bridge through which an on-chain system interacts with the external real world. Because a blockchain is itself a closed, deterministic state machine, it cannot directly perceive the prices of external assets or the occurrence of real-world events. It must therefore rely on an oracle network to bring this critical data into the system securely. The design principle of the data layer is to eliminate the fragility of a single data source and to ensure that the price information fed into the system is genuine, manipulation-resistant, and highly available.

The core of the oracle problem is how to introduce subjective or external information into a trustless system without undermining the system's overall security [30]. If a derivatives protocol with billions of dollars in locked value depends on a single exchange's API to obtain its mark price, that API becomes the critical point of fragility for the entire system. An attacker need only use a flash loan to create a brief price spike in a single market to trigger large-scale erroneous liquidations and reap excess profit.

To honor this design principle, the layered architecture must adopt a distributed oracle network. Such a network aggregates quotes from multiple independent data providers and multiple exchanges and removes outliers, thereby producing a robust composite price. In addition, oracle nodes must stake assets, using an economic incentive mechanism to ensure that they report data honestly. If a node submits erroneous data that deviates from consensus, its staked assets are slashed. Through this cryptoeconomic design, the data layer converts trust in *authenticity* into a reliance on economic rationality and decentralized consensus, thereby providing a solid data foundation for the clearing and execution above it.

28.4.5 The procedural legitimacy of the governance layer

Any complex financial system must evolve continuously as market conditions change, which inevitably involves parameter adjustments, contract upgrades, or the deployment of new features. The governance layer is responsible for managing these processes of change. In the monolithic black box, governance is implicit, and management can change the rules at will without notifying users. The design principle of the governance layer is therefore this: all rule changes must be public and thoroughly debated, and their execution must be verifiable and predictable.

Decentralized governance mechanisms aim to eliminate the unilateral control of a few insiders over the system. In an ideal layered architecture, governance power is distributed to token holders or to a council of representatives elected by the community. Any major change to the system—adjusting the base fee rate, revising the liquidation penalty ratio, or upgrading the core smart contracts—must be initiated through an on-chain proposal and put to a vote over a prescribed period [31]. On-chain governance participants may face legal-liability risk: the CFTC's enforcement precedent against Ooki DAO shows that DAO token holders who participate in governance voting may be deemed members of that unincorporated association and thereby bear joint legal liability for the association's debts [32].

More important still, the governance layer must introduce a time-lock mechanism. Even after a change proposal passes, it cannot take effect immediately but must wait a fixed cooling-off period in a time-lock contract. This cooling-off period provides an escape window for users who disagree with the change, allowing them to withdraw their assets safely before the rules change. Through this design, the system not only constrains governance power but also grants users the ultimate right to vote with their feet. The governance layer is no longer a center of power that issues commands unilaterally, but an engine of rule evolution based on consensus and transparent procedure.

28.4.6 Modular design of the interfaces between layers

The success of a layered architecture depends not only on the robust design within each layer but also on how the five layers coordinate securely and efficiently. The last design principle therefore concerns the interfaces between layers: the system must maintain a high degree of modularity and composability, so that the layers can evolve independently while the risk of cross-layer attacks is guarded against.

Modularity means that each layer should provide services externally through a standardized interface while hiding its internal implementation details. The clearing and settlement layer, for example, need not know which specific matching algorithm the matching and execution layer uses; it need only receive trade records in a particular format. Likewise, the matching engine need not care whether the asset-custody layer uses multi-party computation or a hardware security module; it depends only on the proof of available balance that the custody layer provides. This decoupled design greatly reduces the system's complexity, letting developers focus on optimizing the performance or security of a specific layer without having to restructure the whole system.

Modularity also introduces a new challenge, however: interface risk. An attacker often will not attack a hardened layer directly but will instead seek vulnerabilities in the passing of information between layers. For example, if the data layer's latency in passing price information to the clearing layer is too high, the liquidation logic may make erroneous judgments based on stale data. Interface design must therefore include strict data validation, timeout handling, and graceful-degradation mechanisms. When a layer fails or is attacked, the system should be able to isolate it through the interface protocol, preventing the risk from spreading to the other layers and thereby safeguarding the ultimate resilience of the entire financial ecosystem.

28.5 A trust-stack audit of current architectures

Applying the trust stack model to the major trading architectures in the current market reveals the trust state of these systems across the five layers of asset custody, matching and execution, clearing and settlement, data and oracles, and governance. A system's trust audit can not only quantify precisely the trust concentration of different architectures but also lay out clearly the engineering path along which the industry is evolving from the monolithic black box toward a layered, transparent architecture. The following examines, in turn, the trust assumptions and verifiability of the typical centralized exchange, the AMM-type decentralized exchange, the app-chain decentralized exchange, and the emerging hybrid architecture at each layer.

28.5.1 The trust stack of a typical CEX

The architecture of a typical centralized exchange is, in essence, a highly integrated monolithic system, and its trust stack exhibits extreme concentration. Under this architecture, over the entire life cycle of completing a trade, the user must deliver complete trust at all five trust layers to a single entity: the exchange operator. The fragility of this trust structure was amply exposed in the several industry crises of 2022, which demonstrated that when trust concentration reaches 100%, the system's single-point-of-failure risk becomes unbearable [33].

At the asset-custody layer, the user must trust the exchange completely to safeguard the deposited funds. Centralized exchanges typically manage customer assets with internally controlled hot and cold wallet systems, which means the operator has the technical ability to move funds at will. Although some exchanges have introduced proof-of-reserves mechanisms in an attempt to improve transparency, such mechanisms have a fundamental limitation. Proof of reserves is usually based on an asset snapshot at a particular point in time combined with a Merkle tree structure; while it can prove that the exchange controls a certain quantity of assets, it often cannot provide proof of liabilities of equal strength [34]. Worse, the transparency of this off-chain system is provided voluntarily by the exchange and can be withdrawn at any time, rather than being enforced by the underlying architecture. The lag of the snapshot also leaves the exchange room to move funds temporarily in order to pass an audit.

At the matching and execution layer, a centralized exchange's internal order book and matching engine are entirely opaque to the outside. The user must trust the exchange

to execute orders fairly, strictly according to price priority and time priority, without exploiting an informational advantage to front-run, and without deliberately delaying retail orders under extreme market conditions to protect the interests of its proprietary market-making desk. Because the matching process lacks verifiable cryptographic evidence, users who suffer unfair treatment often cannot prove it.

The clearing and settlement layer is likewise an opaque black box. Post-trade clearing of the books (the recomputation of positions, P&L, and margin), mark-to-market and funding settlement, and the triggering and execution of liquidations are all carried out inside the exchange. Although the exchange usually discloses its margin-calculation formulas and liquidation-triggering rules, the specific execution of these rules cannot be audited. During violent market swings, the exchange holds the implicit power to modify the mark-price algorithm or adjust liquidation parameters. The original intent of this flexibility may be to prevent cascading liquidations from triggering a systemic collapse, but in substance it grants the exchange the unilateral power to decide the fate of users' assets.

The data layer and the governance layer further intensify the concentration of trust. The exchange is both the sole provider of market data and the sole maker of the system's rules. Users have no way to verify the authenticity of order-book depth, nor any way to guard against fake volume misleading market sentiment. As for governance, fee adjustments, the listing of new contracts, and modifications of market-maker agreements are all decided unilaterally by the exchange's management, without due process or community oversight.

Table 28-3 maps the foregoing analysis onto the five layers of the trust stack, presenting the trust state and verifiability of a centralized exchange at each layer. The *trust state* column reflects the degree of trust the user must deliver to the exchange operator at that layer, while the *verifiability assessment* column measures the technical feasibility of an external observer independently confirming the operating condition of that layer. Reading the two columns together reveals a central contradiction: the centralized exchange demands complete trust from users at every layer yet offers effective external means of verification at almost none.

Layer	Trust state	Object of trust	Verifiability assessment
Asset custody	Complete trust	Exchange operator	Very low; proof of reserves offers only a limited static snapshot
Matching and execution	Complete trust	Exchange operator	None; the internal matching-engine logic is invisible to outsiders
Clearing and settlement	Complete trust	Exchange operator	Low; the rules are public but the specific execution cannot be audited
Data and oracle	Complete trust	Exchange operator	Low; data authenticity relies entirely on the platform's self-discipline
Governance	Complete trust	Exchange management	Very low; rule changes lack transparency and any mechanism for participation

Table 28-3. A trust-stack audit of a typical centralized exchange (Data source: compiled by the author)

Table 28-3 shows clearly that the typical centralized exchange exhibits, without exception, a state of *complete trust* across all five layers, with verifiability ranging from *very low* to *none*. This severe mismatch between trust and verifiability means that users have the fewest means of verification precisely where they most need protection. Taken together, the trust concentration of a typical centralized exchange reaches 100%. This mode of fully merging the trading, custody, clearing, and settlement functions within a single entity would be unthinkable in the traditional financial system [35]. As discussed in Section 28.3, traditional finance has, through institutional evolution, built a framework of mandatory functional separation. In pursuing the highest performance, the monolithic black box architecture of the centralized exchange sacrifices every attribute of trust minimization.

28.5.2 The trust stack of AMM-type DEXs

AMM-type decentralized exchanges mainly serve spot swaps and are not in the same category as the perpetual futures derivatives on which this book focuses; they are included here only within the trust-stack audit, as one link in the evolutionary lineage of four architecture types, and are not elaborated. Their core achievement is a high degree of trust minimization at the asset-custody layer through smart contracts: user assets are locked in an on-chain liquidity-pool contract, deposits and withdrawals strictly follow public code logic, no single entity holds the private-key authority to misappropriate funds, and the custody-default risk of a centralized exchange is thereby fundamentally eliminated.

Yet while the automated market maker solves custody-layer trust, it introduces new challenges at the other layers. The matching and execution layer abandons the central limit order book (CLOB) and instead relies on algorithmic pricing such as the constant-

product formula; this eliminates trust in a centralized matching engine, but the execution process is exposed in the public mempool, making trades highly vulnerable to MEV attacks, in which validators can extract value from users through front-running or sandwich attacks. At the clearing and settlement layer, because the AMM mainly serves the spot, its logic is relatively simple, usually taking the form of instant swaps of tokens within the liquidity pool; once extended to derivatives, however, it must rely on an external oracle to supply the mark price for computing the funding rate and triggering liquidations, thereby shifting the trust risk to the data layer. The data and oracle layer thus becomes the most fragile link in this type of architecture: once the oracle is manipulated or fails, it can lead to large-scale erroneous liquidations, and even multi-source aggregation and economic staking cannot eliminate this risk entirely. The governance layer generally adopts a DAO model, which improves transparency but faces the problems of voting power concentrating among large holders and low decision-making efficiency.

Table 28-4 summarizes the foregoing analysis within the trust-stack framework. AMM-type DEXs display a marked divergence: the custody layer achieves a breakthrough close to zero trust, while the execution and data layers retain residual trust of differing forms. Unlike the uniform *complete trust* of the centralized exchange in Table 28-3, their trust distribution varies as a gradient across the layers.

Layer	Trust state	Object of trust	Verifiability assessment
Asset custody	Highly trustless	On-chain smart contract	High; the contract code is publicly auditable, and asset flows are traceable on chain
Matching and execution	Partially trustless	Algorithmic pricing and block-ordering parties	Moderate; the pricing logic is transparent, but transaction ordering is subject to MEV
Clearing and settlement	Highly trustless (spot) / partially trustless (derivatives)	Smart contracts and external oracles	High (instant spot settlement) / moderate (derivatives depend on oracles)
Data and oracle	Partial trust	Decentralized oracle network	Moderate; multi-source aggregation, but a manipulation window remains
Governance	Partially trustless	DAO and token holders	Moderate; the voting process is transparent on chain but may be dominated by large holders

Table 28-4. A trust-stack audit of AMM-type decentralized exchanges (Data source: compiled by the author)

The key pattern that Table 28-4 reveals is *solid at the base, weak at the top*: the custody layer and the spot clearing layer reach *high* verifiability, but this gradually gives way

to *partial trust* as one moves up the layers. The MEV problem at the matching and execution layer and the manipulation window at the data and oracle layer constitute the two persistent bottlenecks of this type of architecture, and they also point to the directions in which the later app-chain architecture improves.

28.5.3 The trust stack of app-chain DEXs

The app-chain decentralized exchanges exemplified by Hyperliquid and the fourth-generation dYdX represent the current frontier of engineering exploration on the trust-performance frontier. By *app-chain*, this chapter means an independent blockchain network customized for a single application scenario, whose consensus mechanism, state machine, and network parameters are all deeply optimized around a specific use (such as derivatives trading). An app-chain differs from a general-purpose rollup in that a rollup inherits the security of the underlying chain and submits state proofs to it, whereas an app-chain maintains an independent validator set and consensus process. An app-chain differs from a traditional sidechain in that a modern app-chain is usually based on BFT consensus (such as Cosmos SDK's Tendermint or Hyperliquid's HyperBFT) rather than a simple multi-signature scheme or merged mining. dYdX v4's migration from StarkEx (an Ethereum-based rollup solution) to an independent app-chain built on the Cosmos SDK is a landmark case of this architectural paradigm shift, motivated precisely by the desire to obtain full customization of the consensus and execution layers so as to move order-book matching fully on chain. These protocols abandon the traditional path of building applications on a general-purpose smart-contract platform and instead develop an independent blockchain optimized specifically for trading. This architectural choice achieves an unprecedented, finely balanced equilibrium across the layers of the trust stack.

At the asset-custody layer, the app-chain architecture exhibits partial trustlessness. Because the app-chain itself typically does not host native assets, users must transfer funds from a mainnet such as Ethereum onto the app-chain through a cross-chain bridge. The security of the cross-chain bridging contract therefore becomes the core trust assumption of this architecture. Historical data show that cross-chain bridges are the infrastructure most prone to attack in the decentralized finance ecosystem; by cumulative amount stolen, at one point more than 40% of major hacking losses in 2022–2023 occurred at this link [36]. An app-chain usually relies on its validator set to jointly maintain the security of the cross-chain bridge, which means that the safety of users' assets shifts from trusting Ethereum's mainnet consensus to trusting the app-chain's local validator set.

The matching and execution layer is where the app-chain architecture's core innovation lies. By natively integrating a central limit order book at the consensus layer, these systems achieve a highly trustless matching process. All order submissions, cancellations, and matches are executed within a block, and the results are visible to everyone and immutable. Although the physical constraints of distributed consensus keep an app-chain's block time at the level of a few hundred milliseconds—unable to reach a centralized exchange's microsecond latency—this performance compromise buys complete transparency at the execution level, an acceptable trade-off for the great majority

of non-high-frequency traders [37].

The clearing and settlement layer likewise achieves a high degree of trustlessness. Post-trade clearing of the books, mark-to-market, and funding settlement all execute deterministically on chain according to uniform rules, and the margin requirements, liquidation-triggering conditions, and the operating mechanism of the insurance fund are all hard-coded into the app-chain's underlying protocol. Every liquidation event can be audited precisely through on-chain data, sharply compressing the room for implicit governance maneuvering that is pervasive in centralized exchanges (though the specific liquidation parameters can still be adjusted through the governance process; see Section 28.6.3).

At the data layer, for crypto-native assets, an app-chain can directly use the trade prices generated by its on-chain order book, thereby freeing itself to a great extent from dependence on an external oracle. Only when derivatives on traditional financial assets are introduced does it again face the trust degradation that oracles bring.

At the governance layer, however, current app-chain projects generally exhibit partial trust. To ensure the system's capacity for rapid iteration and its ability to respond to sudden events, the core development team often retains control over protocol upgrades and the adjustment of key parameters. Although these projects usually promise to follow a roadmap of progressive decentralization, at the present stage the concentration of governance power remains a link in their trust stack that cannot be ignored.

Table 28-5 maps the trust state of the app-chain architecture onto the five-layer model. Unlike the centralized exchange's across-the-board *complete trust* and the AMM-type DEX's *solid at the base, weak at the top* pattern, the app-chain architecture displays a distinctive distribution—*solid in the middle, under pressure at both ends*: the three middle layers of matching, clearing, and data all reach *highly trustless*, while the custody layer at one end (constrained by cross-chain bridge security) and the governance layer at the other (constrained by concentrated team control) each retain trust assumptions of differing degrees.

Layer	Trust state	Object of trust	Verifiability assessment
Asset custody	Partially trustless	Cross-chain bridge contract and validator set	Moderate; traceable on chain but exposed to the security risk of cross-chain bridging
Matching and execution	Highly trustless (residual validator-MEV risk exists)	On-chain order book and consensus mechanism	High; all matching logic and results are fully public on chain, but validators still have the potential ability to manipulate transaction ordering
Clearing and settlement	Highly trustless	Protocol-level hard-coded rules	High; the liquidation process strictly follows the code and is auditable after the fact
Data and oracle	Partially trustless	On-chain native data-generation mechanism	Moderate; crypto-asset prices are generated by on-chain trades but face the risk of self-referential price manipulation
Governance	Partial trust	Core development team and early validators	Moderate; the decision process has limited transparency and relies on the team's self-discipline

Table 28-5. A trust-stack audit of app-chain decentralized exchanges (Data source: compiled by the author)

Comparing Tables 28-3, 28-4, and 28-5, a clear evolutionary trajectory emerges: from the centralized exchange's across-the-board *complete trust* over all five layers, to the AMM-type DEX's first breakthroughs at the custody and clearing layers, and then to the app-chain architecture's extension of *highly trustless* to the matching and data layers. The trustlessness assessment of the data and oracle layer, however, requires a more cautious correction: the JELLY (JELLYJELLY) incident of March 2025 showed that even when prices are generated entirely by on-chain trades, the price of a low-liquidity token can still be maliciously manipulated, forcing the market-making vault (HLP) to take over the liquidated position and bear an enormous unrealized loss (peaking at about $13.5 million), which exposed the inherent fragility of the self-referential pricing mechanism. The incident was ultimately quelled only by a vote of the validator set to force a delisting and settle at an off-market price (see Section 28.6.4), which simultaneously exposed discretionary intervention at the governance layer [38]. By the trust concentration metric proposed in Section 28.1.4, the app-chain architecture retains a dependence on a single entity at only the governance layer and part of the custody layer (the cross-chain bridge) among the five layers of the trust stack, giving an equal-weighted trust concentration of about 30% to 40%, markedly below the centralized exchange's 100%. Under the weighting scheme proposed in Section 28.1.4, because the app-chain achieves a high degree of trustlessness at the three middle layers of clearing,

matching, and data and retains a dependence only at the custody and governance layers, its weighted trust concentration is about 25% to 30%. The difference between the two measures arises not from the middle layers (which are trustless under both measures, contributing close to zero to the residual concentration) but from the weighting's reallocation of weight to the custody layer (0.35): the app-chain's core residual risk—the cross-chain bridge—happens to sit at this custody layer whose weight is raised, so the weighted measure in fact highlights the cross-chain bridge's status as the decisive residual risk rather than systematically producing a more optimistic figure. Table 28-5 also lays bare, however, the app-chain's two structural weak points: the cross-chain bridge risk at the custody layer and the team's concentrated control at the governance layer, and these two links constitute the main bottlenecks to the further decentralization of the app-chain architecture.

28.5.4 The trust stack of hybrid architectures

Between the centralized exchange's monolithic black box and the app-chain's fully on-chain approach, an emerging model known as the hybrid architecture is gradually taking shape. The core design philosophy of this architecture is selective trustlessness: applying strict de-trusting at the layers most likely to cause catastrophic consequences, while retaining centralized, high-efficiency operation at the layers with the highest performance requirements.

The hybrid architecture first undertakes a thorough reform at the asset-custody layer. Users' funds are no longer deposited into a wallet controlled by the exchange but are instead held in an independent on-chain smart contract or a regulated third-party MPC custody system. This design ensures that the exchange cannot unilaterally misappropriate customer assets and greatly reduces the risk of the operator doing so. Under extreme market conditions, however, on-chain asset withdrawal may face frictions such as delay from network congestion, spiking gas fees, and cross-chain bridge congestion, so that the theoretical *instant exit* cannot always be realized frictionlessly in practice. The combination of multi-party computation and multi-signature technology requires that asset transfers pass through multi-party authorization, substantially raising the security boundary of the custody layer [39].

At the clearing and settlement layer, the hybrid architecture tends to move rule execution on chain. After the exchange completes matching off chain, it submits the results to an on-chain smart contract for the final confirmation of clearing the books, settling funds, and liquidating positions. This design guarantees the transparency and immutability of the rules, and users can audit the system's solvency and the fairness of its liquidation logic at any time.

At the matching and execution layer, however, the hybrid architecture continues to retain an off-chain centralized engine. This is because, for institutional investors and high-frequency market makers, microsecond latency and extremely high throughput are indispensable. By keeping the matching process off chain, the hybrid architecture can offer liquidity depth and a trading experience fully consistent with those of a traditional centralized exchange. Although this retains, to some degree, a trust black box at the execution level, by introducing cryptographic techniques such as zero-knowledge proofs,

some hybrid architectures have begun to attempt to provide after-the-fact verifiability of matching results without sacrificing performance.

The trust state of the data layer and the governance layer in a hybrid architecture usually falls between the two extremes. Data are still generated mainly by a centralized engine, but because settlement occurs on chain, fabricating fake volume becomes more difficult and costly. The governance layer may adopt a mechanism combining a multi-signature council with a time lock, providing users an exit window while preserving responsiveness.

The trust distribution of the hybrid architecture displays an *interleaved* character quite unlike that of the three preceding architectures: rather than pursuing uniform de-trusting across all layers, it directs limited engineering resources precisely at the layers with the most severe consequences (custody and clearing) while retaining centralized, high-efficiency operation at the performance-sensitive layer (matching). This strategic trade-off is clearly reflected in the table below: two starkly different trust states—*highly trustless* and *partial trust*—coexist within the same architecture, forming a sharp layered contrast.

Layer	Trust state	Object of trust	Verifiability assessment
Asset custody	Highly trustless	On-chain smart contract or third-party MPC custody	High; asset segregation is enforced by code or multi-party authorization
Matching and execution	Partial trust	Off-chain centralized matching engine	Low to moderate; some systems introduce ZKPs to provide after-the-fact verifiability
Clearing and settlement	Highly trustless	On-chain smart contract	High; the settlement rules are transparent and immutable on chain
Data and oracle	Partial trust	Generated by a centralized engine, constrained by on-chain settlement	Moderate; the cost of fabricating fake data is raised by on-chain settlement
Governance	Partial trust	Multi-signature council combined with a time lock	Moderate; rule changes are procedurally constrained but still led by a few

Table 28-6. A trust-stack audit of the hybrid architecture (Data source: compiled by the author)

Table 28-6 reveals the core design logic of the hybrid architecture: the verifiability of both the custody layer and the clearing layer reaches the *high* grade, which ensures the safety of user assets and the fairness of rule execution—even if the matching engine is operated entirely by a centralized entity, the final ownership of funds is still adjudicated by on-chain contracts. Although the matching and execution layer retains *partial trust*, techniques such as zero-knowledge proofs are gradually closing this trust gap. Comparing Tables 28-3 through 28-6, the evolutionary path of the four architectures along the

trust stack is plainly visible: from the centralized exchange's across-the-board opacity, to the AMM-type DEX's breakthrough at the base, to the app-chain's extension in the middle, and then to the hybrid architecture's precise trade-offs across the layers. The hybrid architecture represents the industry's pragmatic choice following a deep understanding of the trust-performance frontier. It acknowledges that different types of market participants have starkly different sensitivities to trust. Retail investors may value the full safety of their assets more highly, whereas high-frequency trading firms are willing to bear a degree of verifiable trust risk for the lowest execution latency. Through modular interface design, the hybrid architecture allows components at different layers to be freely combined, pointing to a path of progressive evolution for the on-chain financial infrastructure of the future.

28.6 Hyperliquid's architectural choices

Turning to the frontier of engineering practice, Hyperliquid is one of the protocols most representative of the current state of the trust-performance frontier. As an app-chain designed specifically for high-performance derivatives trading, each of its architectural choices—from building its own layer-1 network and a built-in central limit order book to the design of its liquidity-provider vault and validator set—is a carefully considered trade-off between trust and performance. Dissecting its architecture shows clearly how the layered-design philosophy comes to ground under real engineering constraints, and at which layers the system compromises in pursuit of the highest performance.

28.6.1 The app-chain and the on-chain order book

Hyperliquid's technology stack is built on a specially customized layer-1 blockchain. Its core architecture comprises an app-chain based on the HyperBFT consensus algorithm, a fully on-chain central limit order book, a bridge contract connecting to Ethereum, and a liquidity-provider vault that handles automated market-making along with an on-chain liquidation engine [40] [6]. The first core question raised by this architectural choice is: why build its own layer-1 network rather than build on an Ethereum layer-2 network?

Building on an Ethereum layer-2 network (such as Arbitrum or Optimism) means one can inherit the security of the Ethereum mainnet but must also inherit its performance constraints. The latency of a layer-2 network is usually greater than one second—a level of latency unacceptable to high-frequency traders and market makers—and it comes with non-negligible gas costs [41]. The performance limits of a layer-2 network stem from its architectural design. To inherit Ethereum's security, a layer-2 network must periodically submit state proofs or transaction data to the Ethereum mainnet, and the speed of this process is limited by the mainnet. Beyond that, transactions on a layer-2 network must ultimately be confirmed on the Ethereum mainnet, which adds to overall latency. In choosing to build its own layer-1 network, Hyperliquid in essence treats the app-chain as a dedicated chip customized for the single purpose of perpetual futures trading, rather than emulating a trading environment on a general-purpose processor. Through deep customization from the underlying network stack to the consensus algorithm, Hyperliquid achieves deep optimization for end-to-end latency [42].

This specialized design means that every component of the system can be optimized for trade processing. The storage structure, for example, can be optimized for the query patterns of the order book, the network protocol for the communication patterns of high-frequency trading, and the consensus algorithm for the rapid confirmation of trades.

The HyperBFT consensus algorithm is the core of this optimization. It draws on the design ideas of the HotStuff consensus protocol [43] and its successors LibraBFT and DiemBFT, but with several key improvements targeting the specific needs of a trading system [44]. Compared with the three-phase commit process of the base HotStuff, HyperBFT adopts the DiemBFT/Jolteon-style two-phase (two-chain) commit rule, which removes one round of inter-validator communication relative to the three-phase version and thereby lowers the confirmation latency of a single block. The two-phase commit is itself a contribution of the predecessor protocols (DiemBFT/Jolteon), not an original invention of HyperBFT relative to HotStuff. On the normal happy path, HyperBFT advances consensus through optimistic responsiveness: the rate of block production depends on the actual communication speed of a validator quorum rather than on a preset fixed timeout parameter, which means that when network conditions are good, consensus latency can be far below the protocol's theoretical upper bound. Optimistic responsiveness governs this happy path (which does not depend on a known time bound), whereas the synchrony assumption is used only in scenarios such as a view change (that is, leader replacement) to guarantee liveness: when network latency exceeds expectations, the protocol falls back to a more conservative consensus path. Whatever the state of the network, the protocol's safety (that is, not producing mutually contradictory state confirmations) is never conditioned on the synchrony assumption.

HyperBFT also introduces two optimizations: pipelined block processing and optimistic execution. Pipelining sends the quorum certificate of the previous block together with the proposal for the next block, so that multiple blocks advance in parallel through the proposal and certification phases. Optimistic execution allows a transaction to begin executing before the block is finally confirmed, further shortening the end-to-end latency the user perceives. Optimistic execution, however, introduces a theoretical risk of block rollback: in extreme cases, an optimistically executed transaction may be reversed because consensus was not reached, and users must distinguish two states—*optimistic confirmation* (low latency but not final) and *final confirmation* (verified by consensus and irreversible). The combined effect of these engineering optimizations lets HyperBFT achieve a median consensus latency of about 200 milliseconds while maintaining the safety guarantee of a BFT protocol. This performance figure, however, is measured under optimistic conditions. Under adversarial conditions—such as some validators going offline or a network partition—block-production latency may spike significantly, and consensus may even stall temporarily. While consensus is paused, the liquidation engine cannot execute protective liquidations, and risk exposure accumulates continuously; this constitutes the inherent fragility of high-performance BFT consensus under extreme conditions.

This technical choice inevitably comes at a cost, however. Building its own layer-1 network means the system must rely on its own validator set for security, rather than

being backstopped by the vast computing power of the Ethereum mainnet. Under this architecture, the system's trust anchor shifts: users no longer trust the overall security of the Ethereum ecosystem but must instead trust Hyperliquid's specific validator set [45]. This shift of trust is the starting point for understanding Hyperliquid's architectural trade-offs; it reveals that in pursuing the highest performance, one must make a degree of compromise at the foundational security layer.

28.6.2 An analysis of the trust state at each layer

As the main instance of the *app-chain* archetype audited in Section 28.5.3, Hyperliquid's overall trustlessness at each layer is consistent with the judgments in that section; this section places it within the trust stack model and audits it, focusing on the increments specific to it. At the asset-custody layer (Layer 1), user assets are transferred from the Ethereum network to the Hyperliquid network mainly through a bridge contract. The security of this process depends entirely on the honesty of the validator set and the robustness of the bridge code [46]. At present, a deposit takes effect only after obtaining signatures from validators representing more than two-thirds of the stake weight [47]. As noted in Section 28.5.3, cross-chain bridges are the infrastructure most frequently attacked in decentralized finance—the Ronin bridge, for example, was hacked in 2022 for a loss of more than $600 million—and they therefore constitute one of the weakest links in Hyperliquid's trust stack.

Although the multi-signature mechanism requiring signatures from more than two-thirds of the validators provides some protection, the safety of user assets ultimately depends on the honesty of the validator set and remains inferior to holding assets directly on the Ethereum mainnet—the specific residual risks of the validator set and the cross-chain bridge are detailed in Section 28.6.4.

At the matching and execution layer (Layer 2), Hyperliquid places its central limit order book entirely on chain, with all orders matched within blocks and the results publicly visible [48], so that users can independently verify whether price-time priority is honored. As audited for the app-chain matching layer in Section 28.5.3, this layer is among the most trust-minimized in the entire system; although its latency of about 200 milliseconds is somewhat slower than a centralized exchange's microsecond response, it is entirely acceptable for the great majority of retail users and for some institutional traders [11].

The rules of the clearing and settlement layer (Layer 3) are likewise hard-coded in the on-chain Hypercore execution state machine: post-trade clearing of the books, continuous mark-to-market, funding payments, and unrealized-P&L updates all execute on chain according to deterministic rules and can be recomputed and checked by any node; every liquidation record of the liquidation engine can also be publicly queried; and the takeover of liquidated positions by the insurance fund and the liquidity-provider vault proceeds transparently on chain [49]. As audited in Section 28.5.3, this transparency far exceeds that of a centralized exchange. Here, however, two levels of *hard-coding* must be distinguished: the execution framework of the liquidation logic (such as the formula for computing the liquidation-triggering conditions and the execution flow) is truly immutable logic hard-coded into the protocol, whereas the specific liquidation

parameters (such as the maintenance margin ratio and the maximum leverage) are variables that can be adjusted through the governance process. The rigidity of the former limits the flexibility of the rules: any framework-level modification requires a protocol upgrade, lacking the buffer that human intervention provides in a centralized system for handling extreme market conditions. During the 2008 financial crisis, the central counterparties of the traditional financial system could flexibly adjust their liquidation rules according to market conditions, and this flexibility was sometimes the key to preventing the spread of systemic risk. Although Hyperliquid's hard-coded rules are more transparent, they may lack the flexibility necessary under extreme market conditions.

At the data and oracle layer (Layer 4), the prices of crypto-native assets are generated directly by on-chain trades without an external oracle, achieving—as audited in Section 28.5.3—an extremely high degree of trust minimization. This self-referential pricing, however, carries an inherent fragility for low-liquidity assets: as the JELLY incident described in Section 28.5.3 showed, an attacker can still distort the perpetual mark price by manipulating the on-chain spot price, so this layer is not unconditionally one of *extremely high trust minimization*. At the governance layer (Layer 5), Hyperliquid is currently highly concentrated: the development team controls protocol upgrades, parameter adjustments, and the strategy of the liquidity-provider vault [50], and although it has declared a path of progressive decentralization, the current concentration of governance makes this the layer where the gap from a centralized exchange is smallest (the governance residual is detailed in Section 28.6.4).

28.6.3 The performance-latency trade-off

In a layered architecture, gains in performance often come at the cost of sacrificing some trustlessness. As a Byzantine fault-tolerant consensus protocol, HyperBFT follows the basic safety principle of BFT consensus by design: it prioritizes safety (that is, never producing mutually contradictory state confirmations) and, when the network partitions or inter-validator communication is severely delayed, chooses to pause block production, sacrificing liveness. This means the system would rather halt service briefly than confirm an erroneous trade result. Hyperliquid's performance optimization is achieved not by relaxing its safety guarantee but through engineering means—reducing the number of consensus communication rounds, optimizing the network topology among validators, and adopting pipelined block processing—to maximize throughput and minimize latency without impairing safety. Compared with a fully centralized matching engine, however, the communication overhead of distributed consensus still constitutes an irreducible lower bound on latency. According to official figures, for geographically close clients, Hyperliquid's median end-to-end latency is 0.2 seconds and its 99th percentile is 0.9 seconds [11]. By contrast, a centralized exchange's matching latency is typically in the microsecond range. A quantitative comparison of Hyperliquid's key performance metrics with these two benchmark architectures can pinpoint its position on the trust-performance frontier. In Table 28-7, four metrics characterize the performance differences among the three architectures from different dimensions: end-to-end latency reflects the user experience, throughput capacity reflects the system's carrying capacity, actual operating throughput reflects the match to market demand, and

finality reflects the time window of trade certainty.

Performance metric	Hyperliquid (app-chain)	Top centralized exchange	Typical Ethereum layer-1 DEX
Matching/consensus latency	~200 ms (end-to-end, consensus confirmation)	~5–50 microseconds (inside the matching engine; ~10–100 ms end-to-end for ordinary remote users)	~12 seconds (block confirmation)
Throughput capacity	~200,000 orders/second	Hundreds of thousands to millions/second	Tens to hundreds/second
Actual operating throughput	No independently verified public peak figure yet	Extremely high, depending on market volatility	Limited by block space
Finality	Single-block finality	Instant confirmation by the internal system	Requires waiting for multiple block confirmations

Table 28-7. A comparison of performance metrics between Hyperliquid and other architectures (Data source: compiled by the author based on [10][11][12])

Note: The "throughput capacity" in the table is measured in orders per second, encompassing all operations such as placing, canceling, and modifying orders. The official documentation states that the current mainnet supports about 200,000 orders per second, with the execution layer being the current bottleneck and the consensus and network stack theoretically scalable to millions of orders per second. As of this writing, there is no independently verified public figure for peak on-chain confirmed transactions (TPS), so the "actual operating throughput" row no longer provides a specific value.

For high-frequency market makers, the latency gap from microseconds to 200 milliseconds is decisive. This explains why, under certain extreme market conditions, Hyperliquid's liquidity depth still struggles to fully match that of leading centralized exchanges, because high-frequency market makers cannot execute optimal market-making strategies amid millisecond-scale market swings. The profit model of high-frequency market-making relies on millisecond-scale reactions to market microstructure, and microsecond-latency arbitrage strategies are entirely infeasible on Hyperliquid. Moreover, the 200-millisecond latency also has a substantive impact on non-high-frequency traders: stop-loss orders may face slippage of several price ticks during violent volatility, market makers' delay in responding with cancellations increases the risk of adverse selection, and the execution-time difference of the liquidation engine under extreme conditions may cause the actual liquidation price to deviate significantly from the trigger price. Although these frictions are not fatal, they constitute a systemic cost that all participants must factor into their strategies.

This latency difference also affects the overall quality of market liquidity: on a centralized exchange, high-frequency market makers provide abundant liquidity, allowing traders to fill large orders near the mid-price, whereas on Hyperliquid high-frequency market-making participation is limited, liquidity comes mainly from ordinary market makers and the protocol's built-in algorithmic market-making vault (such as the HLP Vault), and the bid-ask spread may be wider.

In terms of throughput, the Hyperliquid mainnet currently supports a processing capacity of about 200,000 orders per second [11] [51]. Although the consensus algorithm and network stack can in theory scale to the order of one million orders per second, the current bottleneck lies in the execution layer [52]. This means that as the number of users and trading volume grow, the system may run up against a performance ceiling. Compared with an Ethereum layer-1 DEX, however, Hyperliquid's performance is already several orders of magnitude greater.

Yet although Hyperliquid makes the block-construction process more controllable by building its own layer-1 network, the problem of maximal extractable value does not thereby disappear. Validators still have, in theory, the ability to order and manipulate transactions. A validator could, for example, prioritize the inclusion of its own transactions or delay others' transactions for gain. This existence of maximal extractable value is an inherent problem of all blockchain systems, but on Hyperliquid the problem may be more pronounced because the validator set is small.

The Hyperliquid team has recognized this problem and is exploring various mitigations. It is considering, for example, introducing a timestamp-based order-sequencing mechanism to reduce validators' room to manipulate transaction ordering. In addition, increasing the number and geographic distribution of validators can further lower any single validator's ability to manipulate maliciously. Hyperliquid's high performance is therefore not without cost: it is a relative advantage obtained by concentrating trust in a specific validator set and optimizing the execution layer to the extreme.

28.6.4 Hyperliquid's residual trust

In assessing the robustness of any decentralized system, identifying its irreducible residual trust plays an important role. For Hyperliquid, the largest residual trust lies first in its validator set. In its early days, the system ran on just five validators; it then expanded to 16 genesis validators, converted to 21 permissionless nodes in April 2025, and by mid-2026 had grown to about 24 active validators (with the team announcing a further expansion to 27) [53] [54]. Although the team plans to keep enlarging the validator set, at the present stage the number and geographic distribution of validators remain relatively concentrated.

This number needs to be compared with other systems to gain a clearer picture. The Ethereum mainnet crossed the milestone of one million active validators in March 2024, with about 32 million ETH staked, accounting for 26% of total supply [55]. Even a relatively small proof-of-stake chain usually has hundreds to thousands of validators. Hyperliquid's twenty-odd validators are, by comparison, quite concentrated. More critically, the number of validators is only a surface indicator of security; the distribution of stake is the core variable that determines actual independent security. According to reports, the validators under the foundation collectively control about 81% of the staked tokens, which means that even though there are nominally twenty-odd validators, the actual concentration of power is far higher than the count suggests—this proportion exceeds both the one-third threshold needed to halt the chain and the two-thirds threshold needed to control the cross-chain bridge, and the system's number of effective independent validators may be far below its nominal validator count. If the validator

set were attacked or internal collusion occurred, the security of the entire app-chain would be gravely threatened, and user assets would face direct risk. This is the core fragility of the self-built app-chain model: the system's security depends entirely on its own validators rather than on a node network as vast and dispersed as the Ethereum mainnet's.

Second, the cross-chain bridging mechanism from Ethereum to the Hyperliquid layer-1 network constitutes another significant point of centralization risk. The upgrade authority over the bridge contract, its multi-signature configuration, and the emergency-pause power all embody a high concentration of trust. Because all deposits and withdrawals of assets must pass through this bridging channel, it becomes, in substance, the critical conduit for the flow of funds through the entire system. If the bridge contract has a vulnerability or is attacked, users may be unable to withdraw funds, or funds may be stolen outright. The security problems of cross-chain bridges have already caused several major losses in the history of DeFi. The attack surface of a cross-chain bridge is not limited to smart-contract vulnerabilities: supply-chain attacks (such as the penetration of a signer's operating environment or the planting of malicious code in a dependency library) and common-mode failures (such as all validators running the same version of the software stack, so that a single vulnerability can breach all nodes at once) constitute equally serious threat vectors.

Finally, the development team's dominant position in governance is a notable compromise within the decentralized-finance narrative. According to reports, the associated foundation controls the vast majority of the staked tokens, which gives the team near-complete control over key decisions such as protocol upgrades and adjustments to the vault-strategy parameters [56]. Since HIP-3 went live on the mainnet in October 2025, qualified builders who meet the staking threshold have been able to deploy their own perpetual markets permissionlessly, so the listing of new contracts has shifted from complete team control toward partial decentralization. This concentrated governance power means that if the team makes a decision adverse to users, users have almost no way to stop it. Although Hyperliquid has proposed mitigation paths—increasing validator diversity, open-sourcing the bridge contract, and gradually transitioning governance power to the community—these remain, for now, plans on a roadmap rather than realized system features. Moreover, there is a reflexivity risk between the governance token's price and governance power: a rise in the token price increases the cost of a governance attack (raising security) but also increases the potential gain a large holder could obtain through governance manipulation (lowering security), and this positive feedback loop makes governance security hard to predict under extreme market conditions.

The problem with this governance concentration is that it leaves the system vulnerable to errors in the team's decisions. If the team decides to revise the liquidation parameters, adjust the fee structure, or change the incentive mechanism for liquidity providers, users cannot express their dissent through voting or any other democratic mechanism. This runs counter to the traditional ideal of decentralized finance but may be a necessary trade-off for achieving high performance.

28.6.5 The generalizability of the app-chain model

Hyperliquid's success has drawn wide industry attention, and its architectural pattern of an app-chain combined with an on-chain central limit order book seems to point to a viable path for decentralized derivatives trading. Other leading protocols have adopted similar approaches: dYdX v4 builds its app-chain on the Cosmos ecosystem, and Vertex is exploring a similar architecture. This architectural pattern is fully replicable technically. Any team with sufficient engineering capability can build an app-chain, implement an on-chain order book, and connect to Ethereum.

What is truly hard to replicate in the Hyperliquid model, however, is the specific detail of its performance optimization and its cold-start strategy for liquidity. The system's ability to process high-concurrency orders steadily at 200-millisecond latency depends on the team's deep customization and continuous optimization of the consensus algorithm and the underlying network stack [57]. This requires profound knowledge of systems programming and distributed systems. At the same time, the resilience its liquidity-provider vault displayed during a market downturn, together with the nearly $3 trillion in full-year trading volume and the fourfold user growth achieved in 2025, is owed more to the team's outstanding execution and its first-mover advantage in ecosystem-building [58] [59] [60]. These soft strengths and this accumulation of market presence cannot be obtained merely by copying the architecture's code.

A more fundamental problem is that the proliferation of the app-chain model may increase systemic risk. If hundreds of similar app-chain exchanges emerge in the future, each needing to assemble and maintain an independent validator set to secure itself, this will inevitably lead to a fragmentation of security. Each app-chain's economic bandwidth and capacity to defend against attack will be limited. Validator incentives may be insufficient to attract enough high-quality participants, causing the quality of the validator set to decline.

This problem can be understood through the economic concept of *the tragedy of the commons*. Validators are a scarce resource, all the more so high-quality validators with security expertise and infrastructure. If every app-chain must attract its own validator set, validators' compensation must be high enough to draw their participation. But this raises the app-chain's costs, which are ultimately passed on to users. At the same time, because validators are dispersed, the security of each app-chain declines, since its validator set is smaller and more easily attacked.

To alleviate this problem, introducing a shared security layer (such as restaking protocols like EigenLayer) may become a direction of future evolution. Letting multiple app-chains share underlying economic security can lower the cost of securing each independently, but it also introduces more complex cross-chain trust dependencies and potential systemic cascade risk. If the shared security layer itself is attacked, all app-chains that depend on it are affected. More seriously, in restaking the same stake secures multiple app-chains (actively validated services, AVSs) at once, so aggregate security is shared rather than simply summed, and slashing across different app-chains—together with operator failures—becomes correlated. This reintroduces precisely the common-mode failure and correlation risk warned of in Section 28.1.3 [61]. This model shifts trust from a single app-chain's validator set to the validator set of the shared

security layer, but it does not fundamentally solve the problem of security fragmentation.

28.7 The engineering judgment of sufficient decentralization

In the architectural design of on-chain financial systems, a common misconception is to treat *decentralization* as a strict binary state or an ultimate goal to be pursued without limit. Starting from the first principles of engineering, however, the core task facing a systems architect is not to maximize decentralization at any cost but to identify precisely and minimize the unacceptable trust assumptions in the system. This paradigm shift from idealism to realism forms the basis of the engineering judgment of sufficient decentralization [62].

The concept of *sufficient decentralization* acknowledges the objective existence of the trust-performance frontier discussed above. Under given technical constraints, every additional degree of on-chain verification depth inevitably incurs a loss of performance. Good architectural design should therefore provide appropriate trust guarantees for different layers and different user groups, rather than constraining the entire system with a single rigid standard.

28.7.1 Defining 'sufficient'

In the context of financial infrastructure, defining *sufficient decentralization* requires returning to the bottom-line logic of protecting users' interests. A system's meeting the *sufficient* standard does not mean it achieves full peer-to-peer consensus in every technical detail; it means it has established effective safeguards at the critical risk nodes, so that the cost to any single entity of acting maliciously far exceeds the gain it could obtain. This security condition can be formalized as a cost-benefit attack model: the cost of acting maliciously includes the slashing of staked assets, the forgone future income from reputational loss (an opportunity cost), and the risk of legal prosecution; the potential gain depends on the size of the manipulable assets (such as the cross-chain bridge's total value locked, TVL, or the liquidation profit that could be triggered). The system's security condition requires that the ratio of the cost of acting maliciously to the potential gain remain above a safety threshold. When this ratio declines as the protocol's TVL grows, the system must maintain its margin of safety by raising the staking requirement or introducing additional penalty mechanisms.

Specifically, a *sufficiently decentralized* architecture must meet bottom-line requirements along three dimensions. First, at the asset-custody layer, the system must guarantee that no single entity can steal a user's assets. This is the least compromisable link in the entire trust stack. However high the system's performance requirements, control over a user's assets must, and can only, be exercised by the user's own cryptographic signature. Second, at the matching and execution and the clearing and settlement layers, the system must ensure that no entity can covertly manipulate the trading rules for private gain. Although the execution at these layers may be performed by a specific validator set or high-performance nodes, the results must be fully verifiable, and any

deviation from the established rules must be detected promptly and penalized. Finally, at the governance layer, the system's rule changes must be predictable, auditable, and allow users to exit. When the protocol needs to be upgraded or key parameters adjusted, it must follow due process, granting users a full right to be informed and a time window to withdraw their assets before an adverse change takes effect.

Understanding what *sufficient* means also requires clarity about which properties are unnecessary. For example, requiring real-time whole-network consensus verification of every high-frequency order in pursuit of maximal decentralization often sacrifices the system's availability. For most trading scenarios, after-the-fact cryptographic verifiability is already *enough* to guarantee fairness. By the same token, fully decentralized permissionless governance may respond slowly to an urgent security incident, whereas a small expert committee constrained by a time lock and community veto power can often find a better balance between efficiency and security. Moreover, pursuing zero MEV may be impractical under the current public-chain architecture; as long as such value extraction is confined to a sufficiently small and controllable range, it can be regarded as an inherent friction cost of on-chain trading rather than a fundamental defect of the system.

28.7.2 The trust needs of different users

The participants in a financial market are not a homogeneous whole. Retail investors, institutional capital, and high-frequency market makers differ markedly in their core demands and their tolerance for various kinds of risk when participating in on-chain trading. Forcing a single unified architecture to satisfy everyone's needs therefore often yields a compromise that satisfies no one. Excellent architectural design should possess enough flexibility to respond to the specific trust needs of different user groups.

For retail investors, the core concern usually centers on the safety of assets and the fairness of the trading environment. What they least want to see is a repeat of an event like a centralized exchange misappropriating funds, or being subjected to *queue-jumping* arbitrage by an internal market maker inside an opaque matching engine. By contrast, retail investors are relatively insensitive to trading latency, and the difference between a few hundred milliseconds and a few microseconds is all but negligible in their subjective experience. At the same time, retail participants usually lack the time or expertise to engage deeply in a protocol's day-to-day governance. For this group, therefore, the optimal architectural configuration is often full trustlessness at the asset-custody layer, complemented by moderate trustlessness at the matching and clearing layers, while a relatively high degree of expert representation is acceptable at the governance layer. Yet there is a systematic gap between retail investors' revealed preference and their stated preference: in a bull market, retail investors systematically favor the platforms offering the highest leverage and the fastest speed, even when those platforms are conspicuously weak in architectural security. This means that an architectural security advantage may need to be realized through nudge mechanisms such as default-option design (for instance, enabling a self-custodial wallet connection by default), rather than relying on users' active, rational choice.

Institutional investors present a more complex map of trust needs. Beyond the non-

negotiable safety of assets, institutional capital is often subject to strict compliance and regulatory constraints that require the system to be highly auditable. Institutions need to be able to demonstrate clearly to auditors and regulators the flow of their assets and the compliance of their trades. This requires the system to achieve a high degree of trustlessness and full transparency at the asset-custody, matching and execution, and clearing and settlement layers. At the same time, institutional investors usually require the protocol to have a formal and predictable governance framework in order to avoid the policy risk of frequent rule changes. As for performance, although institutional investors do not demand the lowest latency as insistently as high-frequency trading desks do, they still need the system to have the throughput capacity to handle large orders and cope with extreme market swings.

The demands of high-frequency market makers lie at the other extreme. As the primary providers of market liquidity, their strategies' profitability depends heavily on extremely low latency and a certain execution speed. In the latency-sensitive domain of high-frequency trading, any additional latency introduced by a consensus mechanism can render a strategy ineffective or expose it to serious adverse-selection risk. High-frequency market makers are therefore often willing to trade a degree of trust for performance. As long as the underlying asset custody is fully secure on chain, they may even accept the use of a centralized off-chain engine for order matching. This core demand for performance is precisely the driving force behind the market's current favor toward the *off-chain matching plus on-chain settlement* hybrid architecture.

This divergence of user needs foreshadows the future direction of on-chain financial architecture. We are unlikely to see a single unified architecture suited to every scenario; rather, we will arrive at an ecosystem in which multiple architectures coexist. In this ecosystem, different protocols serve specific target user groups precisely by choosing different landing points on the trust-performance frontier.

28.7.3 A gradation of trust

Building on the definition of the *sufficient decentralization* bottom line and the analysis of different user needs, we can construct a more fine-grained framework for grading trust. This framework discards the simple dichotomy of *trust* versus *trustlessness*, dividing the trust assumptions in a system into four progressive levels and matching each level with a corresponding technical implementation path.

At the very bottom, and the most solid link, is the *zero trust* level. At this level, the system's security is guaranteed entirely by code logic and cryptographic primitives, and the user need not trust any external entity, validator set, or governance committee. This level of trust usually applies only to the asset-custody layer. Through self-custodial smart contracts deployed on a decentralized underlying network, users retain complete control over their assets. As long as the underlying consensus mechanism has not been breached, no one can transfer the assets without the user's private-key signature. This is the foundation for building the entire trustless financial system.

One level up is *verifiable trust*. At this level, some of the system's computation or execution may be performed by a specific centralized entity or a few nodes in exchange for higher processing performance. This execution privilege, however, is not unconstrained.

The executor must submit cryptographic proofs (such as zero-knowledge proofs) or publish the full transaction logs to the base chain, so that anyone can verify the correctness of the execution results after the fact. If the executor attempts to deviate from the established rules, a challenge mechanism allows honest nodes to submit a fraud proof, thereby overturning the erroneous state and imposing severe economic penalties on the malicious actor. This model is widely applied at the matching and clearing layers of layer-2 scaling solutions and app-chains, greatly raising the system's throughput while guaranteeing the correctness of results.

The third level is called *exitable trust*. This trust model applies mainly to the system's governance layer. In a complex financial protocol, wholly rigid rules often struggle to cope with a rapidly changing market environment, and a measure of flexibility is necessary. Adjustments to the protocol's parameters or upgrades to its code may be led by a centralized core team or by a DAO representing a small interest group. To prevent this power from being abused, the system must enforce a time-lock mechanism. After any critical rule change is proposed, it must pass through a mandatory waiting period before taking effect. This waiting period provides a window for users who disagree with the change to exit safely, allowing them to withdraw their assets under the old rules. Combined with an emergency-veto mechanism, the broad community can, in extreme cases, block a malicious upgrade proposal.

At the highest level is *replaceable trust*. At certain stages that must rely on external input—such as the oracle layer that obtains off-chain asset prices—the system inevitably needs to trust a data provider. To lower this single-point-of-failure risk, the architectural design should introduce diversified data sources and a seamlessly switchable substitution mechanism. The system should not be bound to a single oracle service but should compute the final result by aggregating data from multiple independent oracles. When a data source is detected to have an anomalous deviation or to have stopped serving, the system can automatically remove it and switch to a backup node. This mechanism, which constrains behavior through redundancy and competition, converts the dependence on a single entity into trust in an entire decentralized service market.

Cross-mapping these four trust levels onto the three-dimensional bottom-line requirements proposed in Section 28.7.1 yields more operational guidance for architectural design. The asset-custody layer, as the first-dimension bottom line ("no single entity can steal a user's assets"), must reach the *zero trust* level, which is the non-negotiable minimum standard. The matching and execution and the clearing and settlement layers correspond to the second-dimension bottom line ("no entity can covertly manipulate the trading rules"), and their minimum requirement is the *verifiable trust* level—execution may be performed by specific nodes, but the results must be endorsable by cryptographic proof. The governance layer corresponds to the third-dimension bottom line ("rule changes are predictable, auditable, and allow exit"), and its minimum requirement is the *exitable trust* level—even if governance power is concentrated in a few hands, users must have a time window to withdraw their assets before a rule change. The data and oracle layer should reach the *replaceable trust* level, using multi-source redundancy to ensure that the system does not depend on any single data provider. This mapping clarifies the minimum admission standard for each layer of the trust stack within the *sufficient decentralization* framework.

28.7.4 The path of progressive decentralization

Having understood the graded architecture of trust, we need to examine the dynamic process by which a system evolves from its initial state toward the ideal architecture. For the vast majority of on-chain financial projects, achieving full decentralization across all layers at the genesis stage is not only very difficult in engineering terms but often impractical. Early product iteration, bug fixing, and market expansion require the core team to have a high degree of agility and control [63]. A more pragmatic and widely adopted path is therefore progressive decentralization.

Progressive decentralization is not a vague slogan but a structured, multi-stage evolutionary strategy. In this roadmap, the core task of the first stage is to establish the basic bottom line of asset safety—that is, to achieve zero trust at the asset-custody layer. This is the minimum admission threshold for anything that calls itself a decentralized finance protocol. The team must ensure that control over the smart contracts cannot be used to misappropriate user funds at will, which usually requires multi-signature schemes, rigorous audits, and immutable custody logic. Only once asset safety is established can the system earn the initial trust of early users.

As the system enters the second stage, the focus shifts to improving the transparency and verifiability of the matching and clearing processes. At this stage, the team may still run a centralized sequencer or matching engine to ensure the user experience, but it must begin to introduce mechanisms that let these components' behavior be externally audited. This may include periodically publishing transaction logs to a data-availability layer, or beginning to develop and integrate zero-knowledge proof techniques. The goal of this stage is to eliminate the internal information black box, so that the fairness of operations no longer depends solely on the team's reputation.

Once the protocol's business model is validated and its technical architecture stabilizes, progressive decentralization enters the third stage: the devolution of governance power. This is a challenging transition period in which the core team must gradually transfer control of the protocol to a broad community of token holders. This usually begins with the adjustment of non-critical parameters, gradually extends to the allocation of treasury funds, and ultimately encompasses the upgrade authority over the core smart contracts. In this process, establishing a healthy representation system, guarding against governance attacks, and maintaining decision-making efficiency are the difficult problems the team must solve.

The final stage addresses the decentralization of external dependencies, particularly infrastructure such as oracles and cross-chain bridges. This requires the system to move from depending on a single trusted party to connecting to diversified decentralized networks—for example, replacing an internally maintained price feed with a decentralized oracle network, or upgrading a multi-signature-controlled cross-chain bridge into a trustless bridging solution based on light clients or zero-knowledge proofs.

Taking today's representative app-chain architectures as an example, many leading projects are at different stages of this progressive path. They have achieved a relatively high degree of trust minimization at the asset-custody and execution layers but still retain significant centralization residue in the allocation of governance power and the security of cross-chain bridging [45]. This state is not the architecture's endgame but

a slice of the process of evolution.

The path of progressive decentralization, however, also harbors structural risks. The greatest hidden danger is that *gradual* may degenerate into a pretext for the team to retain power indefinitely. When the core team grows accustomed to the efficiency and convenience of centralized governance, or when vested-interest groups reap rich returns under the current architecture, they often lack any intrinsic motivation to truly advance the last few steps of decentralization. Empirical observation supports this concern: among major DeFi protocols, the proportion that complete the final stage of their decentralization roadmap (such as fully removing admin keys and transferring upgrade authority to permissionless governance) is extremely low, and most protocols stall once they reach the comfort zone of *sufficient decentralization*. Beyond technical feasibility, therefore, the community's continued oversight, transparent roadmap commitments, and a hard-coded timetable for the transfer of power are all key forces for ensuring that progressive decentralization does not stray from its stated goal.

28.8 Modularity, composability, and the trust market

Let us place the choice of a layered architecture back within the macro framework for assessing market quality: as the analytical model established in Chapter 25 holds, the architectural design of any trading system must ultimately withstand examination along five dimensions—transaction costs, price discovery, resilience, fairness, and accessibility. The following analyzes how a layered architecture systematically reshapes these five dimensions, what new forms of risk it introduces, and how architectural design can, at a deeper level, substitute for traditional regulatory intervention.

28.8.1 How layered architecture reshapes the five dimensions of market quality

When we turn our attention to how a layered architecture performs in actual operation, we find that it does not provide an immediate, Pareto-optimal improvement along every dimension. On the contrary, it displays a clear trade-off: it achieves structural breakthroughs along the defensive dimensions (resilience and fairness) but must pay a certain performance cost along the offensive dimensions (transaction costs and price discovery). This trade-off is, in essence, the concrete projection of the *trust-performance frontier* discussed above onto the assessment of market quality.

Along the transaction-cost dimension, a layered architecture inevitably raises explicit costs in the short term. Empirical comparative studies of centralized and decentralized exchanges show that the fixed costs of on-chain trading (mainly gas fees) significantly burden small trades and produce a persistent arbitrage deviation between centralized and decentralized exchanges [64]. Yet if we lengthen the assessment horizon and factor in implicit costs, the conclusion reverses. In the monolithic black box architecture, users have in fact been paying a steep *trust premium* for the exchange's credit risk all along—a premium that stays hidden in calm markets but manifests, when a crisis erupts, as the complete zeroing-out of assets. The persistent arbitrage deviation between centralized and decentralized exchanges recorded in the empirical study by

Barbon and Ranaldo [64] provides quantitative corroboration of this cost structure, while the significant withdrawal premium that briefly appeared at major centralized exchanges after the FTX collapse reflects the existence of this implicit trust cost from another angle. By de-trusting the asset-custody layer and the clearing and settlement layer, a layered architecture fundamentally eliminates this implicit systemic-risk cost. Moreover, empirical data also show that, because fixed costs can be amortized over larger notional amounts, a decentralized architecture can often offer more competitive all-in transaction costs for large trades [64].

Along the price-discovery dimension, the impact of a layered architecture is relatively neutral but holds long-term upside. The efficiency of price discovery depends heavily on the symmetry of information and the depth of liquidity. Although the monolithic black box architecture can provide extremely high instantaneous liquidity through low-latency matching, its opacity often conceals the true dynamics of order flow. A layered architecture decouples the data layer from the matching layer, making all trade records and order-book states fully transparent and verifiable on chain. This transparency eliminates information asymmetry, enabling all market participants to price on the basis of the same genuine data and thereby promoting, over the long run, a more robust and fair process of price discovery.

Along the resilience dimension, a layered architecture brings a structural improvement. Research on systemic risk notes that the most dangerous vulnerability in a complex financial network stems from *the correlation of failures* [65]. In a centralized exchange, all functional modules share the same root of trust, which makes a local failure highly prone to cascading into a systemic collapse. A layered architecture severs this chain of failure propagation through modular design. Even if the matching engine suffers an outage or the oracle network briefly fails, as long as the underlying asset-custody smart contract remains intact, users' principal faces no risk of loss. This architectural property of downgrading *systemic destruction* to *localized functional degradation* endows a trading system with powerful resilience to survive extreme market turmoil. A layered architecture, however, does not eliminate systemic risk entirely; it changes the path of transmission. In a layered architecture, economic contagion may unfold along a specific cascade path: oracle failure $\rightarrow$ the liquidation engine cannot price correctly $\rightarrow$ large-scale erroneous liquidations or delayed liquidations $\rightarrow$ the insurance fund is rapidly exhausted $\rightarrow$ the auto-deleveraging mechanism forcibly harvests profitable positions $\rightarrow$ market confidence collapses $\rightarrow$ run-like withdrawals. The core value of a layered architecture is to downgrade this *instantaneous total collapse* to *gradual stress transmission*, in which each layer's fault-isolation mechanism provides a window of response time; but the ultimate elimination of systemic risk still depends on the adequacy of each layer's risk buffers.

Along the fairness dimension, a layered architecture systematically clears away the conflicts of interest inherent in a centralized exchange. We have diagnosed in detail above how a centralized exchange exploits its multiple roles as rule-maker, referee, and player to gain informational privilege. A layered architecture forcibly separates these functions: the matching logic is encoded as an immutable smart contract or the consensus of an independent app-chain, the liquidation parameters are determined by transparent on-chain governance, and the market-maker agreements are no longer a secret of

behind-the-scenes maneuvering. This architecture-level dispersion of power deprives unfair behaviors—front-running, order queue-jumping, and selective liquidation—of their technical feasibility.

Along the accessibility dimension, a layered architecture inherits the permissionless character of blockchain technology and achieves genuine global access. The traditional trust architecture depends heavily on geographic jurisdiction and licensing barriers, which objectively exclude a large number of potential market participants. A layered architecture establishes trust through code rather than institutions, allowing any user with an internet connection to access the core financial infrastructure on equal terms. Although the current on-chain interaction interface still poses a certain threshold in user experience, in terms of the distribution of access rights a layered architecture has unquestionably expanded the boundaries of the financial market enormously.

Table 28-8 summarizes the five dimensions above, presenting the structural reshaping of market quality by a layered architecture. The *structural impact* column describes the effects already observable at the current stage, while the *long-term evolutionary trend* column makes forward-looking judgments based on the development path of the underlying technology. Reading the two columns against each other helps distinguish the efficiency costs a layered architecture must bear in the short term from the structural gains it can deliver over the long run.

Market-quality dimension	Structural impact of layered architecture	Long-term evolutionary trend
Transaction costs	Explicit execution costs rise slightly (constrained by on-chain verification overhead), but the implicit systemic trust premium is greatly reduced	As modular blockchain scaling technology matures, all-in transaction costs will keep falling
Price discovery	Instantaneous pricing efficiency is slightly affected by latency, but on-chain data transparency fundamentally improves information symmetry	Trends over the long run toward more genuine, more robust price discovery, reducing the room for manipulation
Resilience	Markedly improved; module isolation eliminates single points of failure, downgrading the risk of total collapse to localized failure	As decentralized sequencers and fraud proofs become widespread, the system's stress resistance will strengthen further
Fairness	Markedly improved; the architecture level eliminates informational privilege and conflicts of interest, and the matching and clearing rules are fully transparent and verifiable	Trends markedly toward higher fairness, with any rule change required to pass through a public governance process
Accessibility	Markedly improved; permissionless global access breaks the geographic and licensing barriers of traditional finance	As technologies such as account abstraction mature, the interaction threshold will fall substantially, achieving financial inclusion

Table 28-8. The structural impact of layered architecture on the five dimensions of market quality

(Data source: compiled by the author)

A core judgment can be distilled from the horizontal comparison in Table 28-8: a layered architecture achieves structural breakthroughs along the three *defensive* dimensions of resilience, fairness, and accessibility, and these improvements stem from a fundamental restructuring of trust assumptions at the architectural level rather than marginal parameter optimization. Along the two *offensive* dimensions of transaction costs and price discovery, a layered architecture must bear the friction costs of on-chain verification in the short term, but as modular scaling technology matures, these costs keep falling. Overall, a layered architecture trades short-term explicit efficiency losses for long-term systemic robustness—a trade-off repeatedly validated in the historical experience of traditional finance's introduction of the CCP and independent custody.

28.8.2 New forms of risk in a layered architecture

Although a layered architecture excels at eliminating the single points of failure of a centralized entity, we must recognize clearly that this paradigm shift does not abolish risk out of thin air; rather, it transfers risk from *trust in entities* to *trust in code and protocol interfaces*. This transfer introduces three entirely new forms of systemic risk, which constitute the core problems that on-chain financial engineering must solve in the future.

Inter-module interface risk is a relatively hidden point of fragility in a layered architecture. In a monolithic architecture, the functional modules run within the same controlled memory space, and data transfer and state updates are synchronous and atomic. In a layered architecture, however, the functions of different layers are executed by independent smart contracts or app-chain networks, and communication among them must rely on asynchronous message-passing protocols. This cross-layer communication introduces the possibility of time lags and state inconsistency, thereby creating a new attack surface. For example, if the price-update interface between the oracle layer (Layer 4) and the clearing layer (Layer 3) has a latency vulnerability, an attacker can exploit this time lag to conduct a flash-loan attack, manipulating the price to trigger liquidations that should not occur. Several DeFi security incidents show that oracle manipulation has become a major cross-layer attack vector. In 2022 alone, DeFi protocols lost more than $400 million to 41 oracle-manipulation attacks [66]. This requires architects, when designing inter-layer interfaces, to introduce multi-signature verification, the time-weighted average price (TWAP), and circuit breakers to isolate the chain reactions caused by a single module's failure.

Second is cross-chain bridging risk, the core challenge that a layered architecture faces in today's multichain ecosystem. When a trading system's asset-custody layer (Layer 1) resides on the Ethereum mainnet while its matching and execution layer (Layer 2) resides on an independent rollup or app-chain, the transfer of assets between the two layers must rely on a cross-chain bridge. A cross-chain bridge is, in essence, a complex consensus converter that must lock assets on one end and mint equivalent mapped tokens on the other. If the bridging protocol's smart contract has a vulnerability, or its validator network is compromised, an attacker can drain the reserve funds inside the bridge directly without breaching the underlying asset-custody contract [67][68].

This risk shows that the security of a layered architecture is determined by its weakest connection point. Future architectural evolution must strive to minimize dependence on third-party bridges and adopt instead trustless cross-chain communication protocols based on zero-knowledge proofs, so as to converge this attack surface fundamentally.

Finally, there is the endogenous risk that system complexity brings. The essence of a layered architecture is to decompose a complex system into multiple subsystems, but this inevitably increases the overall topological complexity of the system. Every newly added module and every new communication link means an increase in potential points of failure. This practice of *introducing complexity for the sake of security* is a perennial paradox in engineering. A highly complex system is often hard to subject to exhaustive formal verification, and its emergent behavior under extreme market conditions becomes harder to predict. In pursuing layered decoupling, therefore, architects must maintain constant restraint over complexity, following the principles of *least privilege* and *minimal design* to ensure that the logic of each independent module is sufficiently clear and its boundaries sufficiently well defined.

Although a layered architecture introduces the new risks above, a simplified comparison of the net risk effect helps assess the direction in which overall security changes. The risk that a layered architecture eliminates—the systemic default of a centralized entity, epitomized by the FTX collapse (a gap of about $8 billion in customer assets at the time of collapse)—far exceeds, in the historical magnitude of losses, the new risks a layered architecture introduces. To date, although the cumulative losses from cross-chain bridge attacks and oracle-manipulation incidents have reached the level of billions of dollars, they are distributed across hundreds of independent events, and the systemic impact of any single event is far smaller than the total collapse of a monolithic black box. This shift in risk profile—from *low-frequency, catastrophic-consequence* to *higher-frequency, manageable-consequence*—is the core value proposition of a layered architecture along the dimension of risk management.

28.8.3 Architecture as governance

When we examine the evolutionary history of traditional finance, we find that every expansion of regulators' power was often meant to patch an inherent defect in the underlying architecture. Because the monolithic black box cannot prove its own innocence, auditors are needed; because intermediaries have an incentive to misappropriate funds, custody licenses are needed; because internal matching may involve self-dealing, compliance checks are needed. A layered architecture, however, offers an entirely new line of thinking: eliminating, through technical means, the very environment that generates the incentive to break the rules, thereby rendering part of external regulation unnecessary. This idea can be summed up as *architecture as governance*. The concept is of a piece with the idea of *code is law* proposed by Lawrence Lessig in his seminal work. Lessig argued that in digital space, software architecture, like the law of the physical world, is one of the four forces that constrain behavior [69]. De Filippi and Wright extended this idea further into the blockchain domain, proposing the concept of *lex cryptographica*—the technical embedding of legal rules through self-executing smart contracts [70]. A layered architecture advances Lessig's theory from abstract norma-

tive discussion to the concrete engineering practice of financial infrastructure: it not only uses code as a means of constraint but, through modular layered design, precisely allocates different types of constraint (cryptographic enforcement, economic incentives, and governance procedures) to different architectural layers.

With respect to asset segregation, the traditional regulatory framework treats the physical segregation of customer funds from proprietary funds as a fundamental bottom line. The FTX collapse, however, demonstrated the notable fragility of institutional segregation that relies on audit reports and executive promises. In a layered architecture, the asset-custody layer is devolved to a permissionless underlying blockchain, and open-source smart contracts enforce strict account logic. This means that, absent authorization by the user's private key, no exchange operator can physically touch these assets. This cryptography-based mandatory segregation is more resistant to attack than any legal provision. Regulators no longer need to expend vast resources verifying an exchange's bank accounts, because the architecture itself already provides sufficient proof of asset safety.

With respect to information disclosure, traditional regulation treats it as a core means of protecting investors, requiring financial institutions to publish voluminous financial statements and trading data periodically. These reports, however, are often subject to lag and easily window-dressed. A layered architecture moves all data of the matching layer and the clearing layer on chain, so that the bookkeeping of every order, the triggering conditions of every liquidation, and the flow of every cent are exposed in real time to the oversight of the whole network's nodes. This natural, irrevocable transparency turns information disclosure from *a regulatory-compliance cost* into *a fundamental property of the system's operation*. Under such an architecture, on-chain transparency gives anyone the technical prerequisite to become an auditor. Transparency, however, is a necessary but not a sufficient condition for effective oversight: interpreting raw on-chain data requires professional analytical ability, and the vast majority of users lack the technical literacy to identify anomalous patterns among massive volumes of trade records. Effective community oversight therefore also requires accompanying incentive mechanisms (such as bug-bounty programs and audit contests) to attract professional security researchers to invest their attention continuously.

Architecture as governance does not mean the complete demise of regulation but a fundamental shift in the regulatory paradigm. As the relevant research notes, modular architecture profoundly affects the effectiveness of regulation [71]. Future regulation should no longer fixate on layering endless rule patches on top of a bad architecture but should turn toward *requiring and verifying good architecture*. The regulator's role will shift from *behavior supervisor* to *architecture certifier*, whose core task is to ensure that a trading system follows the principle of layered decoupling from the outset, to ensure the security of the inter-layer interfaces, and to ensure the degree of decentralization of the governance mechanism. This paradigm shift, however, faces real obstacles: the current regulatory framework takes legal entities, not technical architectures, as the object of regulation; the transition to architectural regulation requires regulators to substantially update their technical-assessment capabilities and to develop new legal instruments suited to decentralized protocols, a process that may take years or longer. This deep integration of technology and regulation will pave the way for a

next-generation financial system that is at once trustless and compliant.

28.8.4 The trust market and composability

The ultimate form of a layered architecture is not merely to decompose the monolithic black box into internal modules but to open these modules to the entire ecosystem, thereby giving rise to a highly competitive *trust market*. In this market, trust is no longer a scarce resource monopolized by some centralized institution but a standardized component that can be priced, traded, and freely combined.

Composability is the core driving force of this vision. Once functions such as asset custody, matching and execution, clearing and settlement, and data oracles are encapsulated as standardized on-chain modules, they become financial primitives that can be freely assembled like standardized components. An emerging decentralized derivatives platform no longer needs to build a complete trust stack from scratch. It can choose to plug directly into the Ethereum mainnet as its asset-custody layer to obtain the highest grade of security, adopt a dedicated app-chain as its matching and execution layer to achieve low-latency trading, and integrate a decentralized oracle network to provide price data. The vision of composability, however, faces significant limitations at the execution level. When different functional modules are distributed across different chains or protocols, cross-module atomic operations (such as simultaneously adjusting the margin of multiple related positions) lose their guarantee of transactional consistency. The complexity of cross-chain margin management (involving the confirmation times of different chains, bridging delay, and state synchronization) makes a unified risk view hard to maintain in real time. Under current technical conditions, therefore, composability is better suited to low-frequency settlement-layer operations (such as cross-chain asset transfers and final clearing) than to high-frequency trading-execution-layer operations.

This free combination of modules will drive the rapid diversification of trading-system architectures. Different project teams can seek different balance points on the trust-performance frontier according to the specific needs of their target users, thereby designing a diversity of customized architectures. More importantly, composability breaks the siloing effect of traditional financial infrastructure. An excellent clearing module can serve multiple different matching engines at once, and a deep liquidity pool can be shared by multiple front-end interfaces. These network effects will greatly lower the barrier to financial innovation and push the on-chain financial system to evolve toward greater specialization and refinement.

In the trust market, the providers of the various modules face full and open competition. The modules that offer higher security, lower latency, or superior governance mechanisms will win adoption by more protocols and higher economic returns. This market-based competition grounded in code and consensus will become the main driving force pushing the on-chain trust infrastructure to evolve continuously. In the end, we will see a global financial-infrastructure network woven from decentralized modules—one with a high degree of survivability and a capacity for self-healing.

28.9 Chapter summary

Starting from the systemic architectural failure of the FTX collapse, this chapter proposed the trust stack method of analysis, decomposing the implicit trust assumptions in a trading system into five layers—asset custody, matching and execution, clearing and settlement, data and oracles, and governance—and, through the trust concentration metric, turning the problem of trust into a quantifiable engineering design variable. On this basis, the characterization of the trust-performance frontier revealed a fundamental engineering constraint: trustless verification and high-performance execution cannot both be optimized within the same architectural layer. Layered architecture is precisely the engineering response to this constraint: by independently choosing the balance point between trust and performance at each layer, the system can reserve sufficient performance headroom for the matching and clearing stages while making no compromise on asset safety.

The institutional evolution of traditional finance has validated the effectiveness of functional separation, and on-chain systems are now realizing a technical upgrade of the same functions through cryptography and consensus mechanisms. The app-chain practice epitomized by Hyperliquid shows that this layered migration is engineering-feasible, but the size of the validator set, the security of cross-chain bridging, and the concentration of governance power still constitute residual trust in the current architecture that cannot be ignored. The judgment framework of *sufficient decentralization* decomposes trust needs into four progressive levels—zero trust, verifiable trust, exitable trust, and replaceable trust—and establishes a clear mapping to the layers of the trust stack, providing a structured methodology for planning the path of progressive decentralization.

The trust stack model and the layered design principles established in this chapter will serve, in the chapters that follow, as a unified analytical tool for assessing the architectures of specific protocols. The next chapter focuses on the risk-management mechanisms of perpetual futures protocols, applying this chapter's layered framework to audit, layer by layer, how key components—liquidation logic, the operation of the insurance fund, and oracle dependence—differ in their implementation and risk profile across architectures.

References

[1] Wikipedia. (n.d.). *Bankruptcy of FTX*. Retrieved from https://en.wikipedia.org/wiki/Bankruptcy_of_FTX

[2] Investopedia. (n.d.). *FTX crypto exchange collapse: Causes, consequences, and lessons*. Retrieved from https://www.investopedia.com/what-went-wrong-with-ftx-6828447

[3] CNBC. (2022, November 13). *Sam Bankman-Fried's Alameda quietly used FTX customer funds without raising alarm bells, say sources*. https://www.cnbc.com/2022/11/13/sam-bankman-frieds-alameda-quietly-used-ftx-customer-funds-without-raising-alarm-bells-say-sources.html

[4] Commodity Futures Trading Commission. (2024, August 8). *CFTC obtains $12.7*

billion judgment against FTX and Alameda. https://www.cftc.gov/PressRoom/PressReleases/8938-24

[5] Federal Reserve Bank of Chicago. (n.d.). *Understanding derivatives: Central counterparty clearing*. https://www.chicagofed.org/-/media/publications/understanding-derivatives/understanding-derivatives-chapter-2-central-counterparty-clearing-pdf.pdf

[6] Hyperliquid. (n.d.). *About Hyperliquid*. Hyperliquid Docs. https://hyperliquid.gitbook.io/hyperliquid-docs

[7] U.S. Securities and Exchange Commission. (2025). *Custody rule modernization: A model framework for crypto asset safeguarding*. https://www.sec.gov/files/custody-rule-modernization-model-framework-121925.pdf

[8] Roy, D., Dubey, A., & Tiwary, D. (2024). Conceptualizing an institutional framework to mitigate crypto-assets' operational risk. *Journal of Risk and Financial Management*, *17*(12), 550. https://www.mdpi.com/1911-8074/17/12/550

[9] Vidal-Tomás, D. (2025). Centralized exchanges & proof-of-solvency: The guardians of trust. *Journal of International Financial Markets, Institutions and Money*, *103*, 102183. https://doi.org/10.1016/j.intfin.2025.102183

[10] Binance. (2017). *Binance exchange whitepaper* (V1.1). https://www.scribd.com/document/369374867/Binance-WhitePaper-En

[11] Hyperliquid. (n.d.). *HyperCore overview*. Hyperliquid Docs. https://hyperliquid.gitbook.io/hyperliquid-docs/hypercore/overview

[12] Ethereum Foundation. (n.d.). *Blocks*. Ethereum.org. Retrieved March 2026, from https://ethereum.org/en/developers/docs/blocks/

[13] Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., & Juels, A. (2020). Flash boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. In *2020 IEEE Symposium on Security and Privacy (SP)* (pp. 910–927). https://doi.org/10.1109/SP40000.2020.00040

[14] DTCC. (2023, May 9). *Celebrating our 50th anniversary*. DTCC Connection. https://www.dtcc.com/dtcc-connection/articles/2023/may/09/celebrating-dtccs-50th-anniversary

[15] Ripatti, K. (2004). *Central counterparty clearing: Constructing a framework for evaluation of risks and benefits*. Bank of Finland. https://www.econstor.eu/handle/10419/211997

[16] Intercontinental Exchange. (n.d.). *Central clearing: Reducing systemic risk*. https://www.ice.com/publicdocs/Central_Clearing_Reducing_Systemic_Risk.pdf

[17] Paddrik, M., & Young, H. P. (2020). *Central counterparty default waterfalls and systemic loss* (OFR Working Paper No. 20-04). Office of Financial Research. https://www.financialresearch.gov/working-papers/files/OFRwp-20-04_central-counterparty-default-waterfalls-and-systemic-loss.pdf

[18] Berndsen, R. (2021). Fundamental questions on central counterparties: A review of the literature. *Journal of Futures Markets*, *41*(12), 2009–2022. https://doi.org/10.1

002/fut.22260

[19] Dukich, R. (2022). *Liquidity management in central clearing: How the default waterfall can be improved*. NYU Stern. https://www.stern.nyu.edu/sites/default/files/assets/documents/Dukich_Liquidity%20Management%20in%20Central%20Clearing%3B%20How%20the%20Default%20Waterfall%20Can%20Be%20Improved.pdf

[20] Heilbron, J. (2024). *Central clearing and trade cancellation: The case of LME nickel contracts on March 8, 2022* (OFR Working Paper No. 24-09). Office of Financial Research, U.S. Department of the Treasury. https://www.financialresearch.gov/working-papers/files/OFRwp-24-09_central-clearing-and-trade-cancellation.pdf

[21] SEC Historical Society. (2010). *An introduction to DTCC*. https://www.sechistorical.org/collection/papers/2010/2010_0701_DTCCServices.pdf

[22] DTCC. (2024). *2024 by the numbers: The value DTCC delivers*. DTCC Annual Report. https://www.dtcc.com/annuals/2024/value/

[23] DTCC. (n.d.). *The Depository Trust Company — DTC*. https://www.dtcc.com/about/businesses-and-subsidiaries/dtc

[24] Commodity Futures Trading Commission. (2026). *The next phase of Project Crypto: Unleashing innovation*. https://www.cftc.gov/PressRoom/SpeechesTestimony/opaselig1

[25] Bank & Finance Consulting Group. (2026). *Designing financial ecosystems: A five-layer architecture for resilience, performance, and public value*. https://bankandfinance.net/designing-financial-ecosystems-a-five-layer-architecture-for-resilience-performance-and-public-value/

[26] Parry, C. (2024). Digital asset custody deciphered: A primer to navigating the challenges of safeguarding digital assets. *Journal of Securities Operations & Custody*, *16*(2), 1–15. https://www.ingentaconnect.com/content/hsp/jsoc/2024/00000016/00000002/art00002 https://doi.org/10.69554/fekb1830

[27] Van Loon v. Department of the Treasury, No. 23-50669 (5th Cir. Nov. 26, 2024). U.S. Court of Appeals for the Fifth Circuit. https://www.ca5.uscourts.gov/opinions/pub/23/23-50669-CV0.pdf

[28] Mavroudis, V., & Melton, H. (2019). Libra: Fair order-matching for electronic financial exchanges. In *Proceedings of the 1st ACM Conference on Advances in Financial Technologies* (pp. 114–128). https://doi.org/10.1145/3318041.3355468

[29] Ram, R. (2025). *The emerging architecture of clearing and settlement in digital finance* (SSRN Working Paper No. 5378367). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5378367 https://doi.org/10.2139/ssrn.5378367

[30] Chainlink. (2025). *The blockchain oracle problem*. Chainlink Education Hub. https://chain.link/education-hub/oracle-problem

[31] Dotan, M., Yaish, A., Yin, H. C., & Tsytkin, E. (2023). The vulnerable nature of decentralized governance in DeFi. In *Proceedings of the 2023 ACM Workshop on Decentralized Finance* (pp. 1–10). https://doi.org/10.1145/3605768.3623539

[32] Commodity Futures Trading Commission. (2022, September 22). *CFTC imposes*

$250,000 penalty against bZeroX, LLC and its successor Ooki DAO. https://www.cftc.gov/PressRoom/PressReleases/8590-22

[33] Wu, X., Deng, W., Quan, Y., & Zhang, L. (2024). Trust dynamics and market behavior in cryptocurrency: A comparative study of centralized and decentralized exchanges. *arXiv preprint arXiv:2404.17227*. https://doi.org/10.48550/arXiv.2404.17227

[34] CoinTracker. (2025). *What is proof of reserves? How crypto exchanges verify asset holdings*. https://www.cointracker.io/learn/proof-of-reserves

[35] Fidelity Digital Assets. (2026). *Navigating digital assets market structure in the UK and Europe*. https://www.fidelitydigitalassets.com/research-and-insights/navigating-digital-assets-market-structure-uk-and-europe

[36] Chainlink. (2025). *Seven key cross-chain bridge vulnerabilities explained*. Chainlink Education Hub. https://chain.link/education-hub/cross-chain-bridge-vulnerabilities

[37] Outlook India. (2026). *Hyperliquid explained: How an app-chain DEX redefines on-chain trading*. https://www.outlookindia.com/xhub/blockchain-insights/hyperliquid-explained-how-an-app-chain-dex-redefines-on-chain-trading

[38] CoinDesk. (2025, March 26). *Hyperliquid's JELLY crisis exposes risks in on-chain price oracles*. https://www.coindesk.com/markets/2025/03/26/hyperliquid-jelly-short-squeeze/

[39] ChainUp. (2026). *Crypto custody guide: MPC, multi-sig & self-custody*. https://www.chainup.com/blog/digital-asset-custody-guide-mpc-vs-multisig/

[40] Nansen. (2024). *Exploring Hyperliquid: Innovation in the perp DEX landscape*. https://research.nansen.ai/articles/exploring-hyperliquid-innovation-in-the-perp-dex-landscape

[41] Xverse. (2025). *What is Hyperliquid? A guide to the high-performance Layer 1*. https://www.xverse.app/blog/what-is-hyperliquid-blockchain

[42] Rock'n'Block. (2025). *How Hyperliquid works: Architecture, order book, HyperEVM*. https://rocknblock.io/blog/how-does-hyperliquid-work-a-technical-deep-dive

[43] Yin, M., Malkhi, D., Reiter, M. K., Gueta, G. G., & Abraham, I. (2019). HotStuff: BFT consensus with linearity and responsiveness. In *Proceedings of the 2019 ACM Symposium on Principles of Distributed Computing (PODC '19)* (pp. 347–356). https://doi.org/10.1145/3293611.3331591

[44] Hyperliquid. (n.d.). *HyperBFT*. Hyperliquid Wiki. Retrieved March 2026, from https://hyperliquid-co.gitbook.io/wiki/architecture/hyperbft

[45] OneKey. (2026, January 26). *Hyperliquid architecture explained: Security & decentralization*. https://onekey.so/blog/ecosystem/hyperliquid-architecture-explained-security-decentralization-19137c

[46] gwrx2005. (2025). *Cross-chain transfer mechanism in Hyperliquid: Architecture and proxy design*. Medium. https://medium.com/@gwrx2005/cross-chain-transfer-mechanism-in-hyperliquid-architecture-and-proxy-design-b3179e636ec7

[47] Hyperliquid. (n.d.). *Bridge*. Hyperliquid Docs. https://hyperliquid.gitbook.io/hyperliquid-docs/hypercore/bridge

[48] ByteTree. (2025). *Hyperliquid: Driven by HYPE, backed by breakthroughs*. https://www.bytetree.com/research/2025/06/hyperliquid-driven-by-hype-backed-by-breakthroughs

[49] Dwellir. (2025). *Building a real-time Hyperliquid liquidation tracker*. https://www.dwellir.com/blog/building-real-time-hyperliquid-liquidation-tracker

[50] Phemex. (2025). *Hyperliquid faces scrutiny over centralized governance model*. https://phemex.com/news/article/hyperliquid-faces-scrutiny-over-centralized-governance-model-27774

[51] Hyperliquid Hub. (2026). *HyperCore throughput update* [Post]. X. https://x.com/Hyperliquid_Hub/status/2021514450042515459

[52] Token Terminal. (2026). *Hyperliquid transactions per second*. https://tokenterminal.com/explorer/projects/hyperliquid/metrics/transactions-per-second

[53] The Block. (2025). *Hyperliquid responds to community concerns over validator issues*. https://www.theblock.co/post/333559/hyperliquid-responds-to-community-concerns-over-validator-issues

[54] Luganodes. (2025). *Hyperliquid: Redefining decentralized trading*. https://www.luganodes.com/blog/hyperliquid-redefining-dex-trading

[55] The Block. (2024, March 28). *Ethereum's 13,900 nodes are running 1 million validators, with 26% of the supply staked*. https://www.theblock.co/post/285262/ethereum-one-million-validators

[56] Ainvest. (2025). *Hyperliquid's centralized governance: A double-edged sword*. https://www.ainvest.com/news/hyperliquid-centralized-governance-double-edged-sword-defi-scalability-2510

[57] gwrx2005. (2025). *Technical architecture comparison: Hyperliquid, dYdX, and Lighter*. Medium. https://medium.com/@gwrx2005/technical-architecture-comparison-hyperliquid-dydx-and-lighter-xyz-2fd005854a7e

[58] TradingKey. (2026). *The Hyperliquid phenomenon: HYPE ecosystem's path to growth*. https://www.tradingkey.com/analysis/cryptocurrencies/more/261633931-hyperliquid-exchange-hype-dex-price-tradingkey

[59] Yahoo Finance. (2026). *Hyperliquid celebrates 4x user growth in 2025*. https://finance.yahoo.com/news/hyperliquid-celebrates-4x-user-growth-112251312.html

[60] CoinDesk. (2026, March 2). *Inside HYPE's bear market resilience*. https://www.coindesk.com/business/2026/03/02/weekend-warriors-how-hyperliquid-became-retails-bear-market-playground

[61] EigenLayer Team. (2023). *EigenLayer: The restaking collective* [Whitepaper]. https://docs.eigenlayer.xyz/

[62] Ritter, V. (2022, November 13). *Minimum viable decentralization (MvD): A framework for evaluating decentralized finance (DeFi) protocols*. BICon. https://bicon.li/en/2022/11/13/minimum-viable-decentralization/

[63] Esber, J., & Kominers, S. D. (2023, January 12). *Progressive decentralization: A high-level framework*. a16z crypto. https://a16zcrypto.com/posts/article/progressive-

decentralization-a-high-level-framework/

[64] Barbon, A., & Ranaldo, A. (2024). On the quality of cryptocurrency markets: Centralized versus decentralized exchanges. *arXiv preprint arXiv:2112.07386*. https://doi.org/10.48550/arXiv.2112.07386

[65] Gambhir, A. (2025). A systemic risk assessment methodological framework for interconnected infrastructure systems. *Nature Communications*. https://doi.org/10.1038/s41467-025-62029-w

[66] Chainalysis. (2023, March 7). *Oracle manipulation attacks rising: A unique concern for DeFi*. Chainalysis Blog. https://www.chainalysis.com/blog/oracle-manipulation-attacks-rising/

[67] Zhao, Q., Wang, Y., Yang, B., Shang, K., & Sun, M. (2023). A comprehensive overview of security vulnerability penetration methods in blockchain cross-chain bridges. *Authorea Preprints*. https://doi.org/10.22541/au.169760541.13864334

[68] Islam, M. J., Islam, S., Hossain, M., Noor, S., & Islam, S. M. R. (2025). Securing blockchain systems: A layer-oriented survey of threats, vulnerability taxonomy, and detection methods. *Future Internet*, *17*(5), 205. https://doi.org/10.3390/fi17050205

[69] Lessig, L. (2006). *Code: Version 2.0*. Basic Books. https://lessig.org/product/code/

[70] De Filippi, P., & Wright, A. (2018). *Blockchain and the law: The rule of code*. Harvard University Press. https://doi.org/10.4159/9780674985933

[71] Henningsson, S., & Eaton, B. D. (2023). Governmental regulation and digital infrastructure innovation: The mediating role of modular architecture. *Journal of Information Technology*. https://doi.org/10.1177/02683962221114429

Designing an Ideal Decentralized Clearing and Settlement Architecture

The first 28 chapters of this book have analyzed, one by one, the individual components of a decentralized clearinghouse and the ways in which each can fail: how novation absorbs bilateral credit risk through an impersonal central counterparty (Chapters 5 and 28); how multilateral netting compresses a large volume of bilateral exposures into net positions (Chapters 1 and 5); how the mark price becomes the measurement basis for unrealized profit and loss and for margin (Chapter 14); how margin and leverage define an account's liquidation boundary (Chapter 11); how the reflexivity of liquidation amplifies individual prudence into systemic fragility (Chapter 11); how the default waterfall absorbs losses layer by layer after a shortfall (Chapters 5, 11, and 28); and how insurance funds and auto-deleveraging (ADL) socialize the tail losses that cannot be eliminated (Chapter 12). Yet clarifying each component in isolation is not the same as knowing how to integrate them, absent a traditional central institution, into a whole that can be trusted.

This integration problem is not abstract. In October 2025, a liquidation cascade triggered the liquidation of approximately $19 billion in leveraged positions within a single day [1]. The mechanics of that cascade were analyzed in full in Chapter 11 and are not restated here; it is cited here to make one fact plain: when the components above are integrated into a single high-leverage, around-the-clock, unattended system, the transmission among those components can itself constitute system-level risk. It is not this chapter's argument but its backdrop.

Chapter 28 closed by establishing the identity of the clearing and settlement layer: the layer in an on-chain system that corresponds most closely to the functions of a traditional central counterparty, whose role is clearing (recomputation of the books) and settlement (commitment of account state), rather than mere default management (Section 28.4.3). That chapter also left three constructive questions to the present one: can an on-chain system achieve a default waterfall as deep as that of a traditional central counterparty? How should risk control be balanced against user experience? And how can incentives be aligned in a system in which no single actor can issue commands (Section 28.3.6)? Traditional finance took more than two centuries, driven by successive crises, to arrive at today's clearinghouse; the tiered structure of its default

waterfall is the product of that evolution [2]. What the first 28 chapters provide this chapter is an inventory of components and a catalog of failures; they have not yet given a scheme for integrating those components into a whole, nor answered the three questions that integration necessarily raises.

These three questions must wait for the integration stage because none can be resolved at the level of a single component. Depth asks not whether any one buffer layer is thick enough, but in what order the layers of capital are arranged and who bears losses first. Degradation asks not how fast or slow any single liquidation is, but whether the system can maintain order when selling pressure, liquidity, and confidence contract at once. Incentive alignment asks not whether some parameter is appropriate, but who bears the ultimate loss when all parameters operate as designed and yet some participants rationally choose not to act. All three point to the coordination among components rather than to the performance of the components themselves, and they can therefore be answered only after those components have been integrated into a whole and their interactions made visible. This is why the chapter first carries out the system integration and only then addresses the three questions.

On this basis, the chapter undertakes two tasks: integration and answering the questions. Integration means assembling, for the first time, the components dispersed throughout the book into a complete decentralized clearinghouse, and showing that the system's true fragility lies not in the components themselves but in the seams between them. Answering the questions means addressing, one by one, along the structure of this system, the three questions Chapter 28 left open. Before that, four premises must be made explicit. First, the fragility of a clearinghouse system lies mainly not in any single component but in the seams between components; each component may be well designed, and what breaks the system is usually the transmission between them. Second, two layers of the ideal decentralized clearinghouse remain absent on-chain today: the skin-in-the-game first-loss layer of the operator's own equity, and the layer of prefunded mutualization among members. These two layers are the source of a traditional default waterfall's depth (Section 28.3.6). Third, there is a class of extreme scenarios that no purely algorithmic system can fully cover: absent a lender of last resort, the total losses may exceed all the buffers the system can currently draw upon, and full loss coverage is no longer an achievable goal [3]. Fourth, there is a further boundary that no design of market incentives can cross: there will always be moments when deterministic code must allow human discretion to intervene. These four points are not pessimistic concessions but the premises of the chapter's entire argument: only by first clarifying where integration is impossible, where losses cannot be fully covered, and where incentives fail can we discuss how to integrate, to the greatest extent possible, the parts that can be integrated.

29.1 The overall architecture of a decentralized clearinghouse

29.1.1 A single deterministic state machine

A clearinghouse system is not a simple superposition of several risk mechanisms but a single, deterministic state machine. Each component dissected separately above is, within this system, merely one stage; its value lies not in operating independently but in being incorporated into an uninterruptible chain of state transitions. Describing the margin-calculation module, netting, marking to market, the insurance fund, and auto-deleveraging separately does not automatically yield a clearinghouse system, just as analyzing each component in isolation does not yield a functioning whole. The components also require a determinate set of connections: the output of the previous stage must be the sole, unambiguous input of the next; wherever a connection is mismatched, the entire system's state diverges. Understanding a decentralized clearinghouse should begin with understanding this chain of state transitions, not with enumerating its components.

To understand the clearinghouse as a state machine is not a convenient metaphor but a definition with substantive consequences. The core promise of a state machine is determinism: given the same initial state and the same sequence of inputs, it necessarily evolves to the same final state, with no degree of freedom that depends on timing, on the executor, or on off-chain negotiation. This is the classic abstraction that distributed systems have used for decades to build fault-tolerant services: a service is expressed as a deterministic state machine so that a set of mutually distrusting nodes, each of which may crash or even submit malicious state, can agree on a single input sequence and thereby present themselves to the outside world as a single, never-failing whole [4]. Byzantine fault-tolerant (BFT) consensus extends this abstraction further, to environments in which nodes may exhibit Byzantine behavior, so that the consistency of the state machine holds even when some nodes deliberately submit false information [5]. A decentralized clearinghouse system is, in essence, this service placed within that abstract framework: its task is to let multiple validators who need not trust one another reach an unambiguous agreement on what each account's state should be at this instant.

Applying this abstraction to a clearinghouse requires distinguishing one level, without which the analytical boundary of the term *state machine* becomes unclear. Determinism is one problem; replicated determinism is another, and between them lies the entire decentralization content of the system. A traditional clearinghouse's computation module can, in principle, also be deterministic: given the same inputs, it too will compute the same books. But its determinism holds only within its own internal copies of the system; external participants can neither obtain that input sequence nor hold that rule set, still less recompute the books to check them. What makes state machine replication truly demanding is not that each node be individually deterministic, but that multiple mutually distrusting nodes, any of which may crash or be compromised, evolve from the same input sequence, under the same rules, to the same state, and that this agreement hold even when some of those nodes fail or submit false information [4]. In

other words, the difficulty of a replicated state machine is not merely that the computed result is correct, but that, absent a trusted arbiter, no participant can raise conflicting valid claims about the state. Byzantine fault-tolerant consensus strengthens this condition further: it requires that even when some nodes deliberately submit forged state, honest nodes can identify and exclude it, so that the replicated state remains unique despite the presence of malicious nodes [5]. A decentralized clearinghouse system is decentralized because of this layer of replication: it requires not merely that a single computed result be accurate, but that no one can privately tamper with that result and that everyone can independently verify it has not been tampered with.

It is in this sense that the difference between a decentralized clearinghouse and a traditional central counterparty is not merely the removal of an intermediary. A traditional central counterparty does not satisfy the determinism promise of a state machine: its risk-control model is confidential, its stress-test parameters are not disclosed, and it retains room for discretion at extreme moments. Chapter 28 used the March 2022 nickel-contract crisis at the London Metal Exchange to illustrate how such discretion conflicts with the definition of neutral infrastructure (Section 28.7). A central counterparty is still trusted not because it can be recomputed but because its authority has another source: ample proprietary capital as an underlying backstop, a complete legal recourse regime as institutional guarantee, and strict regulation under continuous scrutiny [6]. Capital absorbs the consequences of error, law provides recourse against breaches of trust, and regulation imposes external constraint on internal loss of control; together these three constitute a form of ex post insurance about responsibility, not ex ante verification of process. An ordinary user has never checked, and could never check, whether a central counterparty's margin model is internally consistent or whether its stress-test parameters are prudent; what the user trusts is never the computation module itself but this institutional framework surrounding it. The trustworthiness of a traditional clearinghouse rests on the three institutional foundations of capital, law, and regulation, not on anyone's ability to recompute the books; the absence of verifiability is what these three institutional foundations compensate for. The central-counterparty institution is itself a transformation of counterparty risk from dispersed bilateral trust into trust in a single regulated professional institution—a transformation of trust, not its elimination [7].

A decentralized clearinghouse loses these three institutional foundations. It cannot appeal to capital, because it has no central-bank-scale proprietary capital; it cannot appeal to law, because its counterparty is borderless code; it cannot appeal to regulation, because it runs on a network that no one can unilaterally halt. What it can appeal to is chiefly determinism itself: expressing the clearinghouse in full as a state machine so that every step of its evolution is uniquely determined by public rules, and so that any node can, from the same inputs, independently recompute the same outputs. Hyperliquid's design is the most complete implementation of this approach to date: it unifies matching, liquidation, and settlement within an on-chain state machine called Hypercore, with HyperBFT consensus reaching agreement on each state transition, so that the entire clearing pipeline becomes a deterministic object recomputable by every validator on the network [8]. The authority of this system no longer derives from an institutional principal behind it but from the fact that each of its steps is independently

recomputable (Section 28.6.2).

This substitution carries an often-overlooked cost that must be made clear at the outset of integration: determinism has a dual effect. Determinism makes the clearinghouse verifiable, but it also makes it predictable; the trigger conditions of a fully deterministic liquidation module, its execution path, and the boundaries of each of its steps can, in principle, all be computed in advance by an adversary with sufficient computational power, the root of the predictability attacks discussed in Chapter 12 (Section 12.4.1). A traditional clearinghouse purchases immunity from such prediction with discretion and secrecy; a decentralized clearinghouse exchanges verifiability for exposure to such prediction. This is not a design oversight but an intrinsic tension of the paradigm: the more a clearinghouse system relies on determinism, the more easily its execution boundaries can be computed in advance by participants able to parse the rules. This is the root of the incentive boundary and residual discretion discussed later in this chapter (Section 29.4).

The core difference between a decentralized clearinghouse and a traditional central counterparty lies not in whether an intermediary is removed but in the different bases of trust in the two: the latter's authority rests on capital, law, and regulation, the former's on the fact that anyone can recompute the books and arrive at the same answer (Sections 28.4.3 and 28.6.2). Every state transition described in the subsequent subsections—the ledger recomputation of clearing, the account-state commitment of settlement, the layer-by-layer bearing of losses in the default branch—must be understood as incorporated into this single state machine. Representing this chain of state transitions in full yields this chapter's core architecture diagram (Figure 29-1).

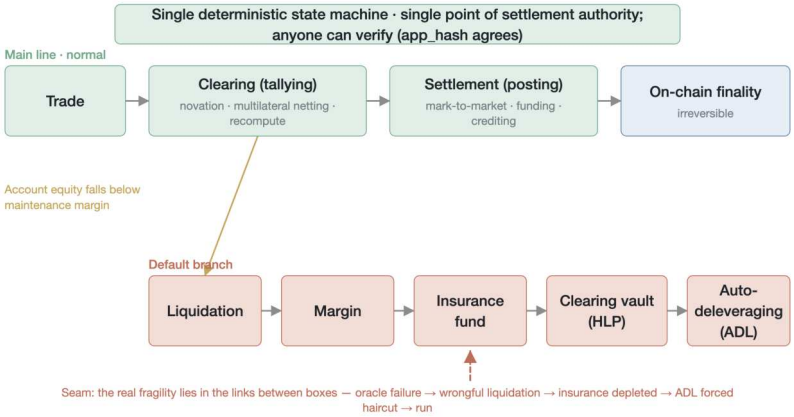


Figure 29-1. The overall state machine of a decentralized clearinghouse: the normal-conditions flow (trade → clearing → settlement → on-chain finality) and the default-management branch (liquidation → margin → insurance fund → liquidation vault → auto-deleveraging), together with the seams between their layers

29.1.2 The clearing stage: novation, multilateral netting, and recomputation

The main line of the state-transition chain processes every trade under normal conditions. After a trade enters the system, it first passes through clearing (recomputation of the books). Clearing, as a stage, consists of three actions each already treated in full earlier; this section does not restate their individual mechanisms but discusses only how they are composed into a single state transition. The first action is novation: the central counterparty interposes itself between buyer and seller, becoming the seller to the buyer and the buyer to the seller, so that a default by either party is no longer transmitted directly to its trading counterparty but is instead absorbed uniformly by the system (Sections 1.1 and 28.4.3). The second action is multilateral netting: multiple trades by the same account in the same contract, and even exposures across multiple counterparties, are consolidated into a single net position change, minimizing the number of entries that must be settled on the ledger (Section 5.2). The third action is recomputation: the system accordingly recomputes the account's position, realized profit and loss, and margin usage, producing a new set of account states. In the vocabulary of traditional clearing, these three steps correspond respectively to updating positions, netting down exposures, and recalibrating margin, and they are the core of a clearinghouse's daily operation [2].

These three actions were treated in separate chapters earlier; what the whole-architecture perspective must supply, and what the earlier discussion never made explicit, is this: within a state machine they are not three independent operations that occur in sequence but a single atomic state transition of the books. By atomic is meant that these three steps are either committed as a whole or rolled back as a whole, with no intermediate state observable from outside in which novation has completed but netting has not, or netting has completed but margin has not been recomputed. For a traditional clearinghouse this is an engineering goal to be approached only through end-of-day batch processing, reconciliation, and manual compensation, because its novation, netting, and margin management belong to different operational stages, even different legal moments, and aligning them into a seamless whole requires layer upon layer of rules, procedures, and end-of-day reconciliation [7]. For an on-chain state machine, by contrast, this atomicity comes from the native constraints of the state machine, because atomicity is inherent to the abstraction itself. Novation, netting, and recomputation can be composed on-chain into a single indivisible transition of the books because they were incorporated from the outset into the same deterministic system, rather than being assigned to three subsystems that must coordinate with one another.

Within this atomic transition there is also an often-overlooked ordering that directly affects how the subsequent default branch is understood: novation must precede netting, and netting must precede recomputation. Novation goes first because only once the central counterparty has become the common counterparty to every trade do exposures originally belonging to different bilateral pairs become eligible to be consolidated; without novation, A's long against B and A's short against C are two claims that cannot offset each other, and only with novation do they both become A's exposure to the system and thus nettable [7]. Netting goes before recomputation because margin

and profit and loss must be computed against the net position, not against the trade flow; if margin were recomputed trade by trade before netting, the result would be a grossly overstated usage, and capital efficiency would fall markedly. This internal ordering—novation before netting, netting before recomputation—is the entire reason clearing can compress tens of thousands of bilateral exposures into a limited number of net states while still measuring each account's risk accurately.

This atomic transition also has a temporal feature: it occurs entirely before any funds actually move. The clearinghouse first computes, on the books, each account's receivable and payable amounts and each account's margin usage accurately; the actual transfer of funds is a matter for the next segment (settlement). This ordering, completing the clearing computation before the settlement commitment, is not an accounting convention but the foundation of risk management: because the clearing computation precedes the settlement commitment, the system can determine, before any funds leave, whether an account is already insolvent and whether it must be routed to the default branch. Understanding clearing as a single atomic transition of the books that occurs before funds move thus draws a clear starting point for the next section's settlement and for the triggering of the default branch.

29.1.3 The settlement stage: marking to market, funding, and finality

After the clearing computation is complete comes settlement (commitment of account state). If clearing translates trading events into accrued states on the books, settlement realizes those accrued states as the accounts' actual states. It, too, consists of several actions treated in full earlier; this section addresses only how they string together into the second segment of the state-transition chain. The system continuously marks open positions to market at the mark price, so that each account's floating profit and loss updates continuously with price (Section 14.4); it charges and pays the periodic funding unique to perpetual futures, continually pulling the contract price toward spot, since perpetual futures have no delivery date and funding is the mechanism that substitutes for expiry-based delivery to maintain the anchor (Section 10.1); at close, it formally books the profit and loss and releases the corresponding margin; and finally, the chain's finality confirms this series of account changes as irrevocable (Section 28.4.3).

What the whole-architecture perspective must supply is the state-machine meaning of the settlement stage: it is the state commitment the clearinghouse makes to the outside world, and the chain's finality is the mechanism that confirms that commitment. Marking to market, funding, and booking are, during the clearing-computation stage, still only numerical evolutions on the books; only when they are settled by the settlement stage and confirmed by consensus into an irreversible stretch of on-chain history do they pass from what the system holds *should be* to what the system holds *already is*. This commitment must be identified separately, and must not be conflated with the clearing computation, because it marks the point at which responsibility transfers: before commitment, an unsettled profit or loss is only an internal computation of the module and may be overwritten or corrected by subsequent inputs; after commitment, it becomes the account's established entitlement, and any change to it can be made only through a new state transition that is likewise verifiable, not revoked without basis.

The settlement stage is therefore the sole boundary at which this system confirms a computed result as an established state. One point is easily misunderstood: finality is never a natural fact but a constructed one. Intuitively, people assume that an asset transfer becomes final the moment it is confirmed, as if irreversibility were an inherent property of the transaction itself; but anyone who has experienced a bank transfer being reversed or a settlement being clawed back because a counterparty went bankrupt knows that *when* finality takes effect is the mechanism most in need of being fixed deliberately. A settlement commitment is irrevocable not because it is physically impossible to reverse but because a set of rules expressly forbids reversal and imposes consequences on anyone who attempts it. In traditional finance this distinction is carried by the legal construction of settlement finality: when a settlement becomes irrevocable is defined jointly by clearinghouse rules, settlement-system rules, and bankruptcy law; the European Union even legislated specifically (the Settlement Finality Directive) to ensure that instructions entering a settlement system are not retroactively unwound because one participant goes bankrupt, thereby preventing systemic risk from spreading through that gap [9]. That this legislation was necessary shows that, absent intervention, this boundary is inherently blurred: the retroactive effect of bankruptcy law would collide head-on with the irrevocability of settlement, and only a dedicated statute carving out an exemption for settlement lets finality hold within that carve-out. On-chain, this same line of *when it becomes irrevocable* is drawn no longer by statute but is carried by the finality of consensus and defined by the consensus protocol itself [5]. Both draw the same boundary: before it, states can still be reordered; after it, no one may renege. The difference lies only in the instrument: traditional finance relies on legal text and court enforcement, on-chain systems on the liveness and safety of the consensus protocol itself.

This comparison reveals a subtlety of on-chain settlement that optimism easily obscures: finality comes in strong and weak forms, and the difference bears directly on the safety of the clearinghouse. Traditional settlement finality is a legal fact, absolute once attained; whereas the finality of many blockchains is probabilistic: a transaction included in a block means only that it very probably will not be rolled back, and the deeper the block, the lower the rollback probability, though not necessarily zero [10]. The divide between these two kinds of finality has its root in distributed-systems theory: for protocols that reach consensus by the longest-chain rule, safety is a decreasing function of the adversary's hashpower or stake share, and the rollback probability decays exponentially with confirmation depth yet never reaches zero; whereas for finality reached by Byzantine fault-tolerant consensus, once a block has been confirmed by a lawful number of validators' signatures, it is, under the protocol's assumptions, absolute and irreversible. The former delivers near-certainty, the latter certainty [10]. For general applications, near-certainty is good enough; waiting a few more blocks pushes the rollback probability down to negligible. But for a clearinghouse system this difference has substantive consequences, and those consequences manifest as an asymmetry. The clearinghouse's loss-bearing has a one-directional execution feature: once a profit or loss is settled by the settlement stage and, on that basis, triggers the next layer of default-branch actions, that layer of action has often already drawn on non-defaulting parties' capital: the insurance fund is tapped, and profitable parties are deleveraged. If at this point a chain

reorganization reverses the profit or loss already booked, the tapped insurance fund and the deleveraged profitable parties are difficult to restore: the prior book state can be rolled back, but the loss-bearing already executed downstream cannot correspondingly be restored. Probabilistic finality therefore leaves a new gap in the loss-bearing of the default branch, arising from a mismatch in which the antecedent is reversible but the consequent is not. This is one reason systems like Hyperliquid choose Byzantine fault-tolerant consensus and pursue deterministic finality that is irreversible once confirmed: a clearinghouse's demand on finality is far more stringent than a general application's, because it cannot afford this asymmetry [8]. Understanding settlement as a state commitment confirmed by finality is therefore not merely an abstract qualitative characterization; it is at the same time a hard constraint on what kind of finality the underlying consensus must provide.

Having clarified the role of finality, one often-confused ordering can be sorted out. Under normal conditions, clearing (recomputation of the books) and settlement (commitment of account state) together constitute all the work this system does: translating a continuous stream of trading events into a continuous stream of account-state changes confirmed by finality. This main line closes on itself: trades come in, states change, finality lands, in a continuous cycle, requiring no default handling. The default branch is a branch rather than part of the main line because it is triggered only when the main line cannot close on itself, that is, when an account's equity falls below maintenance margin under marking to market and the clearing computation, preceding the settlement commitment, determines it to be insolvent. The main line applies to states in which an account is still solvent; the default branch applies to states in which an account is already insolvent. The next section analyzes the latter.

29.1.4 Default management: liquidation and the loss waterfall

Beyond the main line lies a default branch, triggered only when an account's equity falls below maintenance margin. Its place in the state machine is a set of conditional state transitions, wholly dormant under normal conditions and activated only when the clearing stage determines that an account can no longer cover its exposure with its own margin. Its trigger point is liquidation: the system forcibly closes the insolvent position, confining the loss within its margin (Section 11.1.3). Liquidation is itself a default-management submechanism, not a clearing function; placed within the state-machine framework, it is the trigger point from the main line into the branch, and the first node in the entire system that turns a single account's individual loss into a problem that may affect other participants.

But liquidation often cannot confine the loss entirely. Market gaps, insufficient depth, and divergence between the mark price and the execution price can all leave the proceeds of closing out insufficient to cover the position's shortfall, leaving an uncovered shortfall (Chapter 12). The branch therefore continues downward into the default waterfall: after margin comes the insurance fund, after the insurance fund the liquidation vault (such as Hyperliquid's HLP), and after the vault, auto-deleveraging (Section 28.3.6; Chapter 12) [8]. Placed within the state-machine framework, each layer of the default waterfall is a conditional state transition: the exhaustion of the previous layer's buffer is the

condition that activates the next; and each layer, when activated, answers the same question in a different way: who next bears this loss that the defaulter cannot bear for itself. The answer at the margin layer is the defaulter itself; at the insurance-fund layer, the public pool prepaid by all traders; at the vault layer, the capital that voluntarily provides liquidity to the system; and at the auto-deleveraging layer, the profitable parties who are forcibly deleveraged. The waterfall is designed as a descending layered structure so that the bearers are arranged in an order running from those who most ought to bear to those who are most blameless, so that non-defaulting parties are affected as late and as little as possible [2].

Reading the waterfall as a series of conditional state transitions also clarifies an easily overlooked point: this branch shares the same state machine, the same determinism, and the same verifiability as the main line. When liquidation triggers, when the insurance fund is drawn upon, and whom auto-deleveraging selects are all uniquely determined by public rules, with no step left to off-chain negotiation or ex post discretion, in sharp contrast to the discretionary room a traditional central counterparty still retains in default handling, room that must be constrained by dedicated governance procedures [11]. The default branch is therefore not a separate logic outside the main line but the continuation of the same system in the world where it cannot pay in full; for that reason, each of its steps must likewise submit to recomputation by all, a point to which Section 29.1.5 returns.

But this branch also has the strictest limit in the entire system, corresponding directly to the third promise in the opening. Auto-deleveraging is the last layer of the waterfall, and its very existence announces one thing: when the total loss exceeds the capital available at all the preceding layers, the system no longer attempts to make up the shortfall but instead directly cuts the profitable parties' positions proportionally, rebalancing the books at a lower level. Chapter 12 has proved that among determinism, robustness to shocks, and fairness, no purely algorithmic default handling can satisfy all three at once, which constitutes a structural lower bound on bad debt: there is always a class of extreme scenarios in which losses cannot be fully covered and can only be mutualized or truncated (Section 12.2.3) [12]. Placed within the state-machine framework, auto-deleveraging is not a mark of failure but this system's second-best deterministic handling, under given constraints, of the fact that full loss coverage is impossible.

It is on the question of *who bears* that the on-chain waterfall reveals its core difference from the traditional default waterfall, and this difference constitutes the first of Chapter 28's three questions. A traditional central counterparty's default waterfall has five layers: the defaulter's margin, the default fund contributed by the defaulter, the central counterparty's own capital (its skin in the game), the default fund contributed by non-defaulting members (mutualization), and finally recovery and resolution (Sections 5.2.3 and 28.6.2) [2]. Empirically, the resources of these five layers are highly concentrated at the front end: in the default waterfall of credit-derivatives central counterparties, the initial margin posted by defaulters accounts for approximately 77.9% of all prefunded resources, while the central counterparty's own skin-in-the-game capital accounts for only about 2.0%; the latter is small in amount, but its position has structural significance, because it is placed ahead of all non-defaulting members' mutualized funds,

forcing the central counterparty to bear losses first (Section 5.2.3) [13]. The on-chain waterfall has only four layers, and what it skips is those two middle layers, allocating loss-bearing directly from the public pool prepaid by all traders to the profitable parties who are forcibly deleveraged, with neither a layer of skin-in-the-game own equity that bears losses ahead of users nor a layer of prefunded mutualization among members to share them. The depth of a waterfall lies not in how many layers it has but in how many layers of capital willing to bear losses first are placed before losses are allocated to non-defaulting profitable parties; the consequence of these two missing layers is the seam that Section 29.2 will audit specifically. This section establishes only the structure of the branch: it is the extension of the main line when it cannot pay in full, each layer is a conditional state transition, and each transition re-designates the bearer of the loss.

29.1.5 Concentrated adjudicative authority, open verification

Combining the main line and the branch (Figure 29-1) yields the overall architecture diagram of the complete clearinghouse system. The focus of this diagram is not the mechanisms inside each box, which have been analyzed separately in Chapters 5, 11, 12, 14, and 28; the focus is the connection relationships between the boxes and the overall constraint the system as a whole must satisfy. Before discussing this overall constraint, an often-confused pair of concepts must be distinguished: adjudicative authority, and the verification of that adjudicative authority.

Adjudicative authority is the power by which some mechanism determines the final state of each account. In this system, this power is single-point: all clearing computation, settlement commitment, and default handling converge on one deterministic computation module, which issues the sole ruling. Formally, this resembles centralization; the discretion over adjudication is indeed not dispersed but concentrated in one place. But the decentralization of a decentralized clearinghouse never promises to disperse adjudicative authority itself among everyone; handing the power to rule to multiple actors ruling separately yields not a clearinghouse system but a divergence of states that cannot converge. A telling point of comparison: a traditional central counterparty is likewise single-point in adjudication, and the decentralized design does not differ from it here; the entire engineering effort of blockchain consensus is to let multiple dispersed nodes agree on the same determinate state, rather than each maintaining a mutually conflicting ledger [4]. What it truly promises is something else: although adjudicative authority is single-point, the verification of that single point can be performed by anyone.

The technical carrier of this principle is an object often passed over in a sentence yet directly affecting the mechanisms that follow: the state fingerprint. Each time the computation module completes a state transition, its complete state is reduced by a cryptographic hash to a fixed-length fingerprint; the key property of this fingerprint is collision resistance: finding two different states that yield the same fingerprint is computationally infeasible, so identical fingerprints are almost equivalent to bit-for-bit identical states [14]. This property appears ordinary but in fact carries the entire weight of the verification logic. What a collision-resistant hash provides is a commitment in

the cryptographic sense: compressing the positions, margins, and profits and losses of tens of thousands of accounts into a fingerprint of a few dozen bytes, so that once this fingerprint is published, the publisher can no longer claim that the state at the time was substantively otherwise, and any subsequent tampering with the state would be exposed as a fingerprint mismatch. The checkability of this commitment is universal and cheap: a verifier need not trust the publisher, nor replicate all of its computational power, but need only run its own independently computed state through the same hash and compare whether the two fingerprints match—so the problem of comparing enormous states bit for bit is reduced to a single comparison of a few dozen bytes. Without such a commitment object, one that anyone can cheaply recompute yet the publisher cannot repudiate, verification by all would be only a slogan; with it, verification becomes, for the first time, an operation an ordinary node can actually undertake [14]. In an on-chain clearinghouse centered on this state machine, this fingerprint is typically written into each block as an `app_hash`: each validator independently substitutes the same batch of inputs into the same set of rules, computes its own state fingerprint, and then compares through consensus; as long as all nodes compute the same `app_hash`, this proves that they have seen the same path traversed by the same clearinghouse system (Sections 28.4.3 and 28.6.2). This is the concrete implementation of *verifiable by all*: verification is not to trust the module's self-declaration but to independently recompute its account states and then check them with an unforgeable fingerprint. Hyperliquid's HyperBFT consensus is a system that institutionalizes this single-point ruling with recomputation by all: the ruling is uniquely produced by the protocol, while every validator continuously recomputes and agrees on the result [8].

This places the trustworthiness of a decentralized clearinghouse on a foundation utterly different from that of a traditional central counterparty. Decentralization lies not in adjudicative authority itself but in the verification of adjudicative authority. This proposition must be set against common understandings to be understood accurately. One common understanding holds that decentralization should hand every decision to a vote of the many, and that the more power is dispersed among more actors, the more decentralized the system. But a clearinghouse is precisely the setting in which this understanding is hardest to sustain, because the books require a single, convergent ruling; if the state of a particular account at a particular moment were also handed to the many to rule on separately, the result would be not a more decentralized clearinghouse but a system that can never be reconciled and never converge. The key here is that what can and must be dispersed is not the ruling itself but the right to verify the ruling. The ruling must be single-point, or the system will not converge; verification must be by all, or the single point cannot be trusted. The entire engineering effort of blockchain consensus, letting multiple dispersed nodes agree on the same determinate state, was never meant to hand the power of decision to each person, but to let each person independently confirm that the sole decision has not been tampered with or manipulated [4]. The decentralization it promises is always the decentralization of verification, not the decentralization of adjudication. A traditional central counterparty is also single-point in adjudication, but its single point is tolerable because external capital, law, and regulation continuously constrain it and hold it to account when it fails; an ordinary user cannot verify the quality of its risk control and can only trust

that external constraint (what Chapter 28 calls regulated opacity, Section 28.7) [6]. A decentralized clearinghouse adopts the opposite design: it does not require users to trust that the single point will not act maliciously; it requires that every step of that single point be public to all participants and recomputable and falsifiable by anyone at any time. The former guarantees the single point with external authority; the latter constrains the single point with universal verifiability. This principle is the entire reason the chain of state transitions in the preceding sections is called a decentralized clearinghouse rather than merely an automated one: an automated clearinghouse need only be deterministic, whereas a decentralized clearinghouse must also be recomputable by all. For this very reason, this chain of verifiability cannot break at any link: even if only one critical input (for example, the source of the mark price) escapes the range of what can be recomputed, the entire chain's decentralization promise fails, which leads to the seam problem analyzed in the next section.

29.1.6 Inter-layer transmission and systemic fragility

Regarding the connections between the boxes, the true starting point for the rest of this chapter is a proposition more worthy of vigilance than single-point authority and verification by all. Chapter 28 traced a cascading-failure path: an oracle fails, causing the liquidation module to be unable to price correctly, leading in turn to large-scale erroneous liquidations or liquidation delays; erroneous liquidations rapidly exhaust the insurance fund; the exhaustion of the insurance fund triggers the auto-deleveraging mechanism to forcibly cut profitable positions; and the deleveraging destroys the market's last loss-absorbing capacity, inviting a collapse of confidence and run-like liquidity withdrawals (Section 28.8.1). Every link on this mechanistic path is, viewed alone, well designed: the oracle has multi-source aggregation, liquidation has trigger thresholds, the insurance fund has an actuarial model, and auto-deleveraging has ordering rules. Yet once they are connected, the disaster grows from the transmission from one link to the next: the failure of the previous link becomes the trigger condition of the next.

This is not an accident unique to on-chain systems but a commonality of all complex financial networks. Systemic-risk research has repeatedly noted that the most dangerous fragility of a financial system often stems not from the size of any single node but from the connection structure among nodes: risk contagion along these connections amplifies a shock that could have been absorbed locally into a global collapse [15]. This research also reveals a nonlinearity especially pertinent to clearinghouses: the robustness and fragility of a financial network do not vary monotonically with the degree of connection but exhibit a phase-transition-like reversal. When shocks are small, dense connections are beneficial, spreading a loss thinly across many counterparties so that no single node is crushed; but once a shock crosses a certain critical scale, the same dense connections abruptly turn from a robust sharing network into a fragile contagion network, and each connection that just now was used to spread losses becomes, at this moment, a conduit that transmits losses in amplified form [15]. This means that the system's safety margin does not improve smoothly as buffers thicken but has a nonlinear critical point: within the critical point everything appears robust, and beyond it the connections themselves turn into carriers of contagion. Chapter 28 noted that the value of a layered architecture lies in severing this contagion chain, downgrading an

instantaneous total collapse into a gradual pressure transmission, but it also conceded that layering does not eliminate systemic risk—it only changes the path of its transmission (Section 28.8.1). This section pushes that insight to its logical conclusion: since risk transmits along the seams, and transmission amplifies phase-transition-like at the critical point, the safety margin of a clearinghouse system is not the simple sum of its layer buffers but is limited singly by its weakest seam, because collapse always opens a gap at the first critical point to be crossed.

This leads to the core proposition that runs throughout this chapter: the strength of a clearinghouse system depends not on its strongest component but on its most fragile seam. The item-by-item analysis of the first 28 chapters directed attention chiefly to the components: each chapter analyzed one mechanism down to the level of formulas, proving how complex each is and how each fails. But what the whole-architecture perspective reveals is a fact that chapter-by-chapter narration does not readily present—that a system-level collapse almost never begins inside some component but always begins at the seam between two components, a seam for which no one is specifically responsible. The seam between the oracle and liquidation, between liquidation and the insurance fund, between the insurance fund and auto-deleveraging, between auto-deleveraging and market liquidity: the cascading path described above is, in essence, the trajectory of these successive seams being breached in turn. Seams are especially dangerous for a structural reason: each component has a clear owner (market makers are responsible for liquidity, the risk-control team for margin parameters, the oracle team for price-feed quality), but the transmission between two components often falls at the edge of everyone's field of view, becoming a link for which no one is specifically responsible. This ownerlessness is not negligence but a structural byproduct of the division of labor: because each team is assigned to optimize its own component and is assessed on its local metrics, the transmission between components is naturally excluded from clear responsibility boundaries, belonging to no job description, so that no one is specifically responsible for its failure. The strength of the system therefore obeys a near-weakest-link constraint: it is limited singly by the weakest seam, not raised by the strongest component; wherever one seam is breached, the individual complexity of the two components on either side cannot prevent a cascade. The entire point of integrating the clearinghouse into a state machine ultimately falls on identifying and reinforcing these seams, not on adding still more complexity to any component that is already complex enough.

The tasks of the remaining sections of this chapter can therefore be precisely defined along these seams. This chapter will answer, one by one, Chapter 28's three questions, and each question is, in substance, a question about some seam. The depth of the default waterfall is the seam between liquidation and the ultimate bearing of loss: starting from the defaulter's margin, through how many layers of capital willing to bear losses first must the loss pass before it reaches non-defaulting profitable parties? This is the subject of Section 29.2. Orderly degradation under extreme scenarios is the seam between the liquidation shock and the market's loss-absorbing capacity: when liquidation selling pressure strikes and no one plays lender of last resort, how does the system convert an instantaneous total collapse into a controlled, gradual failure? This is the subject of Section 29.3 [3]. Incentive alignment, in turn, is the seam between

deterministic code and the human discretion that can never be replaced by code no matter what: where do pure market incentives fail, and how can the residual discretion be constrained to its narrowest? This is the subject of Section 29.4. Reading the three questions as three seams is the key step by which this chapter turns from local mechanism analysis to whole-architecture analysis, and it is also the first time the catalog of failures discussed piecemeal earlier is gathered into a unified map of fragility.

29.2 The missing layers of the default waterfall

The end of Section 29.1 decomposed Chapter 28's three questions into three seams, of which the first is the most direct and the first to arise: the depth of the default waterfall is the seam between liquidation and the ultimate bearing of loss. Liquidation stops the loss at the defaulter's own margin, and this is the branch's trigger point; but liquidation often cannot confine the loss entirely (market gaps, insufficient depth, and divergence between the mark price and the execution price all leave an uncovered shortfall), so the loss exceeds the range covered by margin, moves down the default waterfall layer by layer, and continues to shift to the next loss-bearing party. Depth is concerned with how many layers this waterfall can draw upon, and from what capital sources, to absorb the loss before allocating it to non-defaulting parties. Chapter 28 condensed this seam into a question—whether an on-chain system can achieve a default waterfall as deep as that of a traditional central counterparty (Section 28.3.6)—and left it to this chapter.

This section's answer is direct, and rather severe: current on-chain systems cannot yet achieve it. This does not mean the chain lacks a waterfall. Section 29.1 already described its four-layer structure—margin, insurance fund, liquidation vault, auto-deleveraging—which is not even fewer in number than the traditional waterfall. The problem is that among these four layers, what is missing is precisely the two key resources that give the traditional waterfall its depth. What is lacking is not imagination in mechanism design but a layer of capital that no clear actor is yet willing to bear: whether the protocol's own equity is willing to sit in the first-loss position. This section proves this judgment through six subsections: first clarifying where depth actually comes from, then examining one by one the various attempts made on-chain to fill these two layers and analyzing why none has yet formed a sustainably operating mechanism, and finally returning to the repeatedly deferred but never avoidable question: what kind of capital is willing to bear the first loss.

29.2.1 The sources of default-waterfall depth

To judge whether the chain can replicate a traditional central counterparty's default waterfall, one must first define what the depth of the traditional waterfall actually means. A common misunderstanding is to understand depth as a large number of layers or a large reserve size. Chapter 28 already gave the complete layer ordering: the defaulter's margin is the first line of defense, the default fund contributed by the defaulter is the second, the central counterparty's own capital is the third (this layer is specifically called skin in the game), the default fund jointly contributed by the remaining non-defaulting members is the fourth, and below that come recovery and resolution procedures (Section 28.6). Among the five layers, what makes this waterfall

qualitatively different from a clearinghouse that relies on margin alone is not that it has as many as five layers but the two layers between the defaulter's resources and the non-defaulting parties' resources: the skin-in-the-game own capital and the members' prefunded mutualized default fund. These two layers are key not because of their size but because of their position.

The significance of the skin-in-the-game layer lies almost entirely in its ordering. Paddrik and Zhang's 2020 statistics on credit-derivatives central counterparties give a starkly contrasting pair of figures: the initial margin posted by defaulting members themselves accounts for approximately 77.9% of all prefunded resources, while the own capital contributed by the central counterparty accounts for only about 2.0% (Section 5.2.3) [13]. It should be noted that these proportions differ by cleared product; the figures cited here are the empirical values for credit-derivatives central counterparties (CCPs), and their absolute magnitudes should not be extrapolated as universal values for all CCPs. The 2.0% proportion is so small as to be almost negligible; by size alone, its actual contribution to loss absorption is minimal. But what traditional central-counterparty rules and regulatory practice retain this layer for is never its thickness but the ordering in which it is placed: it is typically positioned after the defaulter's own resources and before the remaining members' mutualized fund, bearing losses ahead of non-defaulting members' capital [2][13]. The entire purpose of this placement is to bind the clearinghouse's own interests to the risk control it operates.

To show why this ordering matters, imagine removing it. How high a margin a clearinghouse sets for members, what collateral it accepts, and how strict a standard it uses for stress testing—these discretionary choices all directly determine whether the waterfall will be breached, and they are all held by the clearinghouse itself. If the clearinghouse's own capital were not within the waterfall, or were placed after all members' mutualized fund, then the losses occasioned by relaxing its standards by an inch would be borne first by member capital rather than by its own capital, and it would then have a clear incentive: to set margins loosely so as to attract more business, transferring risk as much as possible to members. This is the basic moral hazard in the central-counterparty business model: the party operating risk control can avoid bearing losses for the failure of risk control. The skin-in-the-game layer is the structural corrective to this moral hazard. Once its own equity is forcibly placed ahead of other members' capital, it is, in making every risk-control discretionary choice, no longer a rule-maker that need bear no consequences but an interested party that will bear losses ahead of its clients. It realigns the operator of risk control with the bearer of the consequences of risk-control failure, which is the origin of its name and the reason its value cannot be measured by size. Theoretically, this layer is understood as a calibration of the central counterparty's incentives: when the clearinghouse can fully pass the cost of risk-control failure onto members, it has an incentive to relax standards and expand business; forcing it to place its own capital in the first-loss position is the most direct way to recalibrate this distorted incentive [6][7]. In other words, the 2.0% corresponds not to the principal loss-absorbing capacity but to a credible commitment: that the clearinghouse believes in the risk-control parameters it sets, believes in them enough to bear the risk of those parameter choices with its own capital.

The significance of the member-mutualization layer lies on another dimension, and it

turns on two points: first, prefunded; second, non-defaulting. Prefunded means that this capital is already in place and locked up before the crisis arrives, and does not depend on any party's ad hoc goodwill or resolve in a stress scenario. What a crisis most tests is the ability to raise funds on the spot, because at that moment market liquidity dries up, each actor prioritizes protecting its own liquidity, and participants generally lack the willingness to provide liquidity in extreme stress scenarios; a mutualized fund prepaid and locked up in advance eases this predicament: in a stress scenario it is a definite resource already on the books, not a promise still to be honored amid the crisis. Non-defaulting means that those who bear the loss are the originally non-defaulting members, not some blameless profitable party. These two points together give the central counterparty, when facing the collapse of a single member, a buffer within its authority that need not be raised on the spot and will not be withdrawn on the spot. It converts a single point of default into a bounded collective loss that the entire membership has agreed in advance to share, rather than letting the loss be randomly allocated across the market to some non-defaulting participant [11]. The prior agreement here is especially important: when members join the clearinghouse and contribute to the default fund, they have already accepted the rule that their contributions may be drawn upon when another defaults and its resources are insufficient; so when losses actually occur and the mutualized fund is drawn upon, no one is requisitioned by surprise; every contributor bears, according to the prior contract, a bounded and foreseeable obligation.

A well-designed default waterfall, by means of these two layers (one that makes the operator bear losses first so as to correct its incentives, and one of capital that the benefiting collective prepares in advance and shares by agreement), forms a buffer between the defaulter's margin and the non-defaulting market participants that is thick enough and clearly ordered. At the normative level, the Principles for Financial Market Infrastructures require central counterparties to hold sufficient prefunded resources, to establish default-management procedures, and to disclose key rules; the CPMI-IOSCO recovery guidance and the FSB resolution guidance further institutionalize recovery tools, loss allocation, and resolution continuity [2][16][17]. The specific layer ordering of skin in the game and the member default fund is embodied, in central-counterparty rules and in empirical waterfalls, as an arrangement whereby the clearinghouse, the defaulting member, and non-defaulting members bear losses in a predetermined order [13]. Depth, fundamentally, is this institutionalized, clearly ordered arrangement of resources; what it measures is not the size of capital but how many layers of preplaced capital can, before losses are allocated to non-defaulting parties, absorb losses first, in a set order and unconditionally. With this understanding of depth (ordering, prefundedness, and unconditionality, not the number of layers or their size), turning to look at the chain, the missing places come clearly into view.

29.2.2 The layer gaps in the on-chain waterfall

Placing the depth standard clarified in the previous section back onto the chain, the four-layer structure described in Section 29.1—margin, insurance fund, liquidation vault, auto-deleveraging—forms gaps at the two most critical points of the traditional waterfall, as soon as one asks after the funding source of each layer.

Take the skin-in-the-game layer. What comes closest on-chain to the form of a default fund is a USDC insurance fund like that of dYdX v4 or Perp v2; these funds do exist in advance and are positioned before users bear losses, and from these two points they seem already to have the appearance a buffer layer should have. But as soon as one asks where these funds come from, the difference emerges: their funding source is almost entirely the accumulation of trading fees [18][19]. This is a public pool collected by the protocol and remitted back into the risk buffer: in essence, it uses the fees that market participants themselves pay, accumulated for loss absorption, rather than the protocol putting up its own equity to bear risk. This distinction appears subtle on the surface, yet it is the whole crux of the skin-in-the-game layer. Every unit of capital in the fee public pool ultimately comes from the fees traders pay; it is a redistribution of market participants' funds within the system, and the protocol is merely the aggregator and manager of these funds: its own financial interest is not placed in the loss-bearing order. No one's equity, like the central counterparty's 2.0%, is forcibly placed in the first-loss position and bears losses ahead of everyone else. So on-chain there is a fund but no skin-in-the-game layer. As the previous section said, what skin in the game obtains is a credible commitment: that the operator is willing to bear the economic consequences of its own risk-control parameters; and a public pool supplied by traders' fees, however thick it accumulates, cannot form this commitment, because the financial fate of the protocol operating the system is not bound to the consequences of risk-control failure. This is what Chapter 28's verbatim judgment refers to: the on-chain waterfall typically lacks a layer corresponding to the central counterparty's skin in the game (Section 28.3.6).

Take the member-mutualization layer. The very last end of the on-chain four-layer waterfall is auto-deleveraging, which, after both the insurance fund and the liquidation vault are exhausted, forcibly cuts the profitable parties' positions to make up the shortfall. Formally, auto-deleveraging also shares out losses and seems to play the role of mutualization, but compared against the two conditions above, it is not equivalent to traditional mutualization in either respect. First, it is not prefunded. Auto-deleveraging is a forced allocation that occurs only ex post; before the crisis no capital was set aside for it: it is not a definite resource already on the books but a resource extracted on the spot from profitable parties' positions when the crisis occurs. Prefunded mutualization avoids the ad hoc fundraising constraint, whereas auto-deleveraging extracts this capital on the spot precisely when the crisis occurs, and it does so by forcibly cutting the profitable positions that other participants could have continued to hold. Second, those who bear the loss are not a non-defaulting collective that agreed in advance to share it but those non-defaulting traders who happen to be on the profitable side; they have never signed any contract agreeing to bear another's default losses: their profitable positions are cut merely because, at that moment, they hold profitable positions and can therefore be forcibly deleveraged. Chapter 28 already made this distinction plain: what auto-deleveraging cuts is profitable positions, and it is not equivalent, in either the population that bears losses or the incentives, to the mutualization in a traditional default waterfall, which is prefunded and jointly shared by non-defaulting clearing members (Section 28.3.6). Prefunded mutualization is capital that a set of actors have agreed in advance to share, that is bounded, and that does not withdraw

amid the crisis; auto-deleveraging is a loss allocation imposed ex post on specific non-defaulting parties. The two have only formal similarity; their economic meaning is not the same.

Placing these two gaps together, one can identify the true structure of the on-chain waterfall. It turns from a fee-supplied public pool directly to the ex post forced socialization of losses onto the profitable side, with the two middle layers (own-equity first loss and member prefunded mutualization) both absent. This layer gap is the fundamental reason the on-chain waterfall is shallower than the central counterparty's. It is shallow not in the number of layers; the four layers described in Section 29.1 are all present. It is shallow in this: when the insurance-fund public pool is exhausted, there exists, within the system's available range, no second layer of buffer composed of protocol equity or member prefunding to draw upon, so it can only immediately turn to auto-deleveraging, allocating tail losses to non-defaulting profitable parties more quickly and more directly (Section 29.1). A default waterfall with a skin-in-the-game layer and a member-mutualization layer still has, after public resources are exhausted, two layers of buffer, each independent and each with its own contributors, catching the loss in turn; whereas the on-chain waterfall presents as a structure lacking middle buffer layers, turning to non-defaulting parties to bear losses as soon as the public fund pool is exhausted, with no other buffer layer in between. The depth of a waterfall should be embodied in how many layers of buffer are still set before losses are allocated to non-defaulting parties, and it is at the point where buffers should most be set that the chain has two gaps. The remaining subsections of this section examine one by one the attempts made on-chain to fill these two gaps—whether by transferring the first loss to external capital providers, using a procyclical asset as a substitute for equity, or trying to migrate a mutualization mechanism on-chain, most of which have failed to operate sustainably—and identify why, to this day, no scheme has truly filled the gaps.

29.2.3 Liquidity providers as the counterparty: outsourcing the first loss

Since no actor is yet willing to place its own equity in the first-loss position, the chain has found a path that circumvents this constraint: outsourcing the first-loss position to capital outside the protocol. GMX and Gains are explicit liquidity-pool counterparty models, in which the trader's counterparty is not another trader but a capital pool funded by liquidity providers (LPs); profit and loss settle between the trader and the pool, and the pool is both the market-making liquidity and the first loss that absorbs shortfalls [20][21]. Synthetix's debt pool and Liquity's stability pool and redistribution mechanism are not the same kind of counterparty pool, but in economic meaning they likewise hand tail-bearing to stakers, stability-pool depositors, or other protocol participants, rather than to the protocol's own equity [22][23]. This design lineage does find several first-loss bearers for the chain, and formally it seems to fill the gap: after all, here is a body of preexisting capital that catches losses ahead of users. But the way it fills the gap exposes the meaning of the missing layer all the more clearly: the first loss is outsourced. The protocol itself still does not sit in the first-loss position; it merely places a class of capital providers—who enter voluntarily and can exit at any time—in the first-loss position.

GMX's experience displays the cost of this outsourcing plainly. Its GLP holders have directly borne losses twice, and on neither occasion was there any protocol-layer buffer ahead of them. The first was an oracle manipulation targeting AVAX in September 2022, which PeckShield estimated caused a loss of approximately $565,000 [24]. This figure is an estimate, and the attacker's identity remains unknown to this day: points that must be stated honestly, because this case differs from the clearly attributed manipulation cases discussed in Chapter 12 and must not be conflated with them; GLP holders received no compensation, and the loss was borne by them. The second was a reentrancy vulnerability in a GMX V1 contract in July 2025, which caused a loss of approximately $42 million [25]; only afterward did the protocol arrange compensation of approximately $44 million [26]. GMX has no insurance fund and no fee-supplied public pool, so the loss is simply GLP's loss, with no protocol-layer loss-absorption arrangement ahead of it [20]. The two events together constitute both the existence proof and the cost proof of the liquidity-provider-as-counterparty model: it does give the chain a first-loss layer that absorbs shortfalls, but this first-loss layer is a set of volunteers who can withdraw capital at any time, not an unconditional first loss borne by the protocol's own equity. Comparing the two events reveals another layer of meaning: compensation is ex post, discretionary, and event-dependent. The 2022 event compensated no one, while the 2025 event arranged excess compensation. Whether and how much compensation is paid depends on ex post decision rather than an ex ante contract, which shows that what GLP holders bear is not a first-loss obligation that is rule-bound and bounded but a risk of indeterminate exposure and uncertain compensation.

On the stablecoin side, crvUSD offers an isomorphic second example. It, too, has no dedicated insurance fund: once a shortfall appears on the minting side, it is borne passively by all crvUSD holders in the form of depeg risk; while bad debt in the isolated lending markets is borne by the lenders of each market individually [27]. Two empirical cases lay this bearing mechanism open. In June 2024, the Curve founder's own CRV-collateralized position was liquidated in a chain of liquidations across several lending protocols (including Curve's LlamaLend) as the CRV price fell by about 30%; he publicly acknowledged that this produced approximately $10 million in bad debt (most of which was later covered when a third party took it over at a discount) [28]. In the October 2025 crash, an isolated CRV-long LlamaLend market left approximately $700,000 in bad debt, borne by that market's lenders themselves [29]. Two easily confused things must be distinguished here: crvUSD's LLAMMA continuous soft liquidation and per-market debt ceilings are preventive designs that confine losses to a single market and keep them from spreading; they are not a loss-absorbing fund used to absorb losses after they occur. Prevention and loss absorption are two different things; isolating markets limits the range over which losses can spread, but it does not itself provide any capital to absorb losses; when bad debt truly occurs in an isolated market, it is ultimately still borne by that market's lenders themselves, with no layer of equity ahead of them bearing the loss. What isolation does is limit the range over which losses spread, not provide loss-absorbing capital for the market where the loss occurs.

Placing GMX and crvUSD side by side, the commonality of this class of design becomes clear: they do not lack first-loss bearers; on the contrary, they explicitly designate the

bearers: GLP holders, stablecoin holders, isolated-market lenders. The problem is that these bearers are all outsourced volunteers, not the protocol's own equity. This is exactly the opposite of the skin-in-the-game layer discussed in the previous section: the entire point of the skin-in-the-game layer is to make the operator of risk control bear losses ahead of its clients, thereby correcting its incentives; whereas outsourcing the first loss does the reverse: it moves the loss-bearing position away from the operator and hands it to a class of capital providers who have no say over the risk-control parameters yet must bear their consequences. What the liquidity-provider-as-counterparty model finds for the chain is not the skin-in-the-game layer but its opposite: the protocol transfers the first-loss position to external capital providers and does not itself sit in the first-loss position. This does not fill the gap but temporarily covers it with the funds of another class of capital providers; and since this class of capital providers can withdraw funds at any time, even this temporary cover is not secure.

29.2.4 The procyclical failure of reflexive-token collateral

If the skin-in-the-game layer is missing, and outsourcing the first loss to liquidity providers merely covers that missing layer temporarily, then an idea closer to the protocol's own equity is to have the protocol's token perform that layer's function—that is, to have token holders stake the token as loss absorption, and to slash that stake before users suffer losses. This mechanistic path has already appeared on-chain, and there is a fairly clear existence proof: Aave's Safety Module. The stkAAVE staked into the Safety Module can be slashed when the protocol runs a shortfall, to make up the gap [30]; its successor, Umbrella, further turns this mechanism into an insurance layer that can be configured per market and underwritten in separate tranches [31]. As a primitive, its significance should not be underestimated: it does prove that a first-loss layer can be constructed on-chain that is prefunded, slashable, and bears losses ahead of users. Structurally, this is closer to the intent of skin in the game than a fee public pool, because what is slashed are holders of an asset directly bound to the protocol's interests: they hold and stake the governance token, which is itself a risk-sharing commitment to the protocol's future, and placing their stake in the first-loss position conveys a commitment to bear the economic consequences of the risk-control parameters. As far as existence goes, the chain is not unable to construct an equivalent of the skin-in-the-game layer; stkAAVE and Umbrella are real counterexamples.

But using a token for this layer has an unavoidable tension, and this tension appears precisely when it is most needed: the value of a token is procyclical. What a central counterparty places in the first-loss position is cash, whose face value does not change when a crisis arrives (a dollar is still a dollar) and whose loss-absorbing capacity is independent of market conditions. A protocol token is the opposite: in the very crisis in which the system truly needs it to absorb losses, the token's price often plunges in the same direction as the assets whose losses are being absorbed. The logic here is closed: the system is under stress, bad debt surfaces, the market panics: these are the moments when the token most ought to bear losses; and the token's price is itself the market's discount of the protocol's prospects, so when the protocol suffers a major risk event, the market's judgment of its prospects deteriorates at once, and the token falls sharply at the same moment. Thus loss-absorbing capacity, at the very instant it is

most needed, instead declines: tokens of ostensibly ample value are staked, but by the time they must be realized to make up the shortfall, their own market value has already fallen; worse still, selling these tokens to make up the shortfall further depresses their price, forming a self-reinforcing price decline.

This is not an abstract worry; on-chain cases already show it. Synthetix long used SNX staking and a debt-pool mechanism to back sUSD, one of the most enduring and representative instances of reflexive-token backing [22]; but in April 2025, sUSD once depegged to about $0.68 [32], and by June 2026, SIP-423 had gone further and proposed freezing and retiring legacy sUSD, redeeming outstanding sUSD with newly minted SNX, and decoupling SNX staking from legacy sUSD obligations [33]. This means sUSD should no longer be described as a current success case still stably backed by SNX staking; more accurately, it has exposed the tail fragility of reflexive backing all the way into a governance exit phase. When the value of the backing asset and the stability of the backed asset affect each other, sUSD's stability was supposed to be supported by SNX's value, while SNX's value in turn hinges on whether the system runs well; once sUSD begins to depeg and the market loses confidence in the system, the two slide down together. Perpetual Protocol offers another, weaker piece of corroboration: its v2, through a matching structure based on Uniswap v3, reduced its reliance on the insurance fund itself [19]. Such design shifts show that a protocol token can be incorporated into a tail-tool portfolio but is hard to use as the principal stress-bearing capital in a default waterfall, because selling or issuing a token that is itself under stress to make up a shortfall amid a crisis often adds further selling pressure, weakening rather than strengthening solvency.

Reflexive-token loss absorption amplifies rather than absorbs shocks in the tail precisely because it builds the system's solvency on an asset highly positively correlated with the system. Cash-form own capital can serve as a buffer on the premise that its value is uncorrelated with the system's health; whereas the value of a protocol token is itself the market's discount of the protocol's future, highly positively correlated with the protocol's own fate. Building the first-loss layer on such an asset is tantamount to letting the buffer capital and the risk object being buffered fail together in the same stress scenario. This does not mean token slashing is worthless: stkAAVE and Umbrella prove it can structurally play the role of skin in the game, and under normal conditions and small-to-moderate shocks it does provide a real, slashable buffer, effective against local shortfalls that will not simultaneously crush the token's value. Its limitation is in the tail: in the extreme scenarios that would breach the entire waterfall and that most test depth, procyclicality weakens its absorbing capacity. In other words, the tokenized substitute the chain has found for the skin-in-the-game layer fails most easily where it least ought to fail, and the depth of a default waterfall measures exactly the loss-absorbing capacity in tail scenarios. A buffer effective only under normal conditions, but whose value falls markedly under stress scenarios, cannot yet be counted as the true depth of the waterfall.

29.2.5 Cross-protocol mutualization and the common-mode risk of restaking

While the skin-in-the-game layer is only outsourced or substituted, in limited fashion, by a procyclical asset, the member-mutualization layer's situation on-chain is even harder. Its on-chain candidate is cross-protocol mutual aid or a decentralized guarantee fund: letting multiple protocols share a backstop guarantee, structurally equivalent to the default fund in a traditional waterfall that is prefunded and jointly shared by members. This direction has been tried repeatedly, but the overall result is not encouraging; most attempts have failed to operate sustainably or to reach scale. To answer the first unresolved question, this history cannot be skipped.

The cross-protocol guarantee that has actually left a sustained payout record on-chain is, most typically, Nexus Mutual [34]. It is one of the few in this field that has indeed paid out to policyholders after events, and so it is often taken as an existence proof that on-chain member mutualization is feasible. But this single case itself illustrates the difficulty of the problem: around it, most other mutual-aid-type attempts have failed to operate sustainably. Mutual-aid-type insurers such as InsurAce and Unslashed often expose the pressure on reserves and governance discretion under correlated shocks. The UST depeg is one example: InsurAce's subsequently disclosed UST Depeg claims statistics show 234 covers corresponding to 173 claim filings, of which 155 were approved and 18 rejected, with net payouts of approximately $11.73 million after deductions, against related premium income of approximately $94,000 [35][36]. This set of facts cannot be simply written up as "claim denials" or "tightened payouts"; more accurately, it illustrates the capital pressure a mutual-aid guarantee faces in the face of systemic, co-triggered events: insured events often occur together, a systemic shock triggers claims on a large number of policies at once, and reserves accumulated in advance may be far smaller than the claim requests that arrive simultaneously. Per-market parametric insurance (such as Y2K, Cozy, and Risk Harbor) tries to underwrite with finer contracts triggered by specific risk events, avoiding the discretion problem of the mutual-aid type, but it in turn faces two other structural constraints. The first is basis risk: there is a gap between the payout trigger condition and the loss the policyholder actually suffers, so triggering a payout does not necessarily cover the actual loss, and an actual loss does not necessarily satisfy the trigger condition. The second is fragmentation of liquidity: the finer the slicing by event and by market, the more underwriting capacity is dispersed into isolated small pools, and the harder it is to form meaningful scale. Most of these protocols have already shut down or pivoted. Placing these successive attempts together, the chain's answer to the member-mutualization gap is: there have been repeated attempts, but most have failed to operate sustainably or reach scale, which shows this layer is not easy to construct.

The newly emerged restaking is the latest primitive in this direction and is often invested with high hopes. It directs the same stake toward multiple applications at once, using a single underlying economic security to provide guarantees for multiple protocols; EigenLayer and Symbiotic are its representatives [37][38]. Formally, this looks much like a cross-protocol shared backstop and seems just right to serve as on-chain member mutualization—one body of capital absorbing losses for many protocols at once. But it has two problems that cannot yet be avoided. The first is empirical: EigenLayer's

actively validated service (AVS) slashing went live on mainnet on April 17, 2025, and its design allows AVSs to set custom slashing conditions, Operator Sets, and Unique Stake Allocation [39]; as of publicly available information on June 25, 2026, it has not yet experienced a real, large-scale loss event to test whether it can pay out as designed under stress. A buffer that has never actually been drawn upon remains reliable only at the design level; and the value of each layer of a default waterfall can ultimately be verified only at the moment it is truly breached and truly drawn upon. A guarantee that has not experienced such a moment cannot yet be counted as usable waterfall depth.

The second problem, and the more crucial, is structural: a shared security layer reconcentrates some risk onto the same underlying stake. This point must be made clear, because it directly bears on the boundary of treating restaking as a mutualization primitive. Traditional member mutualization works on an implicit premise—that members' defaults are, to a large extent, mutually independent, so that when one member falls, the rest are healthy and able to share. The structure of restaking weakens this premise. Chapter 28 already issued a clear warning on this: restaking makes the same stake guarantee multiple applications at once, and the aggregate security is shared rather than simply summed, so slashings and operator failures across different applications become mutually correlated, which reintroduces exactly common-mode failure and correlation risk (Section 28.6.5). To be fair, EigenLayer's Unique Stake Allocation tries to confine a given AVS's slashing to the portion of stake the operator has allocated to that AVS, so as to isolate the slashing risk of unrelated tasks [39]; this reduces direct contagion between unrelated AVSs but does not eliminate the higher-level correlations of the underlying collateral asset's price, operator overlap, liquidity-withdrawal pressure, and simultaneous slashing requests. A body of economic security relied upon jointly by multiple protocols, once it shrinks in a systemic shock, correspondingly shrinks the guarantee it provides to all related protocols; if multiple protocols relying on it come under stress simultaneously, slashing or exit demands on the underlying stake arise simultaneously, and their scale may be insufficient to cover the multiple simultaneous requests. It is a new primitive but far from a verified solution: untested in the tail and still carrying the very correlation problem it was meant to ease.

29.2.6 The current absence of an equally deep waterfall on-chain

Taking the preceding five subsections together, Chapter 28's most direct question can, within the range of the main on-chain clearing, perpetual, and lending protocols examined here as of June 25, 2026, only reach one direct conclusion: a default waterfall as deep as that of a traditional central counterparty cannot yet be achieved on-chain (Figure 29-2).

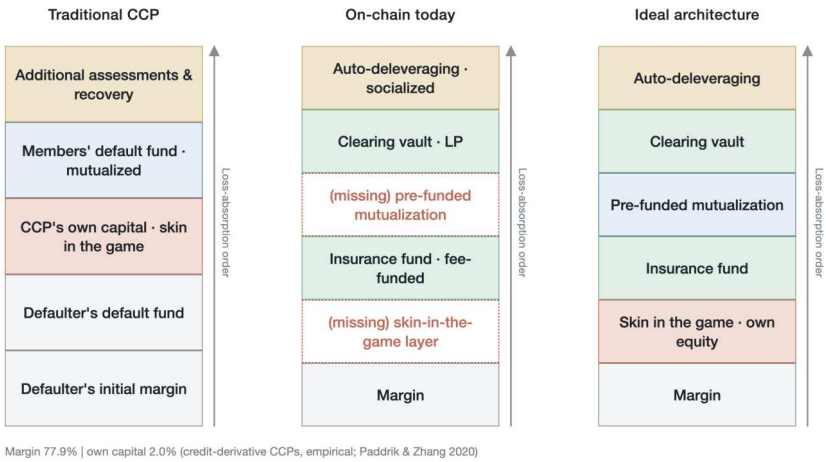


Figure 29-2. A comparison of default-waterfall depth: the traditional central counterparty's five layers, the on-chain status quo's four layers, and the ideal architecture (restoring the skin-in-the-game own-equity layer and the prefunded mutualization layer)

The weight of this conclusion is accurately grasped only after *depth* has been reclarified. Depth is never the number of layers, nor size, but how many layers of preplaced capital can, before losses are allocated to non-defaulting parties, absorb losses first, in a set order and unconditionally. Figure 29-2 places the three side by side: the traditional central counterparty's five-layer waterfall sets, between the defaulter's margin and the non-defaulting parties, the two layers of skin-in-the-game own capital and member prefunded mutualization; the on-chain status quo's four-layer waterfall turns from a fee-supplied insurance fund directly to auto-deleveraging of the profitable side, with those two middle layers absent; and an ideal decentralized clearinghouse needs to restore these two layers. Table 29-1 then compares, item by item, the funding source and ultimate bearer of each layer, so that the position of those two gaps comes clearly into view: in the traditional waterfall, both the protocol-own-equity column and the member-prefunded-mutualization column have clear contributors and bearers, whereas the two corresponding on-chain columns either read as a fee public pool, which is not equity, or point to profitable parties requisitioned ex post. The two exhibits together illustrate the same thing: the on-chain waterfall is not at a disadvantage in the number of layers; what is truly lacking is the two layers of resources that are correctly positioned, clearly ordered, and unconditionally in place.

Waterfall layer	Traditional central counterparty	On-chain status quo
First-loss position · the defaulter itself	Defaulting member's initial margin (approximately 77.9% for credit-derivatives CCPs)	Account margin
Own-equity first loss (skin in the game)	Central counterparty's own capital (approximately 2.0% for credit-derivatives CCPs, placed ahead of member mutualization)	Absent — the insurance fund is fee-supplied, not the protocol's own equity
Prefunded mutualization	Default fund prefunded and jointly shared by non-defaulting members	Absent — auto-deleveraging cuts profitable positions ex post, not prefunded mutualization
Backstop and socialization	Additional assessments and recovery tools	Liquidation vault (liquidity-provider capital) → auto-deleveraging (socialization onto the profitable side)

Table 29-1. Funding sources and ultimate loss-bearers of each layer of the default waterfall: a comparison of the traditional central counterparty and the on-chain status quo (Data source: compiled by the author based on Paddrik & Zhang 2020 [13]; the approximately 77.9% and approximately 2.0% figures are empirical shares of prefunded resources, vary by cleared product, and should not be extrapolated as universal values for all CCPs)

The chain has sought substitutes for these two gaps, and this section has examined them one by one; none has yet formed a sustainably operating mechanism. For the skin-in-the-game layer there are two substitutes: outsourcing the first loss to liquidity providers, and using a slashable token as an equity substitute. The former merely transfers the loss-bearing position to external capital providers rather than the protocol itself, and moreover the party bearing that position is a class of voluntary capital providers who can withdraw capital at any time; the latter, though structurally closest to skin in the game, carries a congenital procyclical defect and fails most easily in the tail scenarios where it most needs to function. For the member-mutualization layer there are also several substitutes: cross-protocol mutual aid, parametric insurance, and restaking. The first two have been tried repeatedly and mostly failed to operate sustainably, while restaking, though a new primitive, runs counter to the intent of dispersion because of common-mode risk and is untested in the tail; and auto-deleveraging at the end of the waterfall, still more, imposes ex post on non-defaulting profitable parties the loss that prefunded mutualization ought to bear. Each substitute touches the correct structural position, but none has yet formed an arrangement of unconditionally committing equity to that layer; they either transfer that position to external capital, or occupy that layer's position with a mechanism that evaporates under stress scenarios, or allocate the loss to actors who ought not to bear it.

The final judgment is therefore clear, and its cause lies not at the technical level. What is lacking is not some complex mechanism, nor engineering feasibility; primitives such as a slashable safety module and a cross-protocol shared guarantee have already appeared, and this section has shown their existence. What is truly lacking is the two most basic and hardest-to-obtain layers of capital: one layer of some actor's own equity willing to bear losses ahead of everyone; and one set of actors willing to prefund jointly, and

share by agreement, before a stress scenario arrives. The substitutes the chain has found for the former have still developed over only a short time and all carry the inherent defect of procyclicality or common mode; the substitutes found for the latter allocate losses ex post to actors who ought not to bear them. Filling these two layers is the precondition for a decentralized clearinghouse to move from *having a default waterfall* to *the default waterfall having sufficient depth*; and the difficulty of filling them was never, from the outset, technical but the basic political-economy question: what kind of capital is willing to bear the first loss. This is the direct answer, on the dimension of depth, to Chapter 28's first unresolved question. It also foreshadows the equally basic and intractable seam the next section will face: when losses ultimately exceed the limit any waterfall can absorb, how should a system without a lender of last resort fail in an orderly way.

29.3 Orderly degradation without a lender of last resort

The second question Chapter 28 left concerns how to balance risk control against user experience (Section 28.3.6). Pursued along this system, this question quickly converges on a harder seam: the seam between the liquidation shock the clearinghouse imposes on defaulting positions and the loss-absorbing capacity the market can provide at that moment (Section 29.1). Liquidation necessarily throws positions onto the market, and the market does not necessarily have sufficient capacity to absorb them; the portion that cannot be absorbed is what orderly degradation must handle. This section answers how a clearinghouse system, when it can neither assume external rescue nor avoid facing its own extreme stress scenario alone, can confine the unavoidable failure at this seam within a controlled range.

29.3.1 The absence of a lender of last resort and the shift in design goals

Traditional finance's extreme moments usually feature a set of external stabilization tools, of which the lender of last resort is only the most conspicuous. A decentralized clearinghouse has no such public credit supporter, a point this book has repeatedly confirmed (Chapter 11, Table 11-2). Its implication is not merely one fewer form of relief; the absence of a lender of last resort means the system cannot outsource its liquidity tail risk to the public sector's ad hoc funding and must face its own extreme stress scenario alone. A central bank can, amid panic, open a liquidity window to an institution facing liquidity-withdrawal pressure but still solvent; a regulator or trading venue can, amid disorder, trigger a halt mechanism; a market maker may, under certain market rules, bear a continuous-quoting obligation. These external supports have no equivalent in an on-chain clearinghouse where no single actor can issue commands.

To show the weight of this absence, one must first list the several tools traditional finance uses to dampen shocks, then confirm one by one which equivalent the chain lacks. The first is the central bank's lender-of-last-resort function, which, through the discount window, injects funds amid panic into institutions that lack liquidity yet remain solvent, cutting off a liquidity crisis before it escalates into a solvency crisis that would other-

wise have turned into liquidity-withdrawal pressure (Chapter 11, Table 11-2); the chain has no actor commanding such a public liquidity-supply mechanism that can create public credit out of nothing, and the balances of the insurance fund and the liquidation vault are both finite and prefunded, and each use depletes them. The second is the exchange's power to trigger circuit breakers and halts; traditional markets can, when prices fall through a preset threshold within a short time, forcibly suspend trading for a few to a dozen or more minutes, providing a time window for margin calls, digesting information, and repricing risk, cutting a stretch out of the positive feedback between panic selling and mechanical liquidation (Chapter 11); on-chain perpetual markets typically run around the clock, and although some protocols have admin pauses, parameter adjustments, or governance emergency channels, these are intra-protocol discretion, not external public authority. The third is market makers' obligation, mandated or institutionally incentivized, to quote continuously; traditional markets, under certain rules, still retain minimal quoting and market-maintenance arrangements; on-chain market making is mostly voluntary capital, which is all the more inclined to pull quotes the more a crisis deepens, so that loss-absorbing capacity evaporates at the very instant it is most needed. The fourth is the margin period of risk buffer; a traditional central counterparty calculates margin over a close-out risk horizon measured in days, leaving a buffer for margin calls, negotiation, and orderly close-out, whereas the chain compresses this time scale to real-time, trade-by-trade, with almost no room to maneuver between discovering insufficiency and executing liquidation (Chapter 11). These tools do not exist at the same intensity in all traditional markets, nor is the chain wholly without internal substitutes; the key difference is that the chain has no external backstop fully equivalent to the institutionalized shock absorbers of public credit, external halt powers, and mandated quoting obligations [3].

Therefore, a clearinghouse system without a lender of last resort must rewrite its design goal from *avoiding failure* to *failing in a controlled way in the worst case*. This adjustment is not an abandonment of risk governance but an acknowledgment of the resource constraint. Chapter 12 has argued that when losses exceed all of the defaulter's own resources and the system lacks sufficient external capital, full loss coverage can no longer be guaranteed by the mechanism itself; Chitra's model of auto-deleveraging likewise gives a formal constraint in the same direction, from the angle that solvency, revenue, and fairness cannot be had together [12]. The question thus shifts from "can no one lose anything" to "who bears the loss, how much, and at what tempo." A system that sets its goal at zero loss is promising a state that cannot be guaranteed by existing internal resources; only a system that sets its goal at controlled failure truly faces its own situation. What "controlled" means can be unpacked into three more concrete engineering goals: predictable (the path and upper bound of losses are written into the rules in advance, rather than adjudicated ad hoc amid the crisis); containable (losses are confined to the range that ought to bear them and do not spill over to unrelated participants); and absorbable (the layer that bears the loss actually has the corresponding resources on hand to absorb it when the event occurs). The conclusion that full loss coverage cannot be unconditionally guaranteed by internal mechanisms does not, in engineering terms, mean the system is doomed to collapse; it means the system must abandon the illusion of leaving no loss and instead shape the unavoidable loss into

a predictable, containable, and absorbable form. Acknowledging this is the starting point of orderly degradation: the system no longer pursues the zero loss that its own resources cannot absorb but instead pursues confining the unavoidable loss within a predictable, containable, and absorbable range.

This is precisely one of the structural deficiencies the Bank for International Settlements pointed out in assessing decentralized finance: on-chain systems generally lack the shock absorbers on which traditional finance relies to dampen shocks: no central-bank liquidity window, no external halt power that a trading venue or regulator can exercise, and no market makers institutionally constrained to quote continuously amid a crisis [3]. A line of research isomorphic to this section's argument likewise notes that, absent such absorbers, the robustness of a decentralized clearing design must be constructed from within the mechanism rather than hoped for from external intervention [40]. The absence of shock absorbers is not the oversight of some protocol but an inherent property of the permissionless architecture; since one cannot seek help outside, the only handling path is to build the capacity to absorb shocks into the clearinghouse's own structure. This sharpens the section's question: to handle that seam between the liquidation shock and the market's loss-absorbing capacity, the chain's most direct and most engineerable internal means are two—confining the shock to as small a range as possible, and turning the shock itself into a design variable that can be actively adjusted. On the shock side, the system cannot conjure thicker market loss-absorbing capacity out of nothing, but it can decide at what scale and at what tempo the shock is thrown onto the market; on the loss-absorbing-capacity side, the system cannot attract external rescue funds, but it can decide which pool is responsible for absorbing, thereby framing the reach of the shock. The former manages the shock as a design variable; the latter delimits the scope of absorption through isolation (one acting on the shape of the shock, the other on the boundary of absorption), and together they constitute the principal internal intervention directions of a clearinghouse system without external aid at this seam. The rest of this section develops these two means in turn, then points out their shared cost.

29.3.2 Risk-pool isolation

The first means is to confine a single market's bad debt within that market's own risk pool, so that it cannot infect elsewhere. Its cautionary counterexample was already dissected in Chapter 12: in the JELLY incident of March 2025, Hyperliquid's liquidity vault HLP was forced to take over a single altcoin's defaulting position, and a risk that should have been confined to a niche market nearly imperiled the entire vault's solvency (Section 12.6) [41]. The problem lay not in the size of the vault but in the commingling of the vault: all markets' risk shared the same loss-bearing pool, so that any one market's tail event could infect participants of all other markets. One pool's capital simultaneously absorbing losses for all markets ostensibly raises loss-absorbing capacity, but in reality it forms a high risk correlation among all markets: once the most fragile market has a tail event, the other markets are affected as well.

The positive design therefore points to a single principle: let each market's risk be borne, as much as possible, by that market's own participants and not spill outward.

This principle already has several mature implementations on-chain, each setting its isolation boundary at a different location, and each mechanism is worth identifying. Aave V3's isolation mode carves out a separate set of constraints for higher-risk collateral: for collateral listed as an isolated asset, the debt borrowed against it is capped by a debt ceiling denominated in a stablecoin, and only governance-approved stablecoins, not arbitrary assets, may be borrowed [42]. These two constraints together lock the bad debt that a high-risk asset may occasion, in advance, within a bounded range measured in the most easily liquidated stablecoin, so that it cannot penetrate the entire lending pool. Euler v2, launched in September 2024, pushed isolation to a fuller position, organizing lending in units of independent vaults, each carrying its own set of risk parameters, so that the risk boundary is defined by the vault boundary and a vault's internal bad debt is held within the vault by that boundary [43]. Compound III organizes the entire protocol into several markets centered on a single base asset, each market having only one borrowable base asset, around which several collateral types are configured; whichever market a lender puts funds into is the only market whose risk it bears, and markets are not interconnected [44]. Morpho Blue goes further, splitting lending into a large number of mutually immutable, mutually isolated markets, each of which, at creation, fixes once and for all its collateral, borrowable asset, liquidation loan-to-value (LTV) ratio, and other parameters, unchangeable thereafter; a lender therefore faces an isolated risk exposure whose boundary and parameters are determined in advance and that no one can adjust after the fact [45]. Hyperliquid itself, in HIP-3, introduced isolated markets in which the deployer bears the risk: whoever launches a new market must post a stake, bringing that market's possible risk, together with its bad debt, into the range borne by the launcher's stake, rather than letting it flow back to the shared vault [46]. There is no basis for saying HIP-3 was directly spawned by JELLY; but as a conceptual point of comparison, the commingling problem JELLY exposed is exactly the kind of problem isolated markets are meant to solve.

The insurance layer is also evolving in the same direction, and this evolutionary thread relates directly to the skin-in-the-game issue discussed in Section 29.2. Aave's Safety Module, already discussed in Section 29.2, whose staked tokens can be slashed in a shortfall event, is a shared first-loss pool—all markets' bad debt can draw on the same stake [30]; its successor design, Umbrella, splits this originally shared, slashable pool into isolated insurance by market, letting stakers choose to provide first-loss guarantees for only one or a few markets, thereby confining the question of *who absorbs this market's bad-debt loss* to those willing to absorb it for that market [31]. From a Safety Module in which one pool is shared and treats all shortfalls alike, to insurance tranched and underwritten per market, the path runs in the same direction, from commingling to isolation; it combines the first-loss guarantee discussed in Section 29.2 with the isolation principle discussed in this section within a single design.

The cost of isolation severing contagion is that it simultaneously causes liquidity fragmentation, and this is the first cost this section must point out. An isolated market cannot call on other markets' depth to absorb its own shock; when it truly has a tail event, the loss-absorbing capacity it can draw upon is limited to the existing resources in its own pool. A commingled pool, though it spreads contagion, also means any one market can borrow the whole pool's depth to absorb a shock; isolation severs this layer

of mutual aid. This cost appears in the same form in each of the primitives above: the debt ceiling Aave V3 sets for an isolated asset, while blocking bad-debt spillover, also limits the absorbing depth such an asset can obtain; the finer Morpho Blue splits markets and the more tightly it locks parameters, the thinner the liquidity a single market can mobilize in a crisis; Umbrella's tranching of shared insurance by market, while making underwriters each responsible for their own, also deprives each market of the possibility of calling on the full thickness of a shared insurance pool. This leads to an unavoidable frontier tension: the finer the pool is split, the more fully contagion is severed, yet the thinner the depth a single market can mobilize in a crisis, and the more easily it is breached by even a modest liquidation; the larger the pool is aggregated, the thicker the loss-absorbing capacity a single market can borrow, yet once one market has a tail event, the risk spreads across the entire network. Depth and contagion are therefore two directions of the same trade-off: isolation trades liquidity fragmentation for contagion blockage (the more strictly the risk boundary is drawn, the fewer the resources available to absorb risk), and there is no optimal solution that simultaneously attains both thicker absorbing depth and contagion blockage.

29.3.3 Active shock management

The second means is to rewrite the price impact that liquidation imposes on the market, from a passively suffered consequence into an actively designed variable. Chapter 11 analyzed the liquidation shock as a passive consequence of cascade dynamics: a price decline triggers liquidations, whose selling in turn depresses the price, and the shock amplifies itself in a feedback loop (Sections 11.3 and 11.5). The idea of active shock management is to move this shock from the end of the causal chain forward to the starting point of design: since liquidation necessarily throws positions onto the market, at what price, at what tempo, and to whom they are thrown are no longer wholly exogenous, uncontrollable variables but parameters that can be designed in advance.

The clearest empirical contrast of this idea comes from two versions of MakerDAO's liquidation system, which happen to sit before and after this means was discovered [47]. Black Thursday, March 12, 2020, was a full exposure of the old system. That day Ethereum fell about 43% in a single day, and on-chain congestion meant that only a very few liquidators could submit bids in the liquidation auctions; the fixed-duration English auction the old system used therefore produced zero-bid outcomes, with approximately $8.32 million in collateral cleared at a price of 0 DAI, and the system was forced to mint more than 20,000 MKR to make up the resulting bad debt [48]. Here two figures must be kept apart: the approximately $8.32 million cleared at zero price is the scale of collateral auctioned off at zero, while the bad debt that ultimately fell within the system's range to bear was approximately $4.5 million to $5.3 million; the two are not the same problem. The root of the failure lay not in the magnitude of the market decline but in the auction mechanism itself compressing the range of loss-absorbing capacity too narrowly. To show this, one must first identify the mechanics of the English auction: an English auction bids upward within a fixed time window, and the highest bidder wins the item; but once the chain is congested and bid transactions cannot make it into blocks, the number of effective buyers able to squeeze into the window drops markedly, the execution boundary closes at the time

limit, and the collateral is then cleared at zero price in a situation where, in effect, no one is effectively bidding [47]. The fixed-duration design limited the entry condition for loss-absorbing capacity to *must submit, amid congestion, a bid that successfully lands on-chain first*, and this entry condition is most restricted when the system most needs buyer participation.

Liquidations 2.0, launched in April 2021, responded one by one to the mechanism failures that day exposed, and its way of responding was to remake every dimension of the auction into an adjustable parameter [47]. It changed the English auction into a Dutch descending-price auction, and the mechanics of a Dutch auction happen to ease the exclusivity that congestion brings: the starting price is set at a multiple above the market price, and the price thereafter decreases over time along a preset curve, so that a liquidator need not bid first amid congestion, contesting scarce block space, but need only wait for the price to fall to a level favorable to it before participating in the transaction to conclude a fill, and the execution boundary of loss-absorbing capacity therefore no longer closes with time in a fixed manner [47]. This descending-price curve is itself a set of adjustable parameters, and this set of parameters happens to correspond to the several dimensions on which the old system failed. The multiple of the starting price relative to the market price determines how high the auction starts, setting an initial price above market for the collateral and then letting the market discover its true absorbing price. The auction duration has a cap; if the price descends through this duration and still no one takes it, the auction is deemed to have reached the cap and must be reset—this duration cap ensures that an auction does not indefinitely keep an unhandled risk position outstanding. The maximum decline the price is allowed to fall along the curve sets a floor, specifying the lowest proportion of the market price to which the price may fall in a single auction; crossing this floor triggers a reset rather than continued low-price disposal, avoiding the collateral being auctioned off at zero in a runaway price decline. This is the execution boundary that directly responds to Black Thursday's zero bids. And the step size of the curve determines whether the price falls smoothly and continuously or jumps down in discrete steps; the finer the step, the closer the price at which loss-absorbing capacity is attracted to the collateral's true liquidation price. Adjusting these several parameters amounts to adjusting the tempo and cost of the whole auction's attraction of loss-absorbers. It also settles fills atomically and in real time, and allows flash loans to participate in bidding: a bidder can, within a single transaction, borrow funds, buy the collateral, and then sell it to repay, with no need to hold cash in advance, so that potential buyers are expanded from the few liquidators holding cash at the time to anyone able, within one transaction, to borrow, buy, and resell, and the range of bidders is therefore greatly widened. It supports partial fills in batches, splitting the collateral in an auction into small portions that can be absorbed in tranches, not requiring a single buyer to absorb all the collateral at once, so that a large liquidation can be jointly digested in tranches by multiple small buyers, and the single-transaction capital required to absorb a large liquidation is broken down, further loosening the demand on a single buyer's financial strength. It is also equipped with a ClipperMom circuit breaker, which can be authorized to trigger a reset or even a suspension of the auction parameters when the oracle price deviates abnormally or the auction exhibits unexpected behavior, serving as a mechanism-level

safeguard. Placing these two versions side by side (Figure 29-4 contrasts the price–time curves of the English and Dutch auctions: the English auction waits for bids within a fixed duration and terminates when the execution boundary expires, while the Dutch auction lets the price descend continuously along a curve, absorbable atomically at any time) reveals the entire point of active shock management: it does not make the liquidation shock disappear entirely but turns the price and tempo of liquidation into mechanism parameters the designer can configure in advance. Every design of the old system that excluded loss-absorbing capacity from the execution range was changed into an adjustable parameter open to more absorbers. Its official design document does not name Black Thursday, but every item of this mechanism corresponds to some failure of that day.

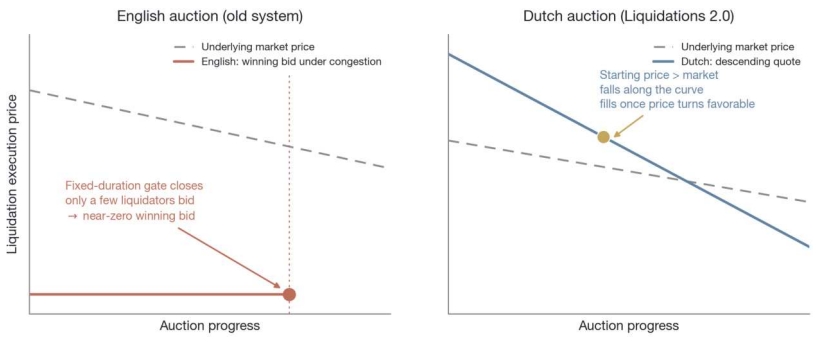


Figure 29-4. A comparison of the price–time curves of the English auction and the Dutch auction (qualitative axes; conceptual illustration, not empirical data)

29.3.4 Other degradation mechanisms

The Dutch auction is only one kind of active shock-management tool; the chain has also evolved several degradation primitives that differ in approach yet all refuse to treat the liquidation shock as an exogenous variable. Liquity adopts an absorption mechanism rather than an auction mechanism: it sets up a stability pool, letting stablecoin holders deposit funds into it in advance; when a vault is liquidated, the stability pool immediately burns the corresponding amount of stablecoin in exchange for the liquidated collateral, and the pool's depositors absorb this collateral in proportion to their shares and share in its profit and loss [23]. Because liquidation is triggered only when a position's collateralization ratio falls below a threshold, that is, when the collateral value is still slightly above the debt, the collateral the stability-pool depositors receive is, under normal circumstances, slightly more valuable than the stablecoin they burn, and this difference is their compensation for bearing loss risk; absorption is therefore not uncompensated bearing but a transaction with consideration. This mechanistic path is also equipped with a supplementary mechanism for the case where the stability pool is exhausted: when the stablecoin in the stability pool is insufficient to offset a liquidation, the debt and collateral of that liquidated vault are proportionally redistributed to the remaining still-active vaults, jointly borne by the owners of those vaults [23]. The value of the two mechanisms together is that they are immediate and do not sell to the open

market, thereby wholly avoiding the loss-absorbing-capacity problem an auction must face; the cost is that they socialize the loss onto the stability pool's depositors and even the remaining borrowers, and the absorbing capacity of the stability-pool main path is limited by the stability pool's balance at the time—once the stability pool is exhausted, the system can only fall back on the more dilutive and more contagious backup path of redistribution.

Euler's soft liquidation designs the liquidation discount as a continuous curve that rises as account health declines, so that liquidation is triggered when a position has just crossed the boundary and the discount is still small [43]. To show its mechanistic advantage, take as a contrast Compound III, which uses a fixed-discount liquidation model: whether a liquidator steps in when a position has just crossed the boundary or when it is already deeply insolvent, the discount obtained is the same predetermined fixed value [44]. The problem with a fixed discount is that, when a position has just crossed the boundary and the discount appears too thick relative to the risk, it incites liquidators to engage in front-running and gas-fee bidding to seize this certain profit. Euler's continuous discount curve is a response to this problem, designing the reward to liquidators to grow continuously as health worsens: roughly, the reward is proportional to the degree of health's deviation from a sound level, with the discount higher the closer the position is to insolvency [43]. The effect of this continuous incentive curve is to attract liquidation the moment it becomes marginally favorable: the discount is no longer a fixed value paid out in one lump after a threshold is crossed but rises continuously from a low level, so liquidation becomes earlier, smaller, and more dispersed, the upper bound of profit a single liquidation can seize is lowered, and the front-running and gas-fee bidding among liquidators to seize a large liquidation are correspondingly weakened. crvUSD's LLAMMA rewrites liquidation as a process occurring continuously along a price range: it distributes collateral over a price band and makes markets there, automatically converting collateral grid by grid into stablecoin along the band as the price falls, and back again as the price rises, so that a position is gradually and smoothly soft-liquidated during a decline rather than being hard-liquidated all at once at some threshold [27]. This mechanism in essence distributes the liquidation shock across a continuous market-making process over an entire price range, converting out only a small portion of collateral for each grid of decline, so the price impact is sliced into multiple smaller increments rather than a one-time large sell-off at some threshold; only in the extreme case where the price falls deep out of this soft-liquidation band and stays there for a long time does it fall into hard liquidation. It essentially converts liquidation from a discrete and irreversible event into a continuous transformation process that is reversible when the price recovers. Ajna takes another simplifying path: it removes reliance on an oracle and instead lets lenders each specify, when depositing funds, the collateral valuation they are willing to accept, so the market's liquidation threshold is determined endogenously by lenders' quotes rather than given by an external oracle that may be manipulated or may fail in a crisis [49]. When a borrowing position is undercollateralized by lenders' endogenous pricing, anyone may initiate a liquidation, launching a Dutch auction on that position and letting a descending-price auction find the absorber [49]; it therefore combines Section 29.3.3's Dutch auction with a set of liquidation trigger conditions that do not rely on an oracle into the same

mechanism combination, separating the liquidation-trigger determination from external oracle dependence. What these designs have in common is that they all refuse to accept the premise that the liquidation shock is an exogenous variable that can only be passively suffered; each starts from a different link, yet all convert the shock that Chapter 11 treated as a passive consequence into a parameter the designer can configure. The relationship of this set of examples to perpetual clearinghouses should be stated: they mostly come from lending and overcollateralized-stablecoin systems and cannot directly prove that perpetual clearinghouses already possess equivalent mechanisms; what they prove is that the price, timing, trigger, and absorber of forced disposal can be engineered and reshaped. Migrating them to a perpetual clearinghouse would still require handling compatibility among position transferability, order-book continuity, isolated/cross margin, market-maker inventory, and predefined loss layers. These paths are not mutually exclusive but decompose the same passive shock into several intervenable links (the destination of the selling, the shape of the incentive, the temporal distribution of liquidation, the source of the trigger determination), each of which was once an exogenous, uncontrollable condition and is now rewritten by some primitive into a designable endogenous variable. The chain's active management of the liquidation shock is therefore not the isolated design of some protocol but an engineering line advancing across multiple protocols from multiple angles at once. Together with Section 29.3.3's Dutch auction, they constitute the existing on-chain toolkit for orderly degradation (Figure 29-3 juxtaposes the effect of these two means on the cascade path: risk-pool isolation severs the lateral contagion, and active shock management slows the vertical escalation).

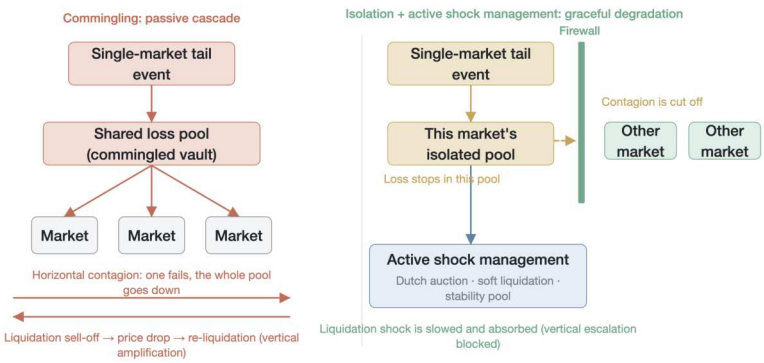


Figure 29-3. Risk-pool isolation and active shock management: from passive cascade to orderly degradation (conceptual illustration)

29.3.5 The three costs of orderly degradation

Every orderly-degradation mechanism has a cost that must be borne. Each means only transfers the loss from one position to another and cannot eliminate the loss itself: this is the architecture-level reappearance of Chapter 12's impossibility theorem (Section 12.2.3). That theorem shows that full loss coverage is no longer attainable; at the architecture level, this means any degradation mechanism can only choose among who

bears, when they bear, and in what form they bear, and cannot make the bearing itself disappear.

The first cost is the dilemma between slowing the auction and deepening bad debt. A Dutch auction slows the tempo to weaken a single shock, but the slower it goes, the greater the risk that the collateral's value continues to fall during the auction, and the deeper the bad debt ultimately left may be. MakerDAO's own design document acknowledges this dilemma almost verbatim, and points in two directions at once: too fast a decline (the price curve falling faster than loss-absorbing capacity enters) means bidders pay a price far below what they would have been willing to pay, producing avoidable permanent bad debt; too slow a decline means the auction price may keep failing to catch up with a still-falling market price, and the collateral keeps depreciating in the slow descent, likewise leaving bad debt; the document even concedes that the choice of price curve remains an active research area, with no single curve optimal in both directions at once [47]. This dilemma cannot be resolved once and for all by tuning parameters: its two ends each point to opposite failures, and it is not a problem with a single optimal solution but a double-bind whose two ends both lead to loss, so the most a designer can do is choose the relatively more bearable of the two losses. A study that quantifies, by simulation, the trade-off between liquidation speed and incentive cost provides a quantitative characterization of this dilemma, confirming that the relationship whereby slower liquidation requires higher incentive cost to attract loss-absorbers is not a qualitative intuition but measurable—each notch of slowing in liquidation requires, to still attract buyer participation at that moment, a higher discount or incentive, and this extra incentive is itself a cost the system bears [50]. The significance of this quantification is that it converts the ostensibly experience-judgeable trade-off of fast versus slow into a cost curve that can be measured yet cannot be eliminated: what faster buys is a smaller risk of market-price decline, and what it pays is a higher immediate incentive cost, and the two can only trade off against each other. Orderly degradation has no uniform standard solution for precisely this reason: the fast-or-slow of the auction sits between two kinds of bad-debt risk, and extremizing the choice in either direction produces loss.

The second cost is that absorption-type mechanisms socialize the loss. What Liquity's stability pool and redistribution trade for immediate absorption is the transfer of the loss from the liquidated party onto the stability pool's depositors [23]; these depositors are not the defaulting party of any position, yet, because they put funds into the stability pool in advance, they share the loss when liquidation occurs. Two different senses of socialization are worth distinguishing here. Under normal circumstances, the collateral depositors receive is slightly more valuable than the stablecoin they burn, and what they take on is less a loss than a transaction with a discounted compensation; only when the market falls deep enough that the collateral, once received, keeps depreciating to less than that stablecoin do depositors truly bear a net loss. The redistribution path goes a step further, allocating a defaulting vault's shortfall to a set of other borrowers who have nothing to do with that default, so the scope of socialization spreads from voluntarily pooled depositors to bystanders who never chose to take on anyone's risk. Socialization is not a defect of this mechanistic path but its mode of operation; it is precisely by spreading the loss thinly onto a set of bearers who are preplaced and largely

voluntary that it trades for immediacy and no selling pressure. That is, absorption-type mechanisms do not eliminate that loss; they merely transfer it from a defaulter forced to bear it onto a set of depositors who agreed in advance to bear it and take collateral at a discount as compensation, or, in the extreme case, onto a set of other borrowers who passively bear the loss; the loss changes bearers, but its total does not diminish.

The third cost is that isolation trades for liquidity fragmentation, a point Section 29.3.2 already made: what isolation trades for by blocking contagion is that each pool can only rely on the depth of its own existing resources to absorb a shock. Placing the three costs side by side reveals that they are, in substance, three facets of the same structural fact: loss can only be redistributed along different dimensions and cannot be eliminated. The fast-or-slow of the auction redistributes the loss along the time dimension: cash it out now at a discount, or drag it into the future and suffer deeper depreciation. Absorption-type mechanisms redistribute the loss along the participant dimension: from the defaulter onto preplaced bearers. Isolation redistributes the loss along the market dimension: deciding whether it is borne by this market alone or spills over to the whole pool. The three together point to the same fact: the core design task of orderly degradation lies not in eliminating these costs but in choosing, under a given market structure, a combination of costs that lets the system survive. They are all projections, on different facets, of Chapter 12's impossibility theorem; since loss cannot vanish into thin air, all engineering effort can only be engineering about how loss is allocated.

29.3.6 The incentive failure of liquidation execution

There is one further cost, more hidden and deeper than the three above, because it leads the discussion into the next section's theme. All these active shock-management mechanisms—auction, soft liquidation, or stability pool—presuppose one premise: that at the moment liquidation needs to occur, there will always be participants willing to execute it. The chain has no liquidators compelled by regulation to perform; liquidation relies on market incentives, and who liquidates depends on whether liquidation is profitable at that moment. To show how fragile this premise is, one must first identify the cost-benefit relationship of on-chain liquidators. A liquidator's decision to execute turns on a cost-benefit difference: the discount or reward obtainable from liquidating this position, minus the gas cost of sending this liquidation transaction on-chain, is worth doing only if the difference is positive. In calm times this difference is large: gas is cheap, discounts are considerable, and liquidators participate actively; whereas it is at the extreme moment when liquidation is most needed that this difference is squeezed from both ends at once. On one end is the gas cost: amid a crisis, on-chain transactions surge, competition for block space intensifies, and to get its transaction on-chain in time a liquidator must keep raising its bid in the priority-fee auction, so gas cost soars and may even exceed the profit of the liquidation itself. On the other end is loss-absorbing capacity: the liquidated collateral must be sellable to realize the discount, and in a crisis market depth precisely evaporates, so a liquidator, even given the chance to dispose of a position, may still show a negative return after selling. With both ends contracting at once, the difference that should have been positive turns negative, so liquidators rationally choose not to participate, and liquidation, at the moment

it is most needed, sees its executors exit.

This cost-benefit relationship has a sharper dynamic, manifesting as a dynamic change on the gas-cost side. The allocation of on-chain block space is itself an auction ordered by priority-fee bids: whoever is willing to pay a higher priority fee for its transaction gets packed first. In calm times this auction is lightly congested, and a liquidator can get on-chain by paying a negligible priority fee; but once a crisis descends and a mass of transactions contests the same batch of scarce block space, priority fees are bid up in competition, and a liquidator, to complete liquidation before the collateral depreciates further, often has to join this bidding up. So the gas cost of a liquidation is not an exogenous constant but a variable that rises endogenously with the degree of congestion: the heavier the congestion, the higher the cost of sending liquidation on-chain, and this occurs at the moment when liquidation most needs to complete quickly. When the priority fee is bid above the height of the liquidation discount, a rational liquidator stops participating—not because it cannot find positions to liquidate but because, on the calculation, it is unprofitable. Congestion and unprofitability here stack into a positive feedback: the deeper the crisis, the more expensive gas and the thinner loss-absorbing capacity, the more unprofitable liquidation, the fewer liquidators willing to execute it, the more unhandled bad positions accumulate, further intensifying the crisis. Liquidators' collective cessation of participation is, in essence, an executor exit that is unorganized yet triggered simultaneously by the same economic calculus.

This is not a theoretical worry but a repeatedly observed fact. In the sharp decline of May 2021, on-chain congestion and soaring gas costs caused some liquidations to fail to execute in time; the liquidation itself was profitable, but the on-chain transaction required to execute it either could not squeeze into a block amid congestion or had its gas cost offset the liquidation profit [51]. In November 2022, a CRV lending position on Aave could not be effectively disposed of by liquidators for a long time amid a targeted market manipulation, ultimately leaving bad debt in that market [52]; liquidators did not fail to exist: rather, when liquidation was unprofitable under such price and liquidity conditions, or when on-chain congestion made it impossible to fill, they rationally chose not to participate. Qin et al. give a systematic empirical characterization of the phenomenon of congestion causing liquidation failures, elevating the fragility whereby liquidation's execution depends on a market condition that may fail precisely in a crisis from an individual case to a measurable general regularity [51]. These two cases sit at the two ends of a spectrum (at one end, gas congestion making a profitable liquidation unexecutable; at the other, market conditions making an executable liquidation unprofitable), and both point to the same structural weakness: the occurrence of liquidation hinges on a condition that is least reliable in a crisis.

This is the structural gap of market incentives substituting for regulatory compulsion. Regulation can command a clearing member to perform, whether or not performance is favorable to it at that moment; whereas a market can attract liquidators only when it is profitable and finds no response when it is not. The divide between mandated performance and incentive-based performance lies here: the former treats executing liquidation as a non-refusable obligation, so a member must perform even at a loss, growing more rigid in a crisis; the latter sets it as an activity one may freely choose whether to join, so a liquidator may exit when it is unprofitable, and the execution

constraint tends to weaken in a crisis, exactly when it most needs to remain rigid. This gap is not the misconfiguration of some incentive parameter but an intrinsic property of the very choice to substitute incentives for compulsion: so long as the occurrence of liquidation depends on whether it pays at that moment, there is always an extreme moment of negative return that makes all carefully designed degradation mechanisms collectively lose their executors. A clearinghouse system that substitutes incentives for commands finds all its capacity for orderly degradation (isolation, auction, or soft liquidation alike) ultimately constrained by the same condition: whether its incentives still hold at the worst moment. And this assumption, whose core is *executors willing to participate at that moment*, is least reliable precisely when it most needs to hold. Orderly degradation thereby reveals its own boundary; this boundary can no longer be resolved by improving degradation mechanisms but must be answered head-on: when the system's stability depends on market incentives, and market incentives reverse in a crisis, what mechanism can this clearinghouse system still rely on? This is precisely the question the next section answers.

29.4 The boundary of market incentives and the constraint of residual discretion

The previous section ended on a gap. At the extreme moment when liquidation is most needed, liquidators may be collectively absent because it is unprofitable, and a clearinghouse system that substitutes incentives for commands finds its capacity for orderly degradation ultimately constrained by whether its incentives still hold at the worst moment. This gap is not an engineering defect that can be resolved by improving code but a boundary problem about what code can and cannot do. It leads to the last seam of this chapter: in a clearinghouse system that runs on deterministic code, there is a seam between the deterministic code and the human discretion that can never be replaced by code no matter what. What this section does is define the exact location of this seam: how far market incentives can reach, from where discretion becomes ineliminable, and how the ineliminable portion of discretion should be constrained to the minimum.

This seam is the hardest of the chapter because it cannot, like the first two, be narrowed by adding a layer of equity buffer or switching to a different auction mechanism. It asks not how to improve the system itself but who fills, and in what way, the remainder once the system's capacity ends. To answer it, one must first understand the first mechanistic premise on which a decentralized clearinghouse operates, then identify the boundary of this mechanistic premise, and finally explain how the scope borne by human discretion beyond that boundary should be compressed and constrained.

29.4.1 Market incentives as a substitute for regulatory compulsion

The first mechanistic premise on which a decentralized clearinghouse operates is the substitution of market incentives for regulatory compulsion. This mechanistic premise is the root-level divide between it and a traditional central counterparty, and to understand it one must start from two utterly different logics of mobilization.

In a traditional central-counterparty system, what makes clearing members fund the default fund and serve as loss-absorbing capacity in a crisis is, in the end, an obligation-based logic. For an institution to obtain clearing membership, it must, according to the rules, contribute its share of the default fund, must satisfy continuous capital and margin requirements, and must take on disposed positions in default-management auctions; these are not choices of whether it is willing but compulsory obligations attached to membership itself, backed by the clearinghouse's membership rules and the regulator's prudential requirements applying pressure jointly (Section 5.3.4) [2]. In comparing the two kinds of default waterfall, this book has already pointed out the strictest facet of this compulsory force: a traditional clearinghouse not only prefunds the default fund but also holds the power to levy statutory additional assessments on members, whereas on-chain protocols' insurance funds generally lack such a compulsory replenishment mechanism (Section 5.4.5, Table 5-2). A central counterparty can therefore, at the moment it most needs resources, conscript directly from a body of rule-bound members; it is this conscriptability that constitutes the reliability source of the mutualized capital in the waterfall's upper layers. That the negative externality of externalizing losses onto the system can be internalized also relies not entirely on participants' goodwill but first on this compulsory institutional arrangement operating as a constraint in the background (Section 12.1.3).

A decentralized clearinghouse has no membership rules that can issue compulsory directives, nor a regulator that can compel performance. What it can rely on is only another logic—mobilization based on self-interest. Liquidity providers voluntarily deposit funds into the liquidation vault because the returns from market making and absorbing liquidations are enough to compensate the risk they bear; liquidators voluntarily execute liquidations because the liquidation discount is profitable to them; arbitrageurs voluntarily close the divergence between the mark price and the spot price because the spread is itself a profit. No link is commanded; the operation of the entire system is built on a complex assumption: that each rational, self-interested participant, in pursuing its own return, happens to take exactly the action the system's stability requires. The clearinghouse commands no one but instead designs a set of payoffs so that dispersed, mutually distrusting participants, who bear no institutional obligation to the system, keep the system running out of self-interest alone. This book's default-waterfall comparison table compresses this divide into the two columns of ultimate loss absorption and liquidation execution: the ultimate backstop in the traditional system is the settlement member bound by statutory obligation, whereas the backstop in the on-chain system is external arbitrageurs and protocol liquidators, relying on the market force of the liquidation discount rather than any directive (Section 5.4.5, Table 5-2).

This substitution has a clear institutional appeal. It requires no trust in any specific actor—no single member constitutes an indispensable pillar of the system's stability, and if any one liquidator exits, another is attracted by the same discount; it also requires no one to obtain permission—becoming a liquidator or liquidity provider needs no qualification review, only an address and a sum of capital. In the language of mechanism design, this system pursues incentive compatibility: designing the rules so that a participant's dominant strategy happens to equal the behavior the system desires, so that the compatibility of incentives itself becomes the enforcement mechanism, with

no need to layer compulsion on top of incentives (Section 12.5.1) [53]. What Chapter 12 pointed out in evaluating Hyperliquid's liquidation vault is the full embodiment of this logic: liquidation profits accrue to liquidity providers, the vault's annualized return becomes directly the market's dynamic pricing of liquidation risk, and risk management thereby turns from the platform's burden into the community's opportunity (Section 12.6.1) [8]. This is also a mirror image of the governance that incomplete-contract theory reveals: since one cannot rely on a complete contract to exhaust and compel all due obligations, one settles for the next best and hands the remaining coordination to appropriately designed property rights and payoffs, letting the parties spontaneously execute, for their own interest, the behavior conforming to the system's goals (Section 12.4.4) [54]. In this sense, a decentralized clearinghouse system is less a law-enforcement system than an incentive structure: it does not stipulate what participants must do; it only ensures that when participants act in their own interest, what they do happens to be what the system needs.

The difference between these two logics is deeper than the literal opposition of compulsion and voluntariness. The obligation-based logic is, in essence, a governance relying on supervision and punishment; it presupposes an authority able to observe whether members perform and to impose sanctions when they default, and the very existence of this authority is an expensive apparatus that must be continuously sustained by capital, law, and institutions. The self-interest-based logic tries to move governance's cost forward to the design stage of the mechanism: once the payoff structure is tuned so that self-interest coincides with the system's goals, there is no longer any need, in real time, for a supervisor watching every participant at every moment, because no one has an incentive to do what the system does not want. This is precisely the ambition of the discipline of mechanism design: to induce, with rules designed ex ante, the behavior that would otherwise require ex post compulsion (Section 12.5.1) [53]. A decentralized clearinghouse pushes this ambition to an extreme that traditional institutions dare not reach: it simply abolishes the supervisory authority and stakes its entire mechanistic premise on the proposition that payoffs alone suffice to substitute for supervision. When the payoffs are sufficient, this is an institutional substitution attractive on a cost-benefit basis, replacing an entire regulatory system requiring trust and continuous maintenance with a payoff rule anyone can recompute. But this institutional substitution harbors an unavoidable point of fragility: the value of a supervisory authority lies in its not depending on whether the supervised still deems performance favorable at this moment; whereas pure payoff incentives stake the system's reliability on whether participants still find it worthwhile at this moment: a condition that varies with market state.

Replacing the obligation-based logic with the self-interest-based logic is a genuine transfer of institutional burden; it removes the entire membership-and-regulation framework on which a central counterparty relies to operate, in exchange for permissionlessness and trustlessness. But this replacement is not without cost. Encapsulated in the removed framework was a compulsory force still effective at the moment of crisis; once it is replaced by voluntary incentives, a hidden problem is inherited along with it: does that coincidence of self-interest and stability hold at all moments?

29.4.2 Procyclicality and predictability attacks

The boundary of market incentives shows itself most clearly precisely at the moment the system most needs them. The first proposition this section advances is: the coincidence of self-interest and stability is a procyclical assumption, holding in calm times and reversing in a crisis.

The fragility of this assumption was seen from one side in the previous section: liquidators rationally exit at the moment liquidation is most needed, because it is unprofitable or the chain is congested (Section 29.3) [51]. But it is not only liquidators who exit. Applying the same logic to the liquidity providers of the liquidation vault yields a stricter conclusion. A liquidity provider keeps funds in the vault to absorb liquidations because, under normal conditions, market-making returns compensate the tail risk it bears; but when volatility suddenly widens, when the mark price begins to jump violently, when successive liquidations are queued waiting to be absorbed, it faces not a business with a fitting risk premium but an exposure that may cause principal loss. A rational liquidity provider withdraws capital at this moment, which is precisely the moment the system most needs loss-absorbing capacity. The incentive has not failed here; it operates fully according to the logic of rational self-interest: it is the correct operation of the incentive that makes the capital meant to serve as a buffer withdraw liquidity at the moment the buffer is most needed. The preceding text already characterized this procyclical resonance of confidence and liquidity: from no one caring to everyone panicking, with almost no smooth transition zone in between (Section 12.3). The first cost of substituting market incentives for regulatory compulsion manifests here: it exchanges a resource that could be conscripted in a crisis for a resource that tends to exit in a crisis. The deeper meaning of the missing shock absorbers the Bank for International Settlements pointed out in assessing decentralized finance lies here: on-chain systems not only lack the central bank's liquidity window but also lack even those private actors that regulation requires to quote and absorb continuously in a crisis, leaving only voluntary capital that can exit as market conditions change (Section 29.3) [3].

The second proposition this section advances is more subtle: the very determinism of a purely algorithmic clearinghouse constitutes an attack surface. A central counterparty's default management is hard to exploit precisely because its critical moments contain discretionary factors; a regulator and market makers can, with unpredictable on-the-spot judgment, add a layer of uncertainty for anyone seeking to profit by exploiting the system's rules. A decentralized clearinghouse adopts the opposite design: all its rules are transparently encoded on-chain, and its reaction to every state is wholly determined and wholly deducible. This determinism is its institutional advantage: anyone can verify in advance how the system will treat them. But the same determinism is also its weakness: anyone can also compute in advance how much it costs to force the system into a particular state, and how much value can be extracted from the system's deterministic reaction in that state (Section 12.4.1). When a system's output for every input can be precisely predicted, it can not only be used but also strategically exploited. An attacker can pre-position ahead of a liquidation or auto-deleveraging that can be algorithmically triggered, letting the system's deterministic action, when pushed into that state, transfer value into the attacker's control. Many of the liquidation cascades and oracle manipulations dissected in Chapters 11 and 14 bear this structure of exploit-

ing predictability: the attacker need not breach any defensive rule; it need only know where and in what manner the defensive rule will trigger, and then exploit the subsequent state transition. The 2022 manipulation of a CRV lending position on Aave is an empirical instance of this structure: the attacker calculated exactly how the system's liquidation rules would react and turned the predictable liquidation logic into an attack parameter (Section 29.3) [52].

This reverse-exploited predictability has a name more general than *attack surface*; Goodhart's law, the Lucas critique, and protective opacity, all introduced next, are different facets of this same phenomenon of determinism being reverse-exploited. When a defensive rule is transparently made public, it turns from a mere rule into the objective function of the attacker's optimization, and the attacker adjusts its strategy precisely against the letter of this rule, making the rule fail exactly where it should most take effect. In analyzing why Hyperliquid's successive ex post patches were always circumvented, Chapter 12 used Goodhart's law and the Lucas critique to make just this point: once a metric is treated as a control target and made public, it ceases to be a reliable metric; defenses formulated on the basis of historical attack patterns fail because of attackers' expectations of those defenses (Section 12.6) [41][55][56]. This explains why a purely algorithmic clearinghouse sits in a loop it can hardly break by itself: its legitimacy requires transparent rules, and the transparency of the rules in turn turns every rule into a target that can be attacked; the more it explains itself clearly, the more fully it exposes its own reaction function. A central counterparty is not as deeply caught in this loop because it retains opacity at the critical points: what regulators and market makers can and cannot do at the moment of crisis is itself a variable the attacker cannot solve in advance. A decentralized clearinghouse, exchanging transparency for verifiability, also gives up this layer of protective opacity: not the oversight of some protocol but the necessary result of the fundamental promise that rules are executed as code and verifiable by all.

These two propositions point to the same deeper judgment: predictability is both the institutional foundation of a decentralized clearinghouse and its fundamental weakness, and this dilemma cannot be escaped by further refining the rules. Writing the rules in finer detail only describes the exploitable states more clearly; raising the incentives higher only briefly delays failure in the face of procyclical capital withdrawal. Market incentives are therefore not a force that can be relied upon unconditionally; they are powerful when marginally favorable but may fail collectively in a systemic crisis, and may even be reverse-exploited because of their own transparency and determinism. Frontier protocols already have direct experience of this: every round of ex post patching against a known attack vector is circumvented by adaptive attackers onto a new attack surface, and this cycle of attack, patch, and new attack is the manifestation of transparently published defensive rules failing in the game (Section 12.6) [41]. Acknowledging this boundary is not to concede the failure of a decentralized clearinghouse but to confirm that its range of capability has a clear boundary. Beyond that boundary lies a class of situations more fundamental than either procyclicality or predictability.

29.4.3 Incomplete contracts and the guarantor of last resort

Beyond the range market incentives can reach lies a most fundamental class of moments: states for which the contract has simply provided no rule. The previous two sections discussed incentives being at some moments not strong enough, or being reverse-exploited; this section discusses another class of problem—that at some moments there is simply no rule to follow.

Economics settled this long ago: contracts are incomplete. Because of bounded rationality and the unpredictability of the future, contracting parties cannot foresee and write down ex ante every future state; there are always situations the rule-drafter never envisioned and therefore never provided for (Section 12.4.4) [54]. Smart contracts push this theorem to the extreme. Called a contract, it is in fact a piece of code deployed on-chain; developers can set parameters only on the basis of their cognizance of the past and their anticipation of a limited set of scenarios, and cannot enumerate every manipulation technique, attack vector, and extreme tail event. When a situation never covered by the code truly descends, the code has no capacity for autonomous judgment: it either executes mechanically by the letter, yielding a result inappropriate or even destructive in that situation, or simply has no corresponding branch to run. The key concept incomplete-contract theory provides for this predicament is residual control rights: the power to decide in situations the contract has not provided for, which must fall to some person or institution (Section 12.4.4) [54][57]. The question is never whether states the contract has not provided for exist, for that is inevitable; the question is who exercises this residual control right when such a state arises.

The bearer a decentralized clearinghouse assigns this residual control right, Chapter 12 calls the guarantor of last resort. It is the role authorized to suspend, change parameters, or even roll back when the system falls into a state the code cannot handle; in a central counterparty this role is played by the clearinghouse operator and the regulator, while in a decentralized system it may fall to a core team holding multisignature authority or to a group of validators (Section 12.4.2). The very existence of this role is a sharp paradox: a system premised on trustlessness and rule-as-code execution must, at the most critical juncture, once again rely on a set of actors' on-the-spot discretion, and once discretion is introduced, the removed trust assumption is reintroduced in implicit form. But precisely because the incompleteness of contracts is structural and ineliminable, this paradox cannot be avoided by further raising mechanism strictness: any system claiming to need no guarantor of last resort at all does not truly not need one but converts that moment with no rule to follow into a systemic failure in the absence of a rule.

This is why the seam of incentives is the hardest of the chapter's four seams. The first two seams, the depth of the waterfall and the tempo of degradation, hard as they are, are after all engineering problems: they can be narrowed by adding a layer of equity buffer or switching to a different auction. This one is not an engineering problem but a problem about power, and it therefore leads a decentralized clearinghouse into its most intractable dilemma. On one hand, if the guarantor of last resort's discretion is too weak, too late, the system will be dismantled in an orderly way by attackers at that moment with no rule to follow, and the insistence on full automation may lead

to systemic failure; on the other hand, if this discretion is too strong, too frequent, it will gradually turn into the very center it was meant to replace, and more dangerously, it will, in the name of protecting the system, carry out selective bailouts. Chapter 12 has explained this layer of danger: when the blow-up of some large account may trigger a chain of liquidations, the guarantor of last resort's participation in trade intervention to protect the insurance fund is, for that large account, a de facto bailout, and for the other participants who acted by the rules, markedly unfair; the too-big-to-fail problem of traditional finance thus reappears on-chain through the emergency-discretion mechanism (Section 12.4.2). A discretion that must be both strong enough to complete crisis handling in extreme stress scenarios and weak enough not to be used for self-dealing or favoritism cannot be obtained by writing the code better; it can only be obtained by designing the shape of this discretion, ex ante, into a particular form. This turns the problem from *whether there should be a guarantor of last resort* into *within what institutional framework of constraint this indispensable and high-risk role should be placed*.

The JELLY incident is an empirical marker of exactly such a moment. When a manipulation of a low-liquidity token pushed the system into a state for which its code had no ready rule to follow, the code could give no answer. In that state, the liquidation vault was forced to catch a position it should not have borne; the oracle faithfully reported the very real market price that had been contaminated by manipulation; and the auto-deleveraging mechanism, which should have absorbed the tail loss, was not triggered, because the liquidation vault had not yet been isolated from the main vault and the denominator of its trigger threshold was computed as the vault's entire enormous assets (Section 12.6). Hyperliquid's ultimate handling was therefore not the execution of any prewritten rule but a discretion unconstrained by preset rules: the validators voted collectively, disregarded the real price the oracle reported, overwrote the market's settlement price to the level before the manipulation, and forcibly closed all related positions on that basis (Section 12.6) [41]. The several mechanisms necessary to hand loss absorption back to the market (asset isolation between the liquidation vault and the main vault, an independent deleveraging trigger threshold, dynamic position caps that adjust with market cap and depth, and the validators' power to delist anomalous assets by on-chain vote) were all filled in only after JELLY (Section 12.6). This ex post filling is itself the most direct acknowledgment that, at the moment of the event, the contract was incomplete: every rule filled in afterward is a retroactive recognition that at the time there was no rule to follow. What it simultaneously exposes is the limit of such ex post rule revision: the later POPCAT incident proved that targeted revision against the previous attack vector cannot prevent an adaptive attacker from finding the next gap (Section 12.6) [41]. This corroborates, from the reverse, the ineliminability of incomplete contracts: the system can make rules for accidents that have already occurred but can never make rules for accidents not yet conceived.

This is the boundary spoken of in the opening, and it can now be stated more precisely: no matter how complex the market incentives are designed, no matter how meticulously the code is written, they cannot cross the boundary of states the contract has not provided for; beyond that threshold, some residual control right must be exercised by a human. Acknowledging this, the ideal architecture no longer faces the question

of *whether discretion can be eliminated* (a foregone absence) but a truly engineerable question: since discretion cannot be eliminated, how can it be constrained to the minimum?

29.4.4 The three constraints on residual discretion

The ideal architecture's handling of this boundary is not to eliminate discretion but to constrain the ineliminable discretion to the minimum. The first proposition this section advances is: acknowledging that discretion cannot be eliminated is by no means the same as providing justification for arbitrary human intervention.

The distance between these two is precisely where the entire value of a decentralized clearinghouse lies. If, on the grounds that discretion cannot be eliminated, broad on-the-spot disposal power were routinely handed back to a center, then this system would regress into the very opaque centralized mechanism it was meant to replace; since the critical moments would still be adjudicated by human actors, the determinism recomputable by any node and the commitment requiring trust in no specific actor would both have only formal significance (Section 29.1). Therefore, acknowledging the ineliminability of discretion must immediately be caught by a constraint in the opposite direction: first, bisect the system's state space. In the large class of states the contract can provide for in advance (normal-condition clearing and settlement, liquidations with clear health, default handling with clear parameters), discretion is not only superfluous but harmful; these states should rely wholly on deterministic code, leaving no room whatsoever for human intervention, because any flexibility here only weakens determinism. Only in the small class of states the contract truly cannot provide for in advance is discretion permitted to intervene. This is the spirit of the two-layer architecture Chapter 12 derived from incomplete contracts: strictly adhere to code day to day, and authorize intervention only in a crisis (Section 12.4.4); and this section, on the basis of this principle, further asks what shape the permitted discretion itself should be given.

The second proposition this section advances is this structure's three constraints: for that portion of residual discretion, the design goal is to make it simultaneously narrow, slow, and visible. First, narrow means that the authorization of discretion is strictly confined within a specific scope and situation; the power the intervenor may exercise is only enough to handle that concrete, contract-unaddressed state, and cannot, under the name of one emergency authorization, do anything outside the authorized scope. An authorization to force settlement of an affected market when the oracle is manipulated is by no means the same as an authorization to rewrite any price at any moment one deems an emergency; the difference between the two is the difference between constrained discretion and unconstrained power. Second, slow means that, except for explicitly defined emergency situations, the exercise of discretion must pass through a timelock constraint that makes it impossible to execute instantaneously and unilaterally: after a decision is made and before it takes effect, a compulsory waiting period is left, giving dissenting participants time to exercise the exit option and withdraw funds, and giving a decision made in haste or under coercion a window to be corrected; and the emergency channel reserved for speed must itself have clear boundaries and an obli-

gation of ex post ratification, and cannot become an unconstrained path that bypasses all constraints (Section 12.4.6). Third, visible means that every exercise of discretion occurs on-chain and can be audited by any node—that is, who did what, when, under what authorization, and on what basis, all leave an undeniable record. Visibility is the natural extension, at the governance level, of the *verification by all* stated in Section 29.1: the discretion over adjudication may be concentrated in a few, but the verification of those few's discretion must be open to all (Section 29.1). The single point of adjudicative authority and the verification by all constrain, under normal conditions, the execution of code; in extreme states they should constrain human discretion.

What each of these three constraints prevents is a different path of discretion running out of control, and none can be omitted. What narrow prevents is the spread of authorization; history repeatedly shows that a power set up to address a particular crisis is all too easily invoked next time to handle situations it should not touch, until that emergency power set up as an exception becomes a routine instrument of rule; slicing the authorization narrow amounts to bounding, ex ante, the upper limit of harm in the worst case, so that even if the intervenor tries to overstep, the range of harm it can cause is limited. What slow prevents is haste and coercion; decisions in a crisis are often made in the few minutes when information is least sufficient and pressure greatest, and a compulsory waiting period gives the market both a right of exit to exercise the exit option and gives the decision itself a chance to be reviewed and overturned, turning an unreviewable unilateral decision into a correctable process. What visible prevents is ex post denial and non-transparent decision-making; if the basis and process of a discretionary act are not fully recorded on-chain, then whether it was to save the system or to shield some private interest can never be distinguished from outside; visibility constrains discretion from being explained however one likes ex post to leaving an immediate trace, so that every intervention must be prepared to accept everyone's review. There is an ordering of priority among the three: narrow constrains the scope of authorization, while slow and visible constrain the procedure of exercising power; an authorization that is sufficiently narrow, even if slow and visible are as yet imperfect, has already had the maximum harm it can cause bounded, whereas an authorization that is too broad, however strict the procedure, merely equips a power that should never have been granted with a set of procedural constraints.

Placing these three constraints on the governance spectrum Chapter 12 sketched, one can identify what exactly the ideal architecture fills. Chapter 12 arranged governance from the wholly opaque G0 to the fully-on-chain-governance-plus-timelock-constraint G3, of which the most mature, G3, with on-chain voting and timelocks, happens to cover the slow and visible spoken of here—it makes discretion slow down and be publicly disclosed (Section 12.4.6). But that spectrum measures transparency, degree of decentralization, and procedural constraint on the exercise of discretion; it answers how slow and how public discretion should be but does not answer how narrow discretion should be. A system reaching G3 may well be procedurally impeccable (every intervention voted on, passed through the timelock, recorded on-chain), yet the scope of authorization broad enough to let the intervenor do almost anything at that moment. Slow and visible constrain the procedure of exercising discretion; narrow constrains the scope of authorization discretion can reach. Narrow is therefore a dimension the

ideal architecture needs to fill beyond G3; it is not the procedural packaging of a broad power but a limitation of the scope of the power itself at the source. A narrow authorization, even with crude procedure, is closer to the ideal than a broad authorization equipped with ever-more-complete voting and timelocks, because the former bounds, at the source, the maximum harm the worst case may cause.

The three constraints of narrow, slow, and visible together give an evaluation standard by which any residual discretion can be measured. What they constrain is not whether a particular intervention is right or wrong but the boundary and exercise conditions of the very power the intervenor holds; a power whose scope is sufficiently limited, whose execution is sufficiently delayed, and whose disclosure is sufficiently full, even if occasionally misused, has its destructiveness bounded by a structural upper limit. This evaluation standard simultaneously redefines what a qualified decentralized clearinghouse means: the mark of qualification is not that it never needs humans but that, when it cannot but need a human, the harm that actor may cause has already been bounded, ex ante, to something sufficiently small.

29.4.5 The JELLY incident and the boundary of the purely algorithmic clearinghouse

Taking the above standard as the analytical baseline and returning to the JELLY incident, its significance divides into two levels. As an event, it is a failure—a failure in which the system was pushed into a state with no rule to follow and had to rely on a group of actors' on-the-spot vote to rescue itself. But as a marker, it indicates where the boundary of the purely algorithmic clearinghouse lies, and this is the last and most important proposition of this section: the boundary lies not in whether discretion is needed but in whether the discretion is constrained narrowly, slowly, and visibly enough.

Measuring JELLY's handling at the time against these three constraints reveals shortfalls in several places. It was not narrow enough: what the validators exercised was a power broad enough to overturn the oracle price and forcibly rewrite the settlement price, not an authorization only enough to handle that specific manipulation state. It was not slow enough: the resolution was reached in an extremely short time with zero opposition (Section 12.6), leaving no time window for dissenters to exit or for a hasty decision to be corrected, and the fast track that should apply only to genuine emergencies was here almost the only track (Section 12.6). Its visibility, too, held only in part: the execution result and on-chain state transition can be tracked, but the decision rationale, the boundary of authorization, the assessment of alternatives, and ex post accountability records were not made public ex ante in an equally full way, and *act first, explain later* replaced *disclose the basis first, then act* (Section 12.6). Add to this that outside auditable information on validator independence and the distribution of governance power was limited, so that a so-called unanimous vote is not by itself sufficient to equate automatically with a sufficiently decentralized governance consensus (Section 12.6) [41], and this discretion fell far short on all three constraints of narrow, slow, and governance visibility. Placing JELLY on the evaluation standard, what it shows is not that the chain should have no discretion but that this discretion was constrained too

loosely.

This is precisely the value of JELLY as a boundary marker: it does not prove that a decentralized clearinghouse is unworkable but marks the conditions under which it works. A misreading would say JELLY proves a purely algorithmic system ultimately needs centralized governance; the reading this section gives is the opposite: JELLY proves that the existence of residual discretion is unavoidable, but the degree to which it is constrained can range from loose to strict, and the quality of a clearinghouse system depends on this degree of constraint. Following this reading, the ideal decentralized clearinghouse should not be imagined as a system that never needs human discretion (a fantasy even incomplete contracts cannot honor); it should be imagined as another institutional arrangement: needing no one most of the time, because the states the contract can provide for are all handed to deterministic code; and in the very few moments when a human is indispensable, that human actor's power is sliced narrow enough, executed with delay enough, and disclosed fully enough that, even given private motives, it can cause only limited harm. The former constitutes the mode of normal operation, the latter the governance arrangement in the boundary state; normal conditions rely on code, boundary states on the constraint of discretion.

Thus the ideal decentralized clearinghouse sits between two attractive but untenable idealized models, negating both. It negates the idealized model of the opaque centralized mechanism, the basis of trust that there is always a sufficiently capable and trustworthy operator to make the right decision for participants at the critical moment, which the earlier parts of this book have repeatedly proved costly through successive failures (Chapter 12). It also negates the idealized model of the purely algorithmic, the full automation that supposes no human need ever intervene so long as the code is written completely enough, which is negated by the law of incomplete contracts: the blank of a contract is not a temporary gap left by insufficiently good current technology but the eternal predicament of any written rule facing an unknown future. The normative goal lies at neither end but in a constrained middle state: hand the states that code can handle to code as much as possible, shrink as much as possible the range of states that code cannot handle and that human actors must take over, and strictly constrain this minimized range with the three constraints of narrow, slow, and visible. This ideal promises neither an omniscient governance actor nor an omnipotent automation system; it requires only that, where reliance on human actors is unavoidable, the harm human discretion may cause be limited to the minimum. It does not pretend the boundary does not exist but places the entire center of gravity of governance on how to govern that boundary.

From this, JELLY also provides a standard for evaluating any real system. Facing a clearinghouse system that calls itself decentralized, the question to raise is no longer the binary one of whether human actors are still needed to intervene, for the answer to that is already fixed as yes; what should be attended to are three questions that can be answered by degree: how many states has it truly handed to code rather than left to discretion? How narrowly is its discretion sliced, how slowly released, how fully disclosed? When it cannot but invoke this residual discretion, where is the worst case bounded ex ante? Standardizing these three questions and using them to evaluate item by item the most frontier systems today, identifying the precise gap between ideal and

reality, is exactly what the next section does.

By this point, the chapter's four seams have each been made clear. The seam of depth is concerned with whose equity bears shortfall losses first; the seam of degradation is concerned with how extreme shocks are isolated and slowed; and the seam of incentives asks how small the scope of that residual power is bounded, and how strong the constraint, when incentives reach their boundary and discretion is unavoidable. Combining this fourth standard—that residual discretion must be narrow, slow, and visible—with the other standards distilled in the preceding three sections, we form a checklist that can be used to measure any real system. The value of a checklist lies not in stating an ideal but in using it to evaluate real systems and identify gaps. This is precisely the task of the final section.

29.5 The ideal-architecture checklist and the Hyperliquid gap audit

Folding the previous section's closing standard on the boundary of discretion into the checklist completes the distillation of the first four sections. Along the four seams, this chapter has each derived a normative principle that an ideal decentralized clearinghouse should satisfy. What remains is to organize these four principles into a reusable evaluation standard and then use it to measure how far the most frontier systems today are from the ideal. This section is therefore both the conclusion of the whole chapter and the transition to the next. It fixes the proposition "what the ideal architecture should look like," unfolded in the preceding text through continuous argument, into a checklist that can be verified item by item, and uses this checklist to conduct an honest and restrained gap audit.

29.5.1 Four normative principles

To distill the first four sections into a checklist, one must first recognize that these four sections are in substance answering four progressively deepening questions, not four mutually isolated ones. They arrange themselves along the same system from external structure to internal incentive: first asking whether this system has been correctly integrated into a whole, then asking whether its depth is sufficient once it begins to absorb losses, then asking whether it can degrade in an orderly way when losses grow too large for the depth to fully absorb, and finally asking where the incentives driving the whole system fail and who takes over after they fail. The four principles are the passing lines of these four questions respectively. The reason for establishing them as principles rather than preferences is exactly this: what each guards is the critical point past which the system, once it fails, degrades along a certain determinate path, so they are necessary conditions, not optional optimizations.

The integrity of the overall architecture is the foundation of all this. The criterion Section 29.1 established is that a decentralized clearinghouse system must be a single, deterministic state machine: clearing, settlement, and the default branch are not three subsystems that can each go their own way but different stages on the same uninterruptible chain of state transitions, where the output of any stage is the determinate

input of the next. The key significance of this principle lies not in determinism itself (a traditional central counterparty's clearing module is likewise deterministic) but in the requirement, unique to decentralization, layered on top of determinism: the discretion over adjudication may be concentrated in a single module, but the verification of every step of this computation module's output must be open to all, so that any node can independently recompute the same books and obtain a bit-for-bit identical result (Sections 28.4.3 and 28.6.2). Once this boundary is breached, the path of degradation is determinate. If adjudication is no longer single-point but split into several separately settling subsystems, the same default may be recorded as inconsistent books on different branches, and default handling degrades from an executable state transition into a set of ex post disputes hard to resolve consistently. If verification is no longer possible for all, external participants can only trust the operator's self-reported results, and the system regresses into an opaque centralized mechanism that happens to be deployed on-chain, losing not performance but the sole credential by which it is called decentralized. The passing line of overall architectural integrity therefore has two halves: one half is the single-point determinism of adjudication, the other the verification-by-all of that determinism; lacking either half, the system no longer has identity. This principle also contains an often-overlooked sub-item, namely whether the system has identified and reinforced its own most fragile seam (Sections 29.1.6 and 28.8.1). Its missing state is likewise determinate: so long as one seam remains unreinforced, what the attacker must do is not to breach the sturdiest module but to head straight for that seam and let the entire system fail as a whole at a link for which no one is clearly responsible.

On this foundation, depth measures how many layers of buffer the default waterfall still sets before losses are allocated to non-defaulting parties. Section 29.2 decomposed the depth of the traditional waterfall into two layers generally absent on-chain. One is a first-loss tranche composed of some party's own equity and positioned ahead of any member or user (skin in the game), whose significance lies not in size but in position; the other is prefunded mutualization in place before the crisis, jointly funded by the parties and dedicated to absorbing the spillover of a single member's default. These two layers are precisely the source of a traditional central counterparty waterfall's depth: the classic hierarchy arranges the defaulter's margin, the central counterparty's own capital, and the members' joint default fund in order ([2]), and empirically the defaulter's own resources make up the overwhelming share, while the own-capital tranche, though small, occupies the key position ahead of member mutualization (Section 5.2.3, [13]). The missing state of each of these two layers exposes why they are necessary conditions rather than decorations. Once the first-loss layer is absent, the designer is no longer bound to the system's safety: the designer can raise leverage and relax listing standards, because the cost of these aggressive parameter choices is borne mainly by users rather than by the designer, so risk appetite shifts systematically in a direction favorable to the designer and unfavorable to users: what the first-loss layer truly constrains is the designer's own risk appetite. Once prefunded mutualization is absent, the system, facing the spillover of a single member's default, has only two equally unfavorable handling paths left: either draw on the profitable positions that ought not to bear loss to make up the shortfall, forcibly spreading an individual default onto the non-defaulting majority; or let the bad debt sit on the books, with the protocol itself

continuously bearing the loss. The former harms fairness, the latter harms solvency, and the significance of prefunded mutualization is precisely to form this common fund in advance, before the crisis, so that the system need not, at that moment, choose between these two unfavorable handling paths. The passing line of this principle is that, before reaching the last resort of forced socialization onto the profitable side, the waterfall must first pass through these two layers, rather than turning directly from a public pool passively accumulated from fees to auto-deleveraging. The conclusion Section 29.2 reached is that, as of the main on-chain systems examined here on June 25, 2026, these two layers have not yet been truly filled; what is lacking is not imagination in mechanism but the political-economy willingness of *what kind of actor's equity is willing to bear the first loss*; whether this willingness can be institutionally induced has spilled beyond the boundary of risk engineering and points to the rule governance of Chapter 30.

Beyond depth, orderly degradation handles the disposal problem after depth is breached. Section 29.3's premise is a fact repeatedly established throughout the book: on-chain systems have no lender of last resort and therefore cannot assume any external liquidity rescue. Layered on the structural constraint that full loss coverage is no longer attainable (Section 12.2.3, [12]), the system's design goal must be rewritten from *avoiding failure* to *failing in a controlled way in the worst case*. The passing line of this principle is composed of two mutually independent yet complementary means: risk-pool isolation, confining a single market's bad debt within that market's pool and severing inter-market contagion; and active shock management, remaking the price impact liquidation imposes on the market from a passively suffered consequence into an actively adjustable design variable (Dutch descending-price auction, continuous soft liquidation, or partial liquidation), replacing the kind of concentrated sell-off, critical-point liquidation that easily invites zero bids. The missing state of each of these two means likewise points to a determinate form of disaster. If isolation is absent, a single edge market's bad debt spreads along the shared capital layer to all markets, amplifying an event that could have been locally absorbed into a system-wide solvency crisis. If active shock management is absent, liquidation regresses into a critical-point one-time sell-off: large collateral is concentrated-sold when buyer liquidity is insufficient, the price is broken further by the liquidation behavior itself, inviting zero bids, and liquidation, far from recovering value, further widens the shortfall and transmits it out through chain liquidations. The qualification of orderly degradation is not that the system never loses but that the system has prepared in advance for the worst case and clearly knows what each preparation must cost.

The boundary of incentives and discretion asks inward after the incentive mechanism that drives the whole system. Section 29.4 pointed out that a decentralized clearinghouse's first mechanistic premise is to substitute the "willing" market incentives for the "must" regulatory compulsion, but this force has a procyclical boundary: it is powerful when marginally favorable but may fail collectively in a systemic crisis, and the full predictability of the system's reactions itself becomes an attack surface. More fundamentally, contracts are incomplete—there are always states the drafter never envisioned and therefore lacking established rules, and in such states, someone must decide. The passing line of this principle is therefore not to eliminate all discretion

(forcing this unattainable goal only regresses the system into an opaque centralized mechanism) but to bisect the state space: hand the overwhelming majority of states the contract can provide for wholly to code, leaving no room for human intervention, and permit discretion only in the minority of states the contract truly cannot provide for, constraining this residual discretion to be simultaneously narrow, slow, and visible: strictly limited in authorization scope, delayed through a timelock in its exercise, and every exercise occurring on-chain and auditable by any node (Section 12.4.6). If any of the three constraints of narrow, slow, and visible is missing, the direction of degradation also differs. Lacking narrow authorization, an intervention that should be confined to an extreme state can seize the occasion to rewrite any parameter, and the emergency power continuously erodes outward into the territory that should have been left to code. Lacking time delay, rules can be rewritten in the instant before users have time to exit, the exit window the timelock was meant to protect disappears, and only a fait accompli remains. Lacking public auditability, discretion turns into a non-public state, and external participants can neither judge whether it has been abused nor hold it to account, so it no longer differs substantively from the unauditable discretion in an opaque centralized mechanism. The dimension of visibility is precisely the natural extension, at the governance level, of the verification-by-all that overall architectural integrity requires.

29.5.2 The Hyperliquid gap audit

Based on the four principles above, one can use them to measure a concrete system. The object of the audit is Hyperliquid, for a fairly clear reason: it is today the on-chain system closest to a complete decentralized clearinghouse with its own on-chain clearing and settlement module, and calibrating with it yields a distance most representative of the frontier level of the entire field. The facts on which the audit relies are all taken from what the preceding text has already treated in full (its consensus and state-machine architecture in Chapter 28, its waterfall layers and the exercise of discretion in the JELLY incident in Chapter 12), so this section does not restate how these mechanisms operate but only places them item by item on the checklist for assessment (Table 29-2).

Checklist item	Ideal standard	Hyperliquid's current status	Assessment
Overall architectural integrity	A single deterministic state machine; single-point adjudication with verification open to all; verification and governance paths do not re-concentrate	HyperBFT consensus and the Hypercore on-chain state machine make state transitions independently recomputable; but validator independence, the bounds of governance discretion, and the disclosure of decision rationales still constitute residual trust assumptions (Sections 28.6.2 and 12.6)	Partially met (state recomputability met)
Waterfall depth	Includes an own-equity first-loss layer (skin in the game) and prefunded mutualization	Margin → insurance fund → HLP vault → auto-deleveraging; the insurance fund is fee-supplied rather than own equity, so there is no skin-in-the-game layer; auto-deleveraging cuts profitable positions rather than providing prefunded mutualization	Partially met
Orderly degradation	Risk-pool isolation and active shock management	JELLY exposed risk-pool commingling; afterward the liquidation vault was isolated from HLP, and dynamic position caps and HIP-3 isolated markets were introduced; auto-deleveraging and partial liquidation exist, but auction-style ex ante shock management is absent	Partially met
The boundary of incentives and discretion	Market incentives; residual discretion that is narrow, slow, and visible	HLP is a permissionless, market-based vault; during JELLY, validators unanimously overrode the oracle price and forced settlement—execution traces are visible, but the decision process and rationale are insufficiently visible, and it was fast (an emergency, no timelock) and broad in scope	Partially met

Table 29-2. The ideal decentralized clearinghouse checklist and the audit of Hyperliquid's compliance (Data source: compiled by the author)

Of the four items, what Hyperliquid comes closest to meeting is the state-

recomputability sub-item of the first, and this is the hardest and most fundamental of the four principles. It is called most fundamental because it is the premise on which the other three stand: the waterfall, degradation, and incentives are all built on the foundation of the ledger being publicly acknowledged by all; if verification is impossible, the latter three, however complexly designed, lose the fulcrum by which external participants can test them, and in a system where no one can independently audit the books, it is meaningless to discuss whether the waterfall is deep enough or degradation orderly. It is called most difficult because it must satisfy two demands pointing in opposite directions at once: adjudication must be sufficiently concentrated to complete that uninterruptible chain of states in milliseconds; and verification must be sufficiently open to let any bystander recompute this high-speed chain bit for bit. Achieving both concentrated execution and open verification in the same system is the hardest tension to reconcile in all of decentralized finance, and it is precisely here that most systems retreat, either sacrificing execution speed for openness or drawing verification into a few actors' authority for speed. Hyperliquid has not retreated on state-machine recomputability: its consensus is produced by HyperBFT, its state maintained by the Hypercore on-chain state machine, the rules of clearing and settlement hard-coded into the protocol and executed identically for all positions, and anyone can pull the on-chain state, independently recompute the books with the same rules, and verify an identical result (Section 28.6.2, [8]). This means it does give, on the point of clearing output being recomputable, the strongest engineering proof. But that state transitions are recomputable does not mean the entire validator organization, governance authorization, and emergency discretion are already trustless; what JELLY exposed is this difference. Therefore the first item cannot be simply judged as met overall but only as met at the state-machine level and partially met at the overall-architecture level. The clearest affirmation in this audit should go to the sub-proposition that "high-speed adjudication and open recomputation can be reconciled by engineering," not to the stronger proposition that "governance and verification power are already fully decentralized."

The other three items are only partially met, and the location of the gap falls, with high consistency, on exactly the absences the earlier sections of this chapter already identified. This consistency shows that these three gaps are not implementation oversights but the appearance, in a real system, of the several structural voids the preceding text predicted from principle. On waterfall depth, it does have four layers (margin, insurance fund, HLP liquidation vault, auto-deleveraging), but measured against Section 29.2's two-layer criterion, both front-placed buffer layers are not in place. Its insurance fund is passively supplied by trading fees, a collected public pool rather than a skin-in-the-game first-loss tranche in which the protocol bears risk with its own equity, which is the concrete landing point of the previous section's *once the first-loss layer is absent, the designer is no longer bound to the system's safety*: a fee pool, however large, does not constitute the designer's commitment to write its own equity ahead of users. Its auto-deleveraging cuts profitable positions, in essence forcibly socializing tail losses onto non-defaulting counterparties rather than a true mutualization funded in place before the crisis (Section 28.3.6): the previous section's *once prefunded mutualization is absent, the former unfavorable handling path chosen between harming fairness and*

harming solvency. In other words, its waterfall skips the two layers the ideal standard requires, turning directly from a public pool to loss socialization. Judging it partially met rather than unmet is because it has indeed built a multi-layer waterfall that can operate in most situations, only with insufficient depth: what is lacking is not the number of layers but the nature of exactly the two most critical layers.

On orderly degradation, its score was driven largely by a single failure. The JELLY incident is exactly the kind of risk-pool commingling Section 29.3 used as a cautionary counterexample: the HLP vault was forced to take over a single altcoin's position and thereby imperiled the whole vault (Section 12.6, [41]). This is the previous section's *once isolation is absent, a single edge market's bad debt spreads along the shared capital layer*, a complete manifestation, with the contagion channel then open by default. After the event, it made a series of reinforcements that all converged toward risk-pool isolation: isolating the liquidation vault from HLP, introducing dynamic position caps, and opening isolated deployer markets through HIP-3 ([46]). But measured against Section 29.3's two means, what it has filled is mainly the isolation means, and it filled it rather passively, as a remedial adjustment after the fact rather than an architecture designed for the worst case in advance, its boundary set according to the type of event that has already occurred and not necessarily equally effective against the next, unforeseen kind of commingling. As for the means of active shock management, it remains incomplete: it has auto-deleveraging, partial liquidation, and a time-weighted mechanism for triggering liquidation, but not a mechanism like the Dutch descending-price auction that actively lowers the price impact by treating it as an adjustable design variable; it has the means to spread losses out but still lacks the mechanism to compress the shock itself before spreading it, and the latter is the core of the lending-side exemplars Section 29.3 showed in MakerDAO [47], Liquity [23], and Euler [43]. Isolation established, active shock management still lacking: hence partially met.

On the boundary of incentives and discretion, the assessment is most complex. Its core capital layer, HLP, is a permissionless, market-based vault into which anyone can deposit funds to serve as counterparty; this side is driven wholly by market incentives, meeting the first half of this principle concerning market incentives. The problem lies on the residual-discretion side. During JELLY it fell into a state with no rule to follow and exercised discretion, the validators voting unanimously to overturn the oracle price and force settlement at the level before the manipulation (the handling is described in Section 29.4.5, [41]). Measuring this discretion against the three constraints of narrow, slow, and visible, the conclusion splits cleanly: the execution traces and final state transition are trackable on-chain, but the decision basis, authorization boundary, assessment of alternatives, and ex post accountability were not disclosed in an equally full way; this is insufficient governance visibility, not unverifiable state execution. It was neither slow nor narrow: this was an emergency intervention made in an extremely short time, with no timelock delay whatsoever, making the previous section's *once time delay is absent, users have no time to exit and only a fait accompli remains* a real-world event; and overturning the oracle pricing and forcibly settling the entire market is far from a strictly limited narrow authorization. State visible, governance insufficiently visible, and both slow and narrow lacking: hence likewise partially met; the slow, narrow, and governance disclosure it lacks all belong to ex ante institutional constraints and could

have been filled with a preset timelock, a limited authorization scope, and a complete on-chain decision record, so the gap in this item belongs especially to the kind that can be converged by engineering.

29.5.3 Unmet items and the frontier of the field

In reading this audit, a common misreading is to read the several *partially met* results as an indictment of Hyperliquid. More accurately, the gaps the audit points out are, none of them, Hyperliquid's own oversight; they are boundaries the entire field has yet to resolve in common. The skin-in-the-game own-equity first-loss layer, true prefunded mutualization, the active shock management not yet mature on the perpetual-clearinghouse side, and discretion not yet constrained narrowly and slowly enough: applying this checklist to measure any other on-chain system yields gaps only more numerous, not fewer. Section 29.2 examined them one by one: as of publicly available information on June 25, 2026, from the fee-supplied-insurance-fund dYdX [18] and the first-loss-outsourced-to-liquidity-providers GMX [20], to the reflexive-token-backed Synthetix now entering the sUSD retirement process [33], and even the crvUSD without a dedicated insurance fund, no perpetual or lending protocol within the range examined here has truly filled those two buffer layers. Section 29.3 also showed that the mature exemplars of active shock management still come mainly from lending-side protocols like MakerDAO [47], Liquity [23], and Euler [43], and remain scarce on the perpetual-clearinghouse side. At the same time, these gaps belong to three different kinds of difficulty (the first loss and mutualization are constrained by political-economy willingness, active shock management by the migration of an engineering paradigm, and constrained discretion by the absence of ex ante institutional design), yet they exist, without exception, before all systems. That Hyperliquid is at the frontier level on state-machine recomputability yet constrained by the same boundary on the other dimensions itself shows that these gaps are structural, not individual: the places where it has not yet met the standard are not where its technology falls short of others' but the common baseline of the entire field.

These several structural gaps are, in substance, the concretization of a single higher-level judgment. Section 29.3 quoted the Bank for International Settlements' assessment of decentralized finance: on-chain systems generally lack the shock absorbers on which traditional finance relies to dampen shocks [3]. This is a highly general diagnosis that does not itself indicate where to improve in a concrete clearinghouse system, and the value of the four-item audit lies in landing this abstract *shock-absorber gap* at the clearinghouse level, translating it into four concrete coordinates that can be tested and improved item by item. Lacking the skin-in-the-game first-loss layer is lacking an endogenous equity absorber that bears losses ahead of users and non-defaulting parties; it differs from a central-bank liquidity window in that it absorbs the designer's own equity loss rather than liquidity temporarily provided by the public sector. Lacking prefunded mutualization is lacking a common fund already formed in advance and reserved for default spillover, for which the chain has yet to form a corresponding mechanism. Lacking constrained active shock management is lacking an absorber that lowers the shock itself and thereby reduces the amount that needs absorbing; its function is akin to market-making arrangements that maintain loss-absorbing capacity amid turbulence and halt

mechanisms in disorder, but it must be rewritten into an intra-protocol, verifiable mechanism. And still-too-fast, still-too-broad discretion is the last absorber, which must be borne by a human actor when all built-in absorbers are breached, not yet engineered and constrained; traditional finance hands this role to a resolution authority with rules to follow, and the chain has yet to form corresponding governance rules for it. The Bank for International Settlements offers a general judgment; this checklist decomposes it into four structural positions that can be filled item by item, and marks for each the absorber it roughly corresponds to in the old world, so the void is no longer a general judgment but four engineering goals with clear specifications.

For this very reason, the true use of this audit is not to judge but to define: it rewrites the question of *how far the frontier system is from the ideal architecture* into a set of engineering goals that can each be separately proposed, separately solved, and separately accepted: which four conditions are still lacking, what structure each condition should have, and by what kind of actor it should be borne. Its clear affirmation of Hyperliquid on the state-recomputability sub-item is the most valuable part of this definition: it proves that the hardest piece of a decentralized clearinghouse, concentrating the discretion over adjudication while opening the state output to recomputation by all, is indeed feasible in engineering and has already been realized by a real running system. Once this foundation is proved buildable, the remaining gaps are no longer questions of *whether possible* but questions to be resolved of *when they are filled, by what kind of actor's equity, and constrained by what governance rules.*

29.5.4 The interdependence of the four principles

One more point must be added, or this checklist is easily misread as four independent items that can each be satisfied separately and are unrelated to one another. They are not. These four principles constitute a progressive dependence within the same system, where the effectiveness of the latter level hinges on whether the former is in place. This means the total quality of the four items is not the simple sum of each item's score but closer to their product along the dependence chain: a shortfall at any link is transmitted and amplified along the dependence relationship.

Depth provides degradation with a buffer space that can be released layer by layer. If the waterfall is itself shallow and both front-placed buffer layers are absent, then the losses that need orderly handling arrive earlier and larger in scale, and the room left for the degradation mechanism to adjust is smaller; the shallower the waterfall, the heavier the execution burden on degradation. Conversely, no matter how complex degradation is, it cannot make up for a deficiency of depth: it can change the way losses are released but not the total amount of loss that needs releasing. So the quality of orderly degradation depends in part on the degree to which depth is filled, which is the deeper reason why, in the previous section, Hyperliquid's waterfall skipping two layers and its degradation mechanism still lacking ex ante shock management simultaneously remain at partially met: the shallow waterfall transfers too heavy an execution burden onto an already incomplete degradation, and the two gaps aggravate each other along the dependence chain.

Degradation in turn depends on incentives to execute. No matter how cleanly risk-

pool isolation is designed or how complexly the Dutch auction's descending-price curve is tuned, these mechanisms need execution actors driven by market incentives to be realized. The liquidator-exit gap at the end of Section 29.3 already showed that once incentives reverse in a crisis and executors exit in sync, even the most orderly degradation design is only an unexecuted design: on-chain congestion pushes up fees, liquidation is unprofitable, and liquidation is delayed or even fails, a point with both empirical cases (such as the CRV position on Aave in November 2022, [52]) and systematic academic characterization ([51]). This layer of dependence also contains a feedback: the shallower the waterfall and the larger the loss pushed to the degradation link, the higher the position risk liquidators must take on and the higher the risk compensation they demand: the same depth gap both aggravates the execution burden of degradation and raises the compensation required for executors to participate.

And incentives themselves are in turn constrained by the boundary of discretion. Market incentives are reliable on the premise that the rules are determinate, predictable to all, and modifiable by no one on the spot. But once discretion is too broad, too fast, too arbitrary, for instance if a position's settlement price can be overturned ex post by an emergency vote, then the very promise that *acting by the rules is profitable*, on which incentives stand, is itself shaken, and rational participants begin to price the risk that *the rules may be changed at any time*, thereby exiting earlier at the moment the system most needs them. Its reflexive feature is especially pronounced: discretion is often exercised in a crisis, at the moment the system most depends on executors' participation in handling, and that one act of overstepping weakens, at the worst point, the determinacy on which executors rely to participate: discretion used for emergency handling instead dissolves the incentives that could have supported emergency handling. Discretion constrained to be narrow, slow, and visible is therefore not only to prevent abuse of power but the foundation of credibility on which market incentives rely to be sustained. Incentives and the boundary of discretion are not two things but two sides of one thing: the more clearly the boundary of discretion is drawn, the more credible the promise of incentives.

Taking this layer of dependence together, what this checklist conveys is not that the four items being individually met suffices but that meeting them must accumulate progressively from the foundation layer up: without the integrity of the overall architecture, it is hard to form a credible default waterfall; without sufficient depth, degradation is hard to keep orderly; without incentives that can keep functioning, no one executes degradation; without constrained discretion, incentives themselves are not credible. The two ends of the dependence relationship are thereby connected into a feedback relationship (the terminal constraint of visible discretion, as the preceding text noted, is itself the extension, at the governance level, of the initial verification-by-all), so this progressive dependence is not a mutually severed linear relationship but a dependence with a self-reinforcing character. An ideal decentralized clearinghouse system is therefore not the simple juxtaposition of a combination of four mechanisms but a structure that can hold only by being joined layer by layer from the foundation up, which also explains why Hyperliquid can hold the state-recomputability foundation sub-item fairly firmly yet simultaneously halt on the overall architecture and the other three items: part of the foundation is already built, but every layer above still depends on whether questions

harder than engineering implementation can be answered.

29.5.5 From risk governance to rule governance

By this point, the chapter has completed two tasks. Integration: assembling, for the first time, the components dispersed throughout the book into a complete decentralized clearinghouse system, and pointing out that the true fragility lies not in the components but in the seams. Answering the questions: along this system's four seams, answering one by one the three questions Chapter 28 left, and fixing the answers into a four-item checklist that can measure any real system. The conclusion this checklist gives corroborates the chapter's main line: even the most frontier Hyperliquid today has only pushed the sub-item of *a recomputable single state machine* to the position closest to met; once validator independence, governance discretion, waterfall depth, and the degradation mechanism are all brought in, the complete architecture still has a distance that can be identified item by item, and behind those distances lie, almost all, not technical problems but the political-economy questions of what kind of actor's equity is willing to bear losses first and what kind of actor exercises discretion in states the code does not cover.

But this points precisely to a proposition the chapter cannot resolve on its own. This chapter has analyzed the clearinghouse as a risk-governance architecture; what it must solve is how losses are absorbed, how shocks are degraded, and how incentives are aligned. Yet none of those three hardest-to-fill gaps can be filled by a more complex contract alone: the skin-in-the-game first-loss layer concerns what kind of actor is willing to bring its own equity under the rules and accept the rules compelling it to lose first; prefunded mutualization concerns what institutional arrangement makes unfamiliar participants jointly inject funds into a pool before a crisis and trust that this fund will be mobilized fairly; constrained discretion concerns how, when code must allow an authorized governance actor to intervene, that actor's power is framed by rules in advance. All these questions have spilled beyond the boundary of risk engineering and turned into another domain—how the rules themselves are written, executed, and governed. That is, to complete this risk-governance architecture, there must first be a governance system that can credibly formulate and execute rules.

The next chapter turns from this to the problem of rule governance. When compliance and governance are no longer constraints imposed on the system from outside but engineered directly into the protocol itself, and when the formulation, alteration, and enforcement of rules all become on-chain verifiable processes, how will this system operate? From a risk-governance architecture to a rule-governance system, the subsequent analysis thereby turns from *how to absorb losses* to *how to govern rules*. What Chapter 30 will analyze is precisely this rule-governance system: embedded compliance and automated governance.

Chapter summary

This chapter completed two pieces of work: integrating the various components dispersed across the book's first 28 chapters into a complete decentralized clearinghouse

system, and, along this system's four seams, answering one by one the three questions Chapter 28 left at its close. The first piece of work is an engineering-level response to the preceding mechanism analysis: the earlier text explained how each component fails, and this chapter answers how, absent a central institution, these components can be reintegrated into a system that can be trusted; the second is a direct answer to three constructive questions, not a restatement of existing mechanisms.

The first conclusion integration yields is that the strength of a clearinghouse system lies not in a single component but in the seams between them. The book's earlier text has analyzed every component to the level of formulas: novation, multilateral netting, marking to market, margin, the reflexivity of liquidation, the default waterfall, the insurance fund, and auto-deleveraging. But what breaks a clearinghouse system is usually not the failure of some one component but the transmission between components: an oracle failure triggers erroneous liquidations, erroneous liquidations exhaust the insurance fund, the exhaustion of the insurance fund triggers auto-deleveraging, and auto-deleveraging in turn weakens the market's loss-absorbing capacity. After expressing this system as a chain of state transitions, the parts truly needing reinforcement become clear: not any single link but the connections between links.

Answering the three questions along these four seams yields four normative principles, which are also four structural positions not yet filled to date. The depth of the waterfall requires setting, before losses are allocated to non-defaulting parties, at least a skin-in-the-game own-equity first-loss layer and a prefunded mutualization layer, and the chain is absent at these two points, able only to turn from a fee public pool directly to forced socialization onto the profitable side. Orderly degradation requires designing for the worst case absent a lender of last resort, severing contagion with risk-pool isolation and confining the liquidation shock within an absorbable range with active shock management, while isolation and auction each carry unavoidable costs. The boundary of incentives and discretion requires handing the overwhelming majority of states to deterministic code and permitting discretion only in the minority of states the contract cannot provide for in advance, constraining residual discretion to be narrow, slow, and visible. Using these four principles as an evaluation standard to measure the most frontier Hyperliquid today, one finds it closest to met on the sub-item of *a single state machine recomputable and verifiable by anyone*, but still with distances identifiable item by item on the overall architecture, waterfall depth, orderly degradation, and the constraint of discretion.

Running through all three answers is the same fact: the hardest parts of a decentralized clearinghouse are mostly not technical problems. Making the loss waterfall deep is hard because of the lack of own equity willing to bear losses first; keeping the degradation mechanism orderly is hard because every cost must ultimately be borne by someone; constraining discretion to the minimum is hard because there will always be a moment when the code must allow an authorized governance actor to intervene. None of these is a problem a more complex contract can solve on its own; they point to how rules are written, executed, and governed. The four-item checklist this chapter gives is precisely the tool for measuring, item by item, the degree to which any real system meets the standard on these four dimensions. When compliance and governance are no longer externally imposed constraints but engineered directly into the protocol itself, in what

way this risk-governance architecture will operate is exactly the domain Chapter 30 (embedded compliance and automated governance) will enter.

References

[1] CoinDesk Research. (2025, October 17). *Market spotlight: Inside crypto's $19 billion liquidation event.* CoinDesk. https://www.coindesk.com/research/market-spotlight-the-19-billion-liquidation-that-shook-crypto

[2] Committee on Payment and Settlement Systems, & Technical Committee of the International Organization of Securities Commissions. (2012). *Principles for financial market infrastructures.* Bank for International Settlements & IOSCO. https://www.bis.org/cpmi/publ/d101a.pdf

[3] Aramonte, S., Huang, W., & Schrimpf, A. (2021). DeFi risks and the decentralisation illusion. *BIS Quarterly Review*, December 2021, 21–36. Bank for International Settlements. https://www.bis.org/publ/qtrpdf/r_qt2112b.htm

[4] Lamport, L. (1998). The part-time parliament. *ACM Transactions on Computer Systems, 16*(2), 133–169. https://doi.org/10.1145/279227.279229

[5] Castro, M., & Liskov, B. (1999). Practical Byzantine fault tolerance. In *Proceedings of the Third USENIX Symposium on Operating Systems Design and Implementation (OSDI '99)* (pp. 173–186). USENIX Association. https://www.usenix.org/conference/osdi-99/practical-byzantine-fault-tolerance

[6] Pirrong, C. (2011). *The economics of central clearing: Theory and practice* (ISDA Discussion Papers Series No. 1). International Swaps and Derivatives Association. https://www.isda.org/a/yiEDE/isdadiscussion-ccp-pirrong.pdf

[7] Cont, R. (2017). Central clearing and risk transformation. *Financial Stability Review, 21*, 127–138. Banque de France. https://publications.banque-france.fr/sites/default/files/medias/documents/fsr21_web.pdf https://doi.org/10.2139/ssrn.2955647

[8] Hyperliquid. (n.d.). *Hyperliquid documentation.* Retrieved June 25, 2026, from https://hyperliquid.gitbook.io/hyperliquid-docs

[9] European Parliament, & Council of the European Union. (1998). Directive 98/26/EC of 19 May 1998 on settlement finality in payment and securities settlement systems. *Official Journal of the European Communities, L 166*, 45–50. https://eur-lex.europa.eu/eli/dir/1998/26/oj/eng

[10] Garay, J., Kiayias, A., & Leonardos, N. (2015). The bitcoin backbone protocol: Analysis and applications. In *Advances in Cryptology – EUROCRYPT 2015* (pp. 281–310). Springer. https://doi.org/10.1007/978-3-662-46803-6_10

[11] Duffie, D., & Zhu, H. (2011). Does a central clearing counterparty reduce counterparty risk? *The Review of Asset Pricing Studies, 1*(1), 74–95. https://doi.org/10.1093/rapstu/rar001

[12] Chitra, T. (2025). *Autodeleveraging: Impossibilities and optimization* (arXiv:2512.01112, v3 revised 2026) [Preprint, not peer-reviewed]. arXiv. https://doi.org/10.48550/arXiv.2512.01112

[13] Paddrik, M., & Zhang, S. (2020). *Central counterparty default waterfalls and systemic loss* (OFR Working Paper No. 20-04). Office of Financial Research. https://www.financialresearch.gov/working-papers/files/OFRwp-20-04_central-counterparty-default-waterfalls-and-systemic-loss.pdf

[14] Merkle, R. C. (1988). A digital signature based on a conventional encryption function. In *Advances in Cryptology – CRYPTO '87* (pp. 369–378). Springer. https://doi.org/10.1007/3-540-48184-2_32

[15] Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review, 105*(2), 564–608. https://doi.org/10.1257/aer.20130456

[16] Committee on Payments and Market Infrastructures, & Board of the International Organization of Securities Commissions. (2017, July 5). *Recovery of financial market infrastructures: Revised report*. Bank for International Settlements. https://www.bis.org/cpmi/publ/d162.htm

[17] Financial Stability Board. (2017, July 5). *Guidance on central counterparty resolution and resolution planning*. https://www.fsb.org/2017/07/guidance-on-central-counterparty-resolution-and-resolution-planning-2/

[18] dYdX Foundation. (2023, November 1). *v4 technical architecture overview*. https://www.dydx.xyz/blog/v4-technical-architecture-overview

[19] Perpetual Protocol. (n.d.). *Perpetual Protocol v2 documentation*. Retrieved June 25, 2026, from https://docs.perp.com/

[20] GMX. (n.d.). *GMX documentation*. Retrieved June 25, 2026, from https://docs.gmx.io/

[21] Gains Network. (n.d.). *Gains Network documentation*. Retrieved June 25, 2026, from https://docs.gains.trade/

[22] Synthetix. (n.d.). *Synthetix documentation*. Retrieved June 25, 2026, from https://docs.synthetix.io/

[23] Liquity. (n.d.). *Liquity v1 documentation*. Retrieved June 25, 2026, from https://docs.liquity.org/liquity-v1

[24] Malwa, S. (2022, September 19). *DeFi trader nets over $500K by using DEX GMX to manipulate Avalanche token*. CoinDesk. https://www.coindesk.com/markets/2022/09/19/defi-trader-nets-over-500k-by-using-dex-gmx-to-manipulate-avalanche-token

[25] Behnke, R. (2025, July 15). *Explained: The GMX hack (July 2025)*. Halborn. https://www.halborn.com/blog/post/explained-the-gmx-hack-july-2025

[26] Okwatch, L. (2025, August 14). *GMX finalizes $44M payout to GLP holders affected by V1 exploit*. crypto.news. https://crypto.news/gmx-44m-payout-glp-holders-v1-exploit-2025/

[27] Egorov, M. (2022). *Curve stablecoin design* [Whitepaper]. Curve Finance. https://docs.curve.finance/developer/resources/whitepaper

[28] Malwa, S. (2024, June 13). *DeFi giant Curve roiled as founder's loans get liquidated; CRV slides 30%*. CoinDesk. https://www.coindesk.com/markets/2024/06/13/crv-

slides-30-as-loans-tied-to-curves-founder-face-liquidation-risk

[29] Rodrigues, F. (2026, April 27). *Curve founder pitches market-based fix for $700K bad debt in contrast to Aave bailout*. CoinDesk. https://www.coindesk.com/markets/2026/04/27/curve-founder-pitches-market-based-fix-for-usd700k-bad-debt-in-contrast-to-aave-bailout

[30] Aave. (n.d.). *AAVE token and the Safety Module* [Documentation]. Retrieved June 25, 2026, from https://aave.com/docs/ecosystem/aave

[31] Aave. (n.d.). *Umbrella* [Documentation]. Retrieved June 25, 2026, from https://aave.com/docs/aave-v3/umbrella

[32] Cointelegraph. (2025, April 28). *What happened to sUSD? How a crypto-collateralized stablecoin depegged*. Cointelegraph. https://cointelegraph.com/explained/what-happened-to-susd-how-a-crypto-collateralized-stablecoin-depegged

[33] Synthetix. (2026, June 12). *SIP-423: sUSD retirement & staking reform*. Synthetix Improvement Proposals. https://sips.synthetix.io/sips/sip-423/

[34] Nexus Mutual. (n.d.). *Nexus Mutual documentation*. Retrieved June 25, 2026, from https://docs.nexusmutual.io/

[35] InsurAce. (n.d.). *InsurAce protocol whitepaper*. https://files.insurace.io/public/InsurAceWhitepaper.pdf

[36] Benzinga. (2022, August 23). *InsurAce announces $12 million insurance payouts to 155 victims of Terra $UST crash*. Investing.com. https://m.uk.investing.com/news/cryptocurrency-news/insurace-announces-12-million-insurance-payouts-to-155-victims-of-terra-ust-crash-2730016?ampMode=1

[37] EigenLayer Team. (2023). *EigenLayer: The restaking collective* (Whitepaper v1.0). https://docs.eigencloud.xyz/assets/files/EigenLayer_WhitePaper-88c47923ca0319870c611decd6e562ad.pdf

[38] Symbiotic. (n.d.). *Symbiotic documentation*. Retrieved June 25, 2026, from https://docs.symbiotic.fi/

[39] EigenCloud. (2025, April 2). *Intro to slashing on EigenLayer: AVS edition*. https://blog.eigencloud.xyz/intro-to-slashing-on-eigenlayer-avs-edition/

[40] Heimbach, L., Schertenleib, E. G., & Wattenhofer, R. (2023). Short squeeze in DeFi lending market: Decentralization in jeopardy? In *Financial Cryptography and Data Security. FC 2023 International Workshops* (pp. 316–331). Springer. https://doi.org/10.1007/978-3-031-48806-1_22

[41] Knight, O. (2025, March 26). *Hyperliquid delists JELLYJELLY after vault squeezed in $13M tussle*. CoinDesk. https://www.coindesk.com/markets/2025/03/26/hyperliquid-delists-jellyjelly-after-vault-squeezed-in-usd13m-tussle

[42] Aave. (n.d.). *Aave V3 technical overview* [Documentation]. Retrieved June 25, 2026, from https://aave.com/docs/aave-v3/overview

[43] Euler Labs. (n.d.). *Euler v2 lite paper: The modular lending platform*. Retrieved June 25, 2026, from https://docs.euler.finance/euler-v2-lite-paper/

[44] Compound. (n.d.). *Compound III documentation*. Retrieved June 25, 2026, from

https://docs.compound.finance/

[45] Morpho Association. (2023). *Morpho Blue whitepaper*. https://github.com/morpho-org/morpho-blue/blob/main/morpho-blue-whitepaper.pdf

[46] Hyperliquid. (n.d.). *HIP-3: Builder-deployed perpetuals* [Documentation]. Retrieved June 25, 2026, from https://hyperliquid.gitbook.io/hyperliquid-docs/hyperliquid-improvement-proposals-hips/hip-3-builder-deployed-perpetuals

[47] MakerDAO. (2021). *MIP45: Liquidations 2.0 — Liquidation system redesign* [Maker Improvement Proposal]. GitHub. https://github.com/makerdao/mips/blob/master/MIP45/mip45.md

[48] Eichholz, L. (2020, March 17). *What really happened to MakerDAO?* Glassnode Insights. https://research.glassnode.com/what-really-happened-to-makerdao/

[49] Ajna Labs. (2024). *Ajna protocol whitepaper*. https://www.ajna.finance/pdf/Ajna_Protocol_Whitepaper_01-11-2024.pdf

[50] Kirillov, A., & Chung, S. (2022). *StableSims: Optimizing MakerDAO Liquidations 2.0 incentives via agent-based modeling* (arXiv:2201.03519). arXiv. https://arxiv.org/abs/2201.03519 https://doi.org/10.48550/arXiv.2201.03519

[51] Qin, K., Zhou, L., Gamito, P., Jovanovic, P., & Gervais, A. (2021). An empirical study of DeFi liquidations: Incentives, risks, and instabilities. In *Proceedings of the ACM Internet Measurement Conference* (pp. 336–350). ACM. https://doi.org/10.1145/3487552.3487811

[52] Sandor, K. (2022, November 22). *Mango exploiter's funds get liquidated after roiling Aave using $20M of borrowed Curve tokens*. CoinDesk. https://www.coindesk.com/markets/2022/11/22/mango-exploiter-gets-liquidated-after-roiling-aave-using-20m-of-borrowed-curve-tokens

[53] Maskin, E. S. (2008). Mechanism design: How to implement social goals. *American Economic Review, 98*(3), 567–576. https://doi.org/10.1257/aer.98.3.567

[54] Hart, O., & Moore, J. (1990). Property rights and the nature of the firm. *Journal of Political Economy, 98*(6), 1119–1158. https://doi.org/10.1086/261729

[55] Goodhart, C. A. E. (1984). Problems of monetary management: The U.K. experience. In *Monetary Theory and Practice* (pp. 91–121). Palgrave Macmillan. https://doi.org/10.1007/978-1-349-17295-5_4

[56] Lucas, R. E., Jr. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy, 1*, 19–46. https://doi.org/10.1016/S0167-2231(76)80003-6

[57] Grossman, S. J., & Hart, O. D. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy, 94*(4), 691–719. https://doi.org/10.1086/261404

Embedded Compliance and Automated Governance

In November 2022, the collapse of the cryptocurrency exchange FTX exposed a shortfall in customer funds of as much as $8.7 billion; the subsequent bankruptcy investigation and criminal trial revealed the complete failure of its internal controls and large-scale misappropriation of assets [1]. This event marked the structural bankruptcy of the trust model underlying centralized crypto-finance institutions. Its core lesson was not merely the isolated failure of a single institution, but the structural fragility of the traditional financial compliance paradigm when confronted with digital-native institutions that hold complete control over assets. Traditional financial compliance is essentially an external enforcement mechanism that relies on ex post penalties; it presupposes that an institution is physically and technically capable of violating the rules and that its incentive to comply derives solely from avoiding those penalties (its logic and premises are detailed in Section 30.1.1). When a centralized institution holds unrestricted private keys to assets and the power to modify databases, the lag inherent in external supervision means that no periodic compliance check can prevent the asset misappropriation or ledger tampering that has already occurred [2]. This systemic risk raises the central question of this chapter: how can regulatory compliance requirements and system governance mechanisms be embedded directly into the architecture of on-chain financial systems, thereby transforming compliance from an external mandate that depends on institutional self-discipline into a built-in guarantee that makes violations physically impossible at the system level?

Embedded compliance proposes a paradigm fundamentally different from traditional regulation. Its underlying logic is to use cryptography and smart-contract technology to design a system that, at the architectural level, excludes the possibility of specific violations [3]. Under this paradigm, rules are no longer legal provisions on paper or entries in a compliance manual; they are compiled directly into the execution logic of smart contracts. If the system requires that customer assets be strictly segregated, this rule is encoded as a contract function that permits withdrawals only when triggered by the customer's own private-key signature, making it technically impossible for the protocol developers or the operating team to misappropriate any assets. This design, which converts behavioral constraints into architectural constraints, achieves a paradigm shift from requiring institutions to obey the law to making it impossible for them to break it. Yet hard-coding rules into smart contracts is not a panacea. The complexity of financial systems requires a balance between the security of immutability and the flexibility to

adapt to changing markets: complete rigidity would leave a system unable to cope with extreme market volatility or protocol vulnerabilities, whereas excessive flexibility could reintroduce the single-point-of-failure and trust risks that administrator privileges entail. Building a robust on-chain financial system therefore requires careful engineering trade-offs in both the embedding depth of rules and the design of governance mechanisms. Crucially, embedded compliance rests on the premise that the correctness of smart-contract code can be guaranteed through means such as formal verification; if the code implementing a compliance rule itself contains a vulnerability, the credibility of the architectural constraint is fundamentally undermined.

This chapter takes an engineering-implementation perspective to systematically construct the theory and methodology for embedding compliance and governance into on-chain architecture. The three-tier compliance spectrum divides compliance mechanisms by embedding depth into hard-coded, parameterized, and reporting-based compliance. Hard-coded compliance applies to non-negotiable security baselines such as asset segregation; parameterized compliance handles operational parameters such as the margin ratio, which must be adjustable yet remain transparent; and reporting-based compliance leverages the openness of on-chain data to enable automated transaction monitoring and real-time risk disclosure. At the governance layer, the dual architecture of timelock and emergency powers strikes a balance between immutability and flexible response through a mandatory delayed review for routine changes and a constrained fast track for extreme situations. Finally, the chapter translates the regulatory frameworks and principles discussed earlier into a concrete inventory of on-chain technical implementations, completing the mapping from regulatory requirements to engineering code.

30.1 The theoretical foundations of embedded compliance

Traditional financial compliance relies on external enforcement to constrain institutional behavior, and this model has exposed structural limitations in cost, timeliness, and cross-border coordination. By embedding compliance rules into the execution logic of smart contracts, on-chain financial systems offer an alternative that shifts from behavioral constraints to architectural constraints. However, different types of compliance requirements differ significantly in how much embedding depth and flexibility they demand, which calls for engineering trade-offs across the three tiers of hard-coded, parameterized, and reporting-based compliance. This section first analyzes the limitations of the traditional model, then sets out the core concept of embedded compliance and the three-tier spectrum, and finally shows how macro-level regulatory principles can be mapped onto concrete on-chain implementation paths.

30.1.1 The limits of the traditional compliance model

The core logic of traditional financial compliance rests on the dual assumptions of trust and verification. Regulators set the rule framework, financial institutions pledge to comply with those rules, and internal compliance departments, external auditors, and regulators then verify compliance through periodic inspections. The fundamental premise of this model is that a financial institution is fully capable, physically and

technically, of violating the rules, and that it chooses to comply mainly out of fear of reputational damage and legal penalties. This model of external enforcement grounded in ex post sanctions, however, is exposing increasingly severe structural limitations.

The most conspicuous limitation lies in the ever-rising cost of compliance. In modern financial systems, compliance has become a high-cost operational burden. According to industry research, compliance costs account for roughly 10% to 20% of a financial institution's operating expenses, and the proportion is higher for some small community banks and regional financial institutions [4]. Since the 2008 global financial crisis, the operating costs that retail and corporate banks devote to compliance have risen by more than 60% [5]. These high costs are ultimately passed on to end consumers through higher service fees and lower deposit rates. In addition, global spending on the anti-money laundering (AML) systems (software and systems) market is projected to reach $51.7 billion by 2028 [6]. Despite the enormous investment, this resource-intensive compliance model has not fully prevented financial crime or systemic risk. To be sure, the high cost of the traditional compliance system is not entirely wasted: macroprudential regulatory frameworks have demonstrable empirical effects in reducing the frequency of systemic crises and limiting their reach. The improvement that embedded compliance offers lies mainly in reallocating the wasteful overhead of "compliance theater" in the traditional model into architectural-level, built-in guarantees, rather than simply replacing every traditional compliance function.

A second structural flaw of the external-enforcement model is its lack of timeliness. Traditional compliance checks are typically based on quarterly reports or annual audits, and such low-frequency, snapshot-style reviews cannot capture violations that occur between two checks. By the time an auditor or regulator discovers a problem, the violation has often persisted for months or even years, and the resulting damage may already be irreparable. The FTX case is an extreme manifestation of this limitation. Although the institution claimed to have strict internal controls, when its de facto controller decided to misappropriate customer assets, no periodic inspection mechanism was able to intercept the behavior before or as it occurred.

This lag further gives rise to the phenomenon of "compliance theater." At many financial institutions, the purpose of compliance activity has gradually shifted from "effectively reducing risk" to "passing regulatory inspections" [4]. Institutions devote substantial resources to filling out forms, generating reports, and designing complex compliance workflows; these activities superficially satisfy regulatory requirements but do not substantively change the business's risk exposure. When compliance becomes a formalistic exercise aimed at meeting regulatory metrics, it not only wastes social resources but also manufactures a false sense of security.

The compliance predicament of centralized cryptocurrency exchanges pushes these structural flaws to an extreme. Unlike traditional banks, the core business of a centralized exchange (CEX) runs entirely on digital infrastructure: order matching is completed on off-chain centralized servers, customer assets are stored in hot and cold wallets whose private keys the exchange controls, and the internal ledger exists as database records that administrators can modify at will. This architecture means that external auditors cannot verify in real time the consistency between the asset status the

exchange claims and the actual on-chain address balances. Even a periodic financial audit is effective only to the extent that the exchange voluntarily provides complete and truthful data, and the collapse of FTX is precisely an extreme confirmation of this architectural flaw.

The fragmentation of cross-border compliance makes these problems even more complex. In globally operating financial markets, different jurisdictions have starkly different and sometimes conflicting compliance requirements. A global financial institution must satisfy dozens of different regulatory frameworks simultaneously, which not only causes compliance costs to grow exponentially but also leaves institutions stretched thin across regulatory regimes, unable to build a unified and coherent risk-control system.

30.1.2 The core concept of embedded compliance

Confronted with the structural failure of the traditional compliance model, on-chain financial systems propose a paradigm shift: embedded compliance. The core of this concept is to convert compliance from a "behavioral constraint" into an "architectural constraint." Under the behavioral-constraint model, rules tell participants "what they should not do," but participants still retain the physical ability to do it; under the architectural-constraint model, the system is designed so that violations become physically and technically impossible.

This concept can be likened to the evolution of building fire codes. Modern building codes no longer rely solely on periodic inspections by fire marshals; instead, they fix firewall materials and sprinkler coverage in the building's blueprint at the design stage, so that the physical structure of the building excludes the possibility of specific fire-code violations. Embedded compliance follows the same engineering logic, moving constraints forward into the system design stage.

In on-chain financial systems, the smart contract serves as the execution engine of this concept. Once a smart contract is deployed to a blockchain, its logic executes strictly according to its code. Code execution is unaffected by conflicts of interest or operational negligence; its output is determined entirely by its inputs and logic. If we encode compliance rules as part of a smart contract, those rules become an inseparable property of the system, giving concrete form, in the domain of financial regulation, to Lawrence Lessig's idea that "code is law" [3][7][8]. When a transaction is broadcast to the network, the system verifies not only its technical validity but also, at the compliance level, whether it is permitted. If a transaction violates the preset compliance logic, the smart contract simply refuses to execute it, making the violation impossible to occur on-chain in the first place.

This shift from ex post detection to ex ante prevention means that compliance is no longer an add-on process separate from the core business but an intrinsic feature of the financial infrastructure itself. It substantially reduces the reliance on human intervention found in traditional compliance, lowers the ongoing operational cost of compliance, and raises the certainty of compliance from the probabilistic level to the level of cryptographic guarantee.

Embedded compliance, however, is not without limits. First, not all regulatory rules are

suitable for, or capable of, being translated into deterministic code logic. Many compliance requirements contain fuzzy boundaries and room for subjective judgment. Second, code itself may contain vulnerabilities; if compliance logic is incorrectly encoded, it may cause legitimate transactions to be rejected or non-compliant transactions to be admitted. More complex still, the regulatory environment and market conditions change continuously, and if rules are hard-coded into the system in a completely rigid form, the system loses the ability to adapt to future change. A deeper limitation is that although embedded compliance eliminates human bias at the execution level, the code itself is written by developers who carry cognitive biases; availability bias, for example, may lead to over-guarding against recently known risks while ignoring unknown ones, and hard-coding a biased judgment into immutable logic effectively locks that bias in—converting it from a "correctable operational decision" into an "immutable system architecture." Moreover, in the traditional compliance framework, legal liability for violations is clearly attributed to the financial institution; once compliance rules are embedded in a smart contract, if the contract's behavior produces a law-violating outcome, the attribution of liability becomes highly ambiguous among the contract deployer, the code developers, the auditing firm, and the governance voters. This "accountability gap" is one of the core practical obstacles to regulators' acceptance of embedded compliance. These limitations raise a key question: how can the rigidity of rules be preserved while endowing the system with the necessary capacity to evolve?

30.1.3 The three-tier compliance spectrum

To resolve the tension between flexibility and certainty in embedded compliance, this chapter proposes the theory of the "three-tier compliance spectrum." This theory holds that compliance requirements should not be hard-coded indiscriminately; rather, compliance mechanisms should be divided into three distinct tiers by "embedding depth," according to the nature of the rule, the consequences of violating it, and the need for flexibility.

These three tiers each strike their own balance between flexibility and certainty, and together they constitute a complete on-chain compliance framework.

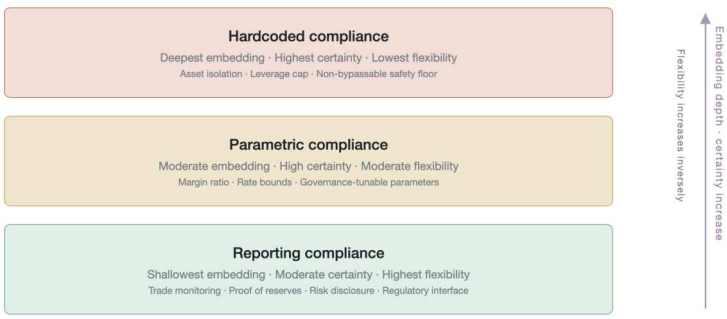


Figure 30-1. The three-tier compliance spectrum and its relationship to embedding depth, flexibility, and certainty [3]

Hard-coded compliance represents the deepest embedding and the highest certainty. In this tier, compliance rules are written directly into the underlying logic of the smart contract, with no interface for modification. Such rules typically concern the system's security baseline; their defining features are that they are "non-negotiable" and that a single violation would lead to severe consequences. Asset segregation, for example, must be hard-coded. The system must ensure through code that, apart from the user's own cryptographic-signature authorization and liquidation events that satisfy strict trigger conditions, no administrator key or protocol backdoor can move a user's funds. An absolute leverage cap is another typical application of hard-coding: by setting an insurmountable maximum leverage multiple at the code level, the system effectively prevents the emergence of extreme risk exposure. The cost of hard-coded compliance is a significant reduction in flexibility, as any modification of the rules requires a complex contract upgrade.

Parameterized compliance sits in the middle of the spectrum, offering a compromise in which "the rule framework is fixed but the specific parameters are adjustable." Many operational compliance rules must be dynamically adjusted in response to market volatility and changes in regulatory policy. The margin ratio, for example, must rise as market volatility increases in order to absorb larger potential losses. Under parameterized compliance, "margin must be collected" and "the formula for computing the margin ratio" are hard-coded, but the specific coefficients within the formula are parameterized. These parameters are stored in a dedicated registry contract and can be modified through an established governance process. This design both guarantees the stability of the compliance framework and endows the system with the flexibility to adapt to environmental change.

Reporting-based compliance represents the shallowest embedding. Not all compliance requirements are suited to serving as ex ante restrictions. For behaviors that require holistic judgment or ex post tracing—such as detecting abnormal trading patterns, identifying wash trading, and preventing market manipulation—forcibly intercepting transactions before execution could seriously impair system throughput and the trading experience of legitimate users. Reporting-based compliance exploits the full transparency and immutability of blockchain data, exporting on-chain data in real time and feeding it into analytical models to achieve automated ex post monitoring. In addition, real-time risk disclosure and proof of reserves belong to this tier; they do not restrict user behavior but satisfy the regulatory demand for transparency by providing ample information visibility.

These three tiers are not mutually exclusive but highly complementary. Hard-coded compliance guards the system's untouchable security baseline, parameterized compliance provides the elasticity needed for day-to-day operations, and reporting-based compliance supplements the ongoing monitoring of complex behaviors. No single tier can cover all compliance requirements; only by organically combining the three can one build an on-chain financial system that is both secure and sustainable.

30.1.4 From regulatory principles to engineering implementation

In Chapter 26, we proposed six core regulatory principles applicable to decentralized finance (DeFi): asset segregation as the baseline, disclosure first, functional regulation, embracing embedded compliance, combining international minimum standards with flexible adjustment, and evidence-based evolution. These macro-level principles constitute the requirements specification for on-chain compliance engineering, and the three-tier compliance spectrum provides a systematic implementation path from "what should be regulated" to "what is technically guaranteed." The next three sections respectively develop hard-coded (Section 30.2), parameterized (Section 30.3), and reporting-based compliance (Section 30.4), and Section 30.7 closes the loop from regulatory requirement to engineering code with a complete mapping table.

30.2 Hard-coded compliance

In the architectural design of embedded compliance, hard-coding represents the deepest constraint. The violation of certain rules would lead to systemic catastrophe, so those rules should possess no flexibility. The core logic of hard-coded compliance is to write these non-negotiable baselines directly into the immutable logic of the smart contract, so that no one, under any circumstances, can physically circumvent them.

30.2.1 When hard-coding applies

The core criterion for judging whether a rule should be hard-coded is whether the consequences of violating it are severe and irreversible. If a violation can inflict serious damage on the system before it is discovered, ex post punishment and compensation lose their meaning. In such cases, making violations physically impossible ex ante through code becomes the only engineering option.

Rules suited to hard-coding typically possess broad universality and a low update frequency. The asset-segregation mechanism is the most typical scenario for hard-coding. Customer assets must be independent of protocol operating funds—a baseline that does not change under any market environment or regulatory framework. If a protocol operator or administrator were allowed to withdraw user assets through some privilege, the system would revert to the traditional trust model, whose fragility the FTX case demonstrated. The absolute leverage cap likewise falls within the scope suited to hard-coding. When leverage exceeds a certain critical value, even a minuscule price movement can trigger liquidation, which in turn causes a liquidation cascade. This systemic risk should not be left to dynamic adjustment by operational parameters but must be capped by an insurmountable constraint at the code level [9]. The conditions for using the insurance fund must also be hard-coded: its funds may be used only to absorb liquidation losses, and any attempt to divert them to other uses should be automatically rejected at the contract level.

By contrast, many operational rules are not suited to hard-coding. The margin ratio must be dynamically adjusted according to market volatility, the determination of abnormal trades depends on complex pattern recognition and subjective judgment, and specific identity-verification standards evolve with the legal developments of different

jurisdictions [10]. Hard-coding rules that require flexible response would trap the system in excessive rigidity, ultimately leaving it unable to meet the market's real needs. The scope of hard-coding must be strictly limited to the core logic that defines the system's baseline for survival.

30.2.2 Hard-coding asset segregation

In the traditional financial system, asset segregation relies on complex institutional design. Take the Depository Trust & Clearing Corporation (DTCC): it ensures the segregation of customer assets from the institution's own funds through strict legal contracts, periodic external audits, and a multi-tiered account structure. Although this segregation mechanism grounded in institutional trust has proven highly reliable over the past 200 years, its essence remains guarding against violations rather than eliminating the ability to violate [11]. When this trust model is transplanted to an offshore cryptocurrency exchange that lacks strong regulation, its inherent fragility is fully exposed.

The asset-custody model of centralized cryptocurrency exchanges amplifies this fragility. Mainstream CEXs commonly adopt an omnibus wallet architecture, pooling all customer assets into a small number of on-chain addresses controlled by the exchange, while individual customer balances exist only as mappings in the exchange's internal database. This means that at the on-chain level, an external observer cannot distinguish the exchange's own assets from customer assets, let alone verify the accuracy of the internal ledger. The exchange holds complete control over the private keys of its hot and cold wallets, order matching is completed off-chain, and asset transfers require no customer authorization. The FTX case is precisely an extreme manifestation of this architectural flaw: Alameda Research was able to directly deploy FTX customer assets for its own trading, while external audits failed to uncover this fact for as long as two years. This model, in which a single entity simultaneously controls asset custody and ledger records, stands in fundamental contrast to the self-custody architecture of on-chain finance.

On-chain financial systems achieve a paradigm shift in asset segregation through smart contracts. The stablecoins or other collateral that users deposit are locked in an on-chain margin contract. The logic of this contract is strictly delimited, permitting changes in the margin balance only under four conditions explicitly defined in code. First, the user cryptographically signs a withdrawal transaction with their own private key. Second, when the user's position health falls below the liquidation threshold, the liquidation engine triggers the forced-liquidation logic. Third, periodic settlement transfers are executed automatically between longs and shorts according to the funding rate formula. Fourth, when the insurance fund is insufficient to cover shortfall losses, the auto-deleveraging (ADL) mechanism forcibly and partially closes the positions of profitable counterparties. Beyond these code-defined liquidation and trading operations (and analogous code-defined paths such as the deduction of trading fees and the settlement of realized profit and loss upon position closure), no discretionary administrative instruction can move a user's assets; neither the protocol development team, nor operational administrators, nor any third party holds a privileged interface capable of

bypassing these paths. These guarantees carry one caveat: when this logic is deployed in an upgradeable proxy contract, they amount only to "version-level hard-coding," whose strength depends on the constraint mechanisms governing the upgrade authority itself (see Section 30.6.3).

Figure 30-2 presents this smart-contract-based asset-segregation architecture: user assets are locked in a non-custodial margin contract, and the trigger conditions and execution scope of all four asset-movement paths are explicitly defined in code.

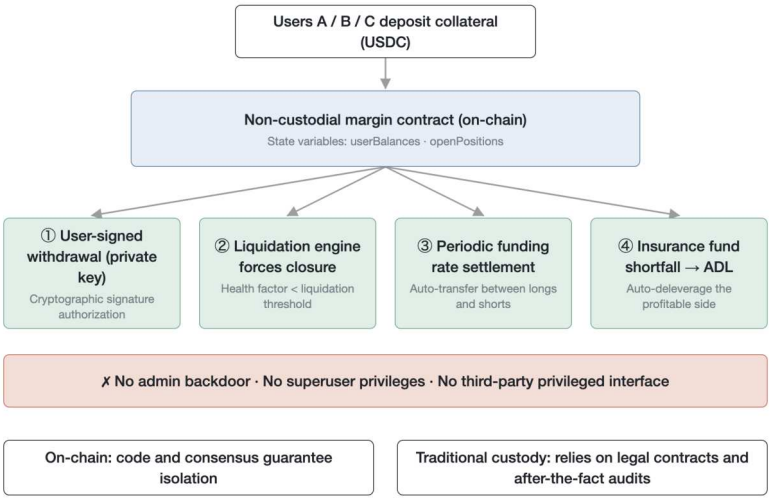


Figure 30-2. The hard-coded design architecture for asset segregation [11]

The core constraint of this design is that it eliminates administrator privileges in the traditional sense. Even the protocol development team that deploys the contract cannot use a backdoor key or superuser privilege to withdraw a user's margin. The security of assets no longer depends on the team's moral integrity or on legal constraints but is guaranteed by the consensus mechanism and cryptographic primitives of the underlying blockchain, such as Ethereum.

The effectiveness of code-level asset segregation rests on a series of operational premises. On the user side, private-key security is the first line of defense: if a user's private key is stolen through a phishing attack, the attacker can withdraw all of the user's assets through a completely "legitimate" signature path, against which the hard-coded segregation logic is powerless. On the system side, the liveness guarantee of the liquidation engine is a critical operational dependency of the segregation mechanism: if the liquidation bots go down, if network congestion prevents liquidation transactions from being confirmed on-chain in time, or if a spike in gas prices makes liquidation unprofitable, the liquidation path fails and the system may incur shortfalls. The reliability and timeliness of oracle price feeds constitute an equally critical premise. In addition, there is a significant gap between "having the right to withdraw" and "being able to withdraw in a timely manner": network congestion may prevent a withdrawal

transaction from being confirmed for a long time, a suspension of cross-chain bridging may block assets from returning to the mainnet, and an emergency-pause mechanism may temporarily restrict an individual user's withdrawals even as it protects the system. Hard-coded compliance eliminates the risk of an administrator misappropriating assets, but it cannot cover these operational-level failure modes.

This purely code-level segregation, however, faces significant challenges in engineering practice. The largest risk exposure often lies in the cross-chain bridging step. In app-chain architectures such as Hyperliquid, a user's native assets (such as USDC on the Ethereum mainnet) typically must be locked through a bridge contract in order to mint an equivalent synthetic asset on the app-chain. A series of major bridge-security incidents between 2022 and 2024 revealed several attack paths: a flaw in a bridge contract's verification logic may allow an attacker to forge cross-chain messages; the leakage of validator private keys or a social-engineering attack can directly seize control of the bridge (in the 2022 Ronin incident, for example, the attacker gained signing control over 5 of 9 validators—four Sky Mavis validator keys were stolen, plus the abuse of a third-party validator's authorization—causing losses of approximately $625 million); relayer spoofing may bypass message-verification mechanisms (as in the message-verification bypass in the Nomad incident); and if an attacker obtains the upgrade authority over a bridge contract, they may fundamentally alter its asset-transfer logic. Any one of these attack vectors could allow an attacker to move assets without a user's signature, thereby circumventing the segregation constraints of the margin contract. For an architecture like Hyperliquid's, which uses a "native bridge" that shares the same validator set as the consensus layer, bridge security and consensus security are fully coupled: the compromise of the validator set means the simultaneous failure of both the bridge and consensus. This risk-concentration property differs fundamentally from a general-purpose cross-chain bridge that uses an independent validator set.

In addition, the upgrade authority over the margin contract itself constitutes another layer of risk. If the contract retains an upgrade authority, the entity holding that authority could, in theory, lift the segregation constraints by replacing the contract logic. This makes the management of the upgrade authority the most sensitive trust assumption in a hard-coded compliance system.

Hard-coded asset segregation also faces a structural challenge arising from the composability of smart contracts. In an on-chain financial ecosystem, when the assets or derivative certificates in a margin contract can be referenced by other protocols or used as collateral, interactions between contracts may circumvent the segregation constraint at the economic level, even if the code logic of each individual contract is correct. System designers face a trade-off between "closed segregation" and "open segregation": the former maintains the strictest segregation guarantee by prohibiting margin assets from participating in any external contract interaction, but sacrifices capital efficiency; the latter permits limited cross-protocol interaction subject to additional constraints, but introduces a new risk dimension in the complexity of security auditing.

Moreover, the discussion of smart-contract asset segregation above implies an important distinction: on a general-purpose public chain such as Ethereum, the logic of a margin contract can be independently audited and formally verified by any third party; but in

an app-chain architecture such as Hyperliquid, the core margin-management logic runs in the consensus-layer state machine rather than in an independently auditable smart contract. This means that the trust assumption for asset segregation shifts from "the correctness of the code logic" to "an honest-majority assumption among validators," and in a network with only about two dozen active validators, this consensus security is far below the security guarantee that the thousands of validators on the Ethereum mainnet provide. The "architectural-level guarantee" described here therefore needs to be qualified according to the consensus-security model of the specific system, rather than treated as a universal conclusion about on-chain finance.

30.2.3 Hard-coding the leverage cap

Leveraged trading is a core function of derivatives markets, but excessive leverage is a primary source of market fragility. In decentralized perpetual futures, if traders are allowed to use extreme leverage (such as 500x or 1,000x), the bankruptcy price of their position lies extremely close to the entry price. Under the constraints of block-confirmation latency on a blockchain network and the discreteness of oracle price updates, the liquidation engine simply cannot complete a forced liquidation before the position's net value reaches zero, which inevitably leads to the rapid depletion of the insurance fund and system shortfalls [12].

Hard-coding the leverage cap is the most direct means of guarding against this kind of systemic risk. In the smart contract's position-opening logic, the system computes the ratio of the requested notional value to the user's available margin. The code explicitly stipulates that if this ratio exceeds a preset absolute cap (for example, 50x), the transaction is automatically reverted. This constraint requires no real-time approval from a risk-control team and does not depend on alerts from a post-trade monitoring system. When a transaction is submitted to the blockchain network, the deterministic execution logic of the Ethereum Virtual Machine (EVM) ensures that any request violating the leverage cap cannot alter the system's global state.

In actual engineering design, the hard-coded absolute cap is typically nested with a parameterized dynamic cap. The hard-coded cap sets a physical ceiling that the system cannot exceed under any extreme circumstances, while the dynamic cap is adjusted in real time according to the current market's liquidity depth, the asset's volatility characteristics, and the user's position size. For example, the contract code might hard-code an absolute cap of 100x, but during periods of high volatility a governance mechanism or automated risk engine would adjust the dynamic cap actually enforced down to 20x. This nested design both preserves the flexibility to respond to market changes and, through hard-coding, establishes an absolute baseline for the system's survival.

The calibration of a hard-coded leverage cap should not be an arbitrary number but should establish a quantitative constraint relationship with the latency parameters of the underlying infrastructure: the maximum leverage must ensure that the liquidation engine completes forced liquidation before the position's net value reaches zero. If the oracle's maximum update latency is 2 seconds and the maximum expected price movement of a major asset within that window is 0.5%, then the critical point

at which the price move exactly exhausts all margin corresponds to about 200x leverage (that is, $1/0.5\% = 200$). But 200x is merely the absolute cap at which the position's net value reaches zero, not a "safe" level of leverage: the usable buffer is in fact the initial margin minus the maintenance margin, and one must further deduct the liquidation delay caused by network congestion and the execution slippage caused by evaporating liquidity. Safe leverage must therefore satisfy safe leverage $<$

$$\frac{1}{\text{maintenance margin ratio} + \text{max adverse move within latency} + \text{execution slippage} + \text{congestion}}$$

and its actual value is far below 200x—which also explains why the earlier examples set the hard-coded absolute cap in the 50x–100x range. In addition, the hard-coded check should not be limited to the initial verification at the moment of opening a position: in a unified margin account, the unrealized loss on one position reduces the account's available margin, causing the effective leverage of other positions to rise implicitly. Even if every position passed the leverage check when opened, the system must continuously verify at each mark-price update whether the effective leverage of the entire account still lies within the hard-coded cap. Finally, a hard-coded leverage cap has one fundamental limitation: it constrains only the leverage within a single protocol. In a composable on-chain financial ecosystem, a trader can open same-direction positions across multiple protocols simultaneously, achieving effective leverage far exceeding the cap of any single protocol. This cross-protocol synthetic leverage cannot be resolved through hard-coded compliance; it requires ecosystem-level leverage tracking at the reporting-based compliance layer for monitoring and early warning.

30.2.4 The cost of hard-coding

Hard-coding rules into a smart contract provides cryptographic-level execution certainty, but this certainty comes at the cost of extreme system rigidity. The immutability of the underlying blockchain architecture means that once a smart contract is deployed, the bytecode it contains resides permanently on the distributed ledger. If a hard-coded rule needs to be modified—even by changing a single small constant value—a completely new smart contract must be deployed, and all users and liquidity must be migrated to it. This migration process is not only costly but also causes severe fragmentation of liquidity.

Over-reliance on hard-coding leaves a system unable to adapt in rapidly evolving financial markets. The microstructure, asset types, and attack vectors of cryptocurrency markets are all constantly changing. A hard-coded rule that appears perfect in the current market environment may, a year later, become a serious constraint that impedes the protocol's development. Only underlying security logic that has withstood the test of time and possesses broad applicability is therefore qualified to be hard-coded.

Furthermore, the security of hard-coding rests entirely on the correctness of the code. If the hard-coded logic contains an implementation vulnerability, a rule intended to protect the system may be exploited by an attacker. In complex DeFi protocols, interactions among smart contracts can produce unforeseeable edge cases. Traditional software-testing methods cannot cover all possible execution paths, which makes formal verification a necessary complementary discipline for hard-coded compliance [13].

Formal verification abstracts a smart contract into a mathematical model and uses logical reasoning to prove whether it satisfies specific properties (such as safety and liveness). Only hard-coded rules that have been rigorously proven mathematically can truly provide a security guarantee that surpasses traditional institutional trust. The engineering solution to this cost of rigidity—the proxy contract pattern and governance constraints—is developed systematically in Section 30.6.

30.3 Parameterized compliance

Most operational compliance rules are not suited to hard-coding, because they must be dynamically adjusted as market conditions and regulatory requirements change. Parameterized compliance offers a compromise in which the rule framework is fixed while parameters can be adjusted through governance, endowing a protocol with the flexibility to adapt to a complex financial environment while keeping its basic logic immutable.

30.3.1 When parameterization applies

Whereas hard-coded compliance applies to security baselines that must never be violated, the applicable boundary of parameterized compliance lies in operational metrics that must be dynamically optimized across different market cycles. The core criterion for judging whether a rule is suited to parameterization is whether its mathematical framework or logical structure is stable while its specific values must respond to changes in the external environment. When market volatility rises sharply, a fixed margin ratio may fail to cover the potential liquidation-slippage risk; when the efficiency of cross-chain arbitrage improves, the settlement frequency of the funding rate may need to be increased accordingly. These scenarios require the system to have a capacity for self-regulation, but that regulation must occur within a preset framework, and the adjustment process must be fully transparent to all participants.

The applicable scenarios for parameterized compliance fall into five categories. The first is the margin ratio: as discussed in Section 29.3, the maintenance margin requirement must vary dynamically with asset volatility and position size in order to guard against liquidation-cascade risk under extreme market conditions [14]. The second is rate parameters, including the settlement frequency, upper and lower bounds of the funding rate, and the size of the risk buffer pool; these parameters directly affect the holding costs of longs and shorts and the protocol's solvency [15]. The third is liquidation parameters, such as the trigger line and execution speed of progressive liquidation, which determine the degree of market impact during deleveraging. The fourth is market-making obligations: an on-chain order book or automated market maker pool must set a minimum liquidity depth, a maximum permissible spread, and obligation-waiver conditions under extreme market conditions in order to ensure the continuous operation of the market. The fifth is access parameters, which concern the compliance requirements of different jurisdictions, such as trading restrictions on specific assets or tiered standards for risk limits.

Hard-coding these operational parameters would leave the system rigid and without

means to respond to sudden events; leaving them entirely to arbitrary modification by off-chain administrators would undermine the trustless foundation of DeFi. Parameterized compliance resolves this contradiction through a parameter-registry mechanism in the smart contract. The rule's execution logic is written permanently into the code, while the specific values are stored in a separate registry contract. Any modification to a value in the registry must go through an explicit governance process and is subject to a timelock, thereby achieving an engineering optimum between flexibility and transparency [16].

The engineering design of the parameter registry follows the principle of separation of concerns. In a traditional monolithic-contract architecture, parameters and execution logic are intermingled, and modifying a parameter often requires redeploying the entire contract—which is not only costly but also prone to introducing new vulnerabilities. In a parameterized architecture, parameters are stored in a dedicated registry contract responsible only for data management and access control, without complex business logic. The execution contract reads parameters from the registry through a standard interface, much like querying a database. This separation allows parameter updates to be carried out independently without affecting the security of the core trading logic. At the same time, the history of parameter changes is preserved completely on the blockchain, forming an immutable audit trail, so that any regulator or user can trace the evolution of the protocol's parameters and verify its compliance.

30.3.2 Parameterizing the margin ratio

The margin ratio is the core line of defense in the risk management of a derivatives protocol. In an on-chain environment, setting the margin ratio is not merely an economic question but a compliance-engineering one. A traditional static margin ratio may cause low capital efficiency during low-volatility periods and fail to provide an adequate buffer during high-volatility periods, and this mismatch is an important structural cause of liquidation cascades [17]. To address this problem, on-chain protocols have introduced a volatility-adaptive margin-computation model and nested the regulatory baseline and the dynamic-adjustment mechanism within the smart-contract architecture.

Figure 30-3 presents the three-tier nesting of margin parameterization—a hard-coded regulatory baseline, a fixed adaptive-formula framework, and governance-adjustable parameter coefficients—which the following text develops tier by tier.

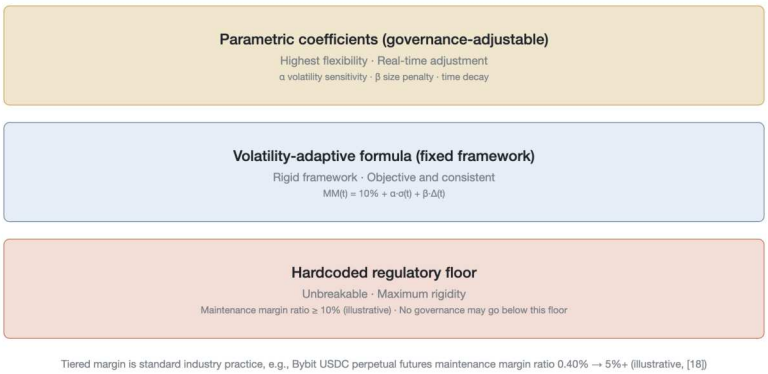


Figure 30-3. The three-tier nested design of margin parameterization [14][18] (the 10% maintenance-margin-ratio floor shown is illustrative)

The bottom of this design is the hard-coded regulatory baseline. At the code level, the smart contract mandates an absolute minimum for the maintenance margin ratio—for example, 10%. No matter how calm the market or how the governance committee votes, the system refuses to execute any operation that would drive the margin ratio below this floor. This hard-coding ensures that the protocol always meets the strictest regulatory requirements, eliminating the possibility that human negligence or malicious manipulation could create systemic risk.

Above the floor is the adaptive-formula layer. The formula for computing the margin ratio is hard-coded into the contract and typically comprises a base rate, a volatility multiplier, and a position-size penalty term. For example, the maintenance margin requirement rises in tiers as the notional value of a position increases; when position value increases from $100,000 to $1 million, the required margin ratio may automatically rise from 0.4% to 0.5% [18]. This logical framework is fixed, ensuring the objectivity and consistency of risk assessment.

In fact, the tiered margin mechanism is not an on-chain invention; mainstream centralized exchanges have long widely adopted a similar tiered margin structure. Take Bybit: the maintenance margin ratio on its USDC perpetual futures rises from 0.40% at the first tier to more than 5.00% at the higher tiers [18]. The low-tier maintenance margin ratios among these CEX tier parameters (e.g., 0.40%, nominally corresponding to maintenance leverage of about 250x, i.e., $1/0.40\% = 250$) fall below the initial-margin level implied by the leverage cap derived from oracle latency in Section 30.2.3 (the maintenance margin ratio and the latency buffer are different measures and should not be equated directly), and the hard-coded floor of an on-chain system constrains the lower bound of the tier parameters, ensuring that even the lowest-tier margin ratio is not below the system's security baseline. Binance, OKX, and other exchanges adopt similar tiered systems, while Deribit has introduced a portfolio margin model in its options market that dynamically computes net risk exposure through hedging relationships. However, these CEX tiered margin systems have one fundamental difference: the authority to change parameters lies entirely and unilaterally with the platform.

An exchange can adjust tier thresholds and margin ratios without any public-notice procedure, and users are informed only after the parameters take effect. By contrast, an on-chain system ensures the process transparency and auditability of every margin-parameter adjustment through a parameter registry and a governance-voting mechanism, which constitutes the core advantage of on-chain parameterized compliance over CEX dynamic-margin systems.

The top layer is the real-time adjustment of parameterized coefficients. The multipliers and thresholds in the adaptive formula are stored as parameters in the registry contract. When a macroeconomic event triggers a structural change in the market, the governance mechanism can adjust these coefficients so that the margin model captures risk more sharply. For example, during a period of sharply rising expected volatility, the volatility multiplier can be raised to tighten leverage system-wide. This three-tier nested architecture allows regulators to be confident of the system's security floor, allows traders to know clearly the computational logic of margin requirements, and leaves protocol operators the tools they need to respond to unknown risks.

In practice, the parameterized coefficients of a volatility-adaptive formula typically span three dimensions. The first is the volatility-sensitivity coefficient, which controls how much the margin ratio increases for each 1% rise in market volatility. This coefficient might be set at 0.15 in a calm market, but during a period of persistently elevated historical volatility the governance committee might vote to raise it to 0.25 in order to guard against a potential liquidity crisis in advance. The second is the position-size penalty coefficient, which reflects the systemic-risk contribution of large positions to the market. When an individual trader's position size exceeds a certain threshold of market liquidity, this coefficient is automatically triggered, requiring a higher margin ratio. The third is the time-decay coefficient, used to gradually loosen margin requirements after the market returns to calm, avoiding excessive liquidity tightening. The calibration of these coefficients should be based on a target coverage probability—that is, ensuring that, at a given confidence level, the margin can cover the maximum expected price movement over the period from liquidation trigger to completed closure—rather than being set arbitrarily by experience. The combined use of these coefficients forms a multidimensional risk-assessment model capable of capturing both changes in market microstructure and macroeconomic shocks.

While improving capital efficiency, the volatility-adaptive margin model introduces a systemic risk that cannot be ignored: procyclicality. When market volatility rises, the adaptive model automatically raises the margin ratio, forcing leveraged traders to add margin or face liquidation. Large-scale liquidations themselves further intensify market volatility, pushing volatility metrics still higher, which tightens the margin ratio further; this positive-feedback spiral of "rising volatility $\rightarrow$ tightening margin $\rightarrow$ forced liquidation $\rightarrow$ further rising volatility" is the core driving mechanism of a liquidation cascade. During Black Thursday in March 2020, this effect was amply demonstrated empirically in DeFi liquidations. To mitigate procyclicality risk, the engineering design of margin parameterization should introduce a countercyclical buffer mechanism: the margin ratio should respond quickly to rising volatility, but its relaxation should be slow and subject to a "look-back period" constraint (asymmetric decay); the magnitude and rate of a single margin-ratio adjustment should be capped to prevent violent tightening

within a short time; and during low-volatility periods, the margin ratio should be preemptively set above the model's raw output to reserve a buffer against future volatility shocks.

30.3.3 Parameterizing rate parameters

The funding rate mechanism is the core engine by which perpetual futures anchor to the spot price. Its design goal is to correct the deviation between the derivative price and the spot index through periodic funding exchanges between longs and shorts. In an on-chain implementation, the funding rate's computation logic is typically hard-coded, but its operational efficiency and risk boundaries depend heavily on several key parameterized settings.

The settlement frequency determines arbitrageurs' capital-turnover efficiency and the market's responsiveness. Traditional centralized exchanges typically adopt a discrete settlement mechanism every 8 hours, whereas in on-chain protocols, constrained by block-generation time and transaction costs, the parameterized choice of settlement frequency faces different engineering constraints [19]. A higher settlement frequency can anchor the price more tightly but significantly increases the computational overhead of oracle price feeds and state updates. Parameterizing the settlement frequency allows a protocol, as the underlying public chain scales or Layer 2 costs fall, to transition smoothly through governance voting toward higher frequencies and even per-block settlement (in a discrete blockchain architecture, per-block settlement is the finest granularity physically achievable) without restructuring the core trading logic. The effect of shortening the settlement interval is not linear: while a higher settlement frequency improves price-anchoring precision, it also raises the capital-efficiency requirements of arbitrage strategies, accelerates market makers' hedging-frequency needs, and may increase the volatility of each individual rate value—second-order effects that must be fully evaluated when adjusting the parameter.

Parameterizing the upper and lower bounds of the rate serves to prevent systemic collapse under extreme one-sided market conditions. When the spot price fluctuates violently, the theoretical funding rate may spike to extreme levels, subjecting market makers or hedgers to unbearable cash-flow pressure. The smart contract sets an absolute upper and lower bound on the funding rate through parameterization; this band is wide enough under normal conditions to accommodate normal arbitrage space but acts as a circuit breaker when an extreme deviation occurs [20]. These boundary parameters can be configured differently according to the underlying asset's liquidity depth and historical volatility characteristics, and can be gradually loosened by the community as the asset matures. However, setting a rate cap embeds a mechanism-stability trade-off that cannot be ignored: the core function of the funding rate is to correct the basis between the perpetual futures and the spot price through a price signal, and when the rate hits its cap and is truncated, the price-correction mechanism fails, the perpetual futures may become persistently unanchored from the spot price, and this instead leads to larger basis accumulation and an eventual violent correction. The setting of the rate cap must therefore strike a balance between "preventing extreme rate pressure" and "maintaining the effectiveness of price anchoring." One possible alternative is to dynam-

ically tighten the open-interest cap as the rate approaches its upper bound, relieving basis pressure from the demand side by restricting the size of new positions.

Parameterizing the size of the risk buffer pool bears directly on the protocol's ultimate solvency. The funding ratio of the insurance fund, the share drawn from liquidation penalties, and the critical value at which auto-deleveraging is triggered when the insurance fund is depleted are all key parameters for maintaining the system's credit. Linking these parameters to the protocol's total value locked or total open interest, and allowing them to be dynamically adjusted through governance, ensures that the protocol's capital adequacy remains within a safe range as it scales.

In practice, the parameterized settings of the insurance fund must confront a structural problem. In the liquidation mechanisms of both centralized exchanges and on-chain protocols, when the liquidation engine executes a forced liquidation, if insufficient market liquidity causes the actual closing price to be worse than the position's bankruptcy price, the resulting shortfall loss is borne by the insurance fund. However, when the cumulative rate of shortfall losses exceeds the rate at which the insurance fund is replenished, the protocol faces solvency risk. The parameterized settings of the insurance fund should therefore be conservative and cannot be computed based solely on the historical average loss rate. The calibration of the insurance fund's size should not rely solely on historical backtesting, because the history of the crypto market spans little more than a decade and its structure continues to change rapidly, so future maximum shortfall losses will almost certainly exceed the historical record. A more robust approach is to base the calibration on forward-looking stress testing, simulating extreme but plausible market scenarios (such as a major asset dropping 30% within 1 hour while market liquidity evaporates, a major stablecoin depegging and causing collateral values to plunge, or oracles failing simultaneously and causing liquidation delays) and computing the total shortfall loss under those scenarios. The risk metric should adopt a measure capable of capturing tail risk, such as conditional value at risk (CVaR), rather than relying solely on the historical backtest of extreme quantiles.

30.3.4 Parameterizing market-making obligations

In a decentralized derivatives market, the continuous supply of liquidity is the foundation for guarding against price manipulation and liquidation cascades. Centralized exchanges typically constrain the behavior of designated market makers through bilateral legal contracts, whereas in an on-chain environment market-making obligations must be converted into parameterized rules that a smart contract can verify and enforce.

Traditional securities and derivatives exchanges have mature institutional frameworks for market-making obligations. The Chicago Mercantile Exchange's (CME) market maker program requires a designated market maker to maintain two-sided quotes at least 85% of the time during a trading session, with a maximum bid-ask spread not exceeding a cap specified for the particular contract and a minimum quote size on each side that meets a specified number of contracts. The New York Stock Exchange's (NYSE) designated market maker (DMM) system imposes stricter continuous-quoting obligations on specific stocks while providing an obligation-waiver mechanism during

periods of extreme volatility: when an underlying asset triggers a circuit breaker or the market enters a volatility-interruption state, market-making obligations are automatically suspended. These traditional frameworks use legal contracts and regulatory licenses as enforcement backstops, and the consequences of default include fines, suspension of eligibility, and even revocation of the market-making license. On-chain systems convert these obligations from legal contracts into parameterized rules in a smart contract; while this improves execution certainty, it also faces the engineering challenge of designing equivalent incentive and penalty mechanisms in an environment beyond legal jurisdiction.

Minimum depth and maximum spread constitute the core quantitative metrics of market-making obligations. A protocol sets, through the parameter registry, the amount of liquidity that must be maintained within a specific price range and the maximum spread permitted between bid and ask quotes. These parameters are not fixed but are set in tiers according to the market-capitalization size and trading activity of the underlying asset. For major assets, the parameter requirements are stricter, so as to provide execution quality comparable to that of a centralized exchange; for long-tail assets, the parameters are relatively lenient, so as to attract early liquidity providers. Market makers' rewards are tied directly to their attainment rate on these parameterized metrics, and the smart contract automatically computes and distributes rewards each epoch.

In a parameterized market-making-obligation framework, the parameters for minimum depth and maximum spread should be set in tiers according to the characteristics of the underlying asset. For major assets such as Bitcoin and Ethereum, the parameters can be set very strictly—for example, requiring market makers to maintain a bid-ask spread of no more than 3 to 5 basis points (bps) within a certain range around the midprice. The cost structure of on-chain market making differs markedly from that of a centralized exchange: each quote update requires submitting an on-chain transaction and paying a gas fee, and the information asymmetry caused by block latency further intensifies market makers' adverse-selection costs. These factors jointly constrain the range within which on-chain market-making-obligation parameters can be set, making it difficult to reach the sub-1-basis-point spread levels seen on centralized exchanges. For newly listed long-tail assets, the parameters can be relatively lenient in order to attract early liquidity providers. This tiered parameter setting embodies a core principle of on-chain finance: optimizing every participant's experience to the greatest extent without introducing systemic risk. Moreover, to prevent market makers from placing a burst of orders at the end of an evaluation epoch merely to satisfy the obligation requirement, the measurement of market-making obligations should adopt a time-weighted liquidity-presence rate rather than a simple attainment-rate assessment based on an epoch-end snapshot.

Parameterizing obligation-waiver conditions is the protective mechanism of the on-chain market-making scheme under extreme market conditions. When the market experiences violent turbulence, oracle-quote delays, or underlying network congestion, forcing market makers to maintain tight quotes could subject them to severe adverse-selection losses. The protocol introduces a parameterized trigger based on volatility or price deviation; when the market state exceeds the preset normal range, the smart contract

automatically loosens the maximum-spread constraint or temporarily suspends the assessment of market-making obligations. This parameterized waiver mechanism protects the capital of liquidity providers while also avoiding the liquidity vacuum that would result from a collective withdrawal of quotes by market makers. However, an automated waiver mechanism faces two engineering-level challenges. The first is manipulation-resistant design: if the waiver trigger condition is based on instantaneous volatility, a market maker might strategically manufacture a volatility signal by submitting a large market order during a thin-liquidity period, thereby triggering the waiver to escape its market-making obligation. The waiver trigger should therefore use a time-weighted average volatility combined with cross-validation across multiple oracles. The second is the liveness guarantee: during extreme market conditions, on-chain infrastructure is most prone to congestion, yet this is precisely the moment when market makers most need the protection of a waiver, and the waiver-triggering transaction itself may fail to be confirmed on-chain in time due to gas competition. For this reason, the system should design a time-based automatic waiver as a fallback mechanism: if no quote-update transaction from a market maker is detected across several consecutive blocks, the system automatically treats the waiver state as in effect, without requiring the market maker to proactively submit a waiver request.

30.3.5 The governance mechanism for parameter adjustment

The effectiveness of parameterized compliance ultimately depends on the power structure that modifies the parameters. If parameters can be arbitrarily tampered with by a handful of insiders, all promises of transparency lose their meaning. Parameter adjustment must therefore be placed within a rigorous governance framework that balances immutability against flexible response.

Parameter adjustment in an on-chain protocol is subject to three tiers of governance constraint. The first tier is routine adjustment executed automatically by code—for example, a volatility metric computed from oracle price feeds automatically updating the margin multiplier without human intervention. The second tier is the governance review of structural parameter changes, which requires an on-chain proposal and community vote and is executed only after a timelock delay [16]. The third tier is an emergency multisig committee mechanism for extreme market crises, which permits rapid adjustment of key parameters within a strictly delimited scope [21]. The complete design and engineering implementation of these three tiers of governance constraint are developed systematically in Section 30.6.

30.4 Reporting-based compliance

Within the three-tier compliance spectrum, reporting-based compliance has the shallowest embedding but the broadest coverage. For requirements whose ex ante restriction is technically difficult to define precisely and whose forcible interception could impair the system's normal operation, reporting-based compliance exploits the immutability and global visibility of on-chain data, converting them into data-driven ex post monitoring and automated auditing that satisfy the regulatory demand for transparency without

adding friction to transaction execution (the four core scenarios are discussed in Section 30.4.1).

30.4.1 When reporting-based compliance applies

As set out in the three-tier spectrum of Section 30.1.3, hard-coded compliance guards non-negotiable security baselines and parameterized compliance adjusts operational rules that must vary flexibly with market conditions; but not all regulatory requirements are suited to being implemented through ex ante restriction. In some scenarios, ex ante restriction is not only technically difficult to define precisely but could also seriously and adversely affect the system's normal operation and market efficiency. For such requirements, reporting-based compliance offers a superior engineering solution: converting compliance from a behavioral constraint into ex post monitoring and automated auditing based on on-chain data.

The core logic of reporting-based compliance is not to restrict participants' specific behaviors but to ensure the visibility, traceability, and analyzability of all behaviors. This mechanism is especially suited to four core scenarios. The first is transaction monitoring and anomaly detection. In complex financial markets, malicious behaviors such as wash trading, price manipulation, or insider trading are often difficult to intercept with a simple rule before a transaction occurs. Forcibly imposing ex ante interception is highly prone to false positives, thereby impeding legitimate market-making and arbitrage activity. The second is proof of reserves. Proving that a protocol holds enough assets to cover user liabilities is a system state that requires continuous verification, not a static metric controllable through a single parameter. The third is risk disclosure. Displaying the system's overall risk exposure and an individual account's distance to liquidation in real time to market participants falls within the category of information accessibility; its value lies in reducing information asymmetry rather than directly restricting trading. The fourth is regulatory reporting. Providing regulators with a complete data record of the system's operation is the basis for satisfying compliance-audit requirements.

In all four scenarios, reporting-based compliance fully exploits the underlying properties of blockchain technology. The immutability of on-chain data ensures the authenticity of historical records, while the global visibility of the data breaks down the data silos pervasive in traditional financial systems. By converting these four requirements into structured extraction and automated analysis of on-chain data, reporting-based compliance not only reduces the system's operating friction but also surpasses the traditional external-enforcement model in both the depth and breadth of compliance.

30.4.2 Transaction monitoring and anomaly detection

Transaction monitoring at a traditional centralized exchange depends heavily on internal data systems, and when regulators conduct an audit, they can obtain only the data snapshots or aggregate reports that the exchange selectively provides. This architecture entails systemic information asymmetry: the completeness and authenticity of the data depend entirely on the honesty of the centralized institution. By contrast, on-chain systems have an inherent architectural advantage in transaction monitoring. All order submissions, execution records, fund transfers, and liquidation events are permanently

recorded on the distributed ledger in the form of cryptographic proofs. For a fully on-chain system, this nearly complete data coverage allows the transaction-monitoring system to reconstruct historical market states with high precision (for hybrid and app-chain architectures whose core state runs on the consensus layer or off-chain, the degree of coverage depends on the observability of that state).

On this foundation of high-precision data, the effectiveness of automated anomaly-detection models is significantly enhanced. In an on-chain environment, every transaction is bound to a specific cryptographic address, and the flow path of funds is fully transparent. By applying machine-learning algorithms and graph-network analysis, a monitoring system can automatically identify complex market-manipulation patterns. For example, if the same address or a cluster of related addresses repeatedly trades on both the buy and sell sides within a very short time, the system can quickly flag it as potential wash trading; or if a specific address builds a large leveraged position that sharply departs from its historical behavior just before a major protocol upgrade or governance proposal is announced, this pattern can be captured in real time and flagged as suspected insider trading [22]. This automated detection mechanism does not introduce additional latency into the consensus path of transaction execution but is completed asynchronously by independent monitoring nodes after the data is committed to the chain. However, the precision of on-chain anomaly detection still faces inherent technical bottlenecks. In an order-book environment, the legitimate two-sided quoting activity of a market maker and wash trading are highly similar in their on-chain data form, since both manifest as frequent trading by the same entity on both the buy and sell sides. Moreover, wash traders can evade correlation analysis by creating a large number of mutually unrelated addresses, and this Sybil attack subjects detection methods based on address clustering to a fundamental accuracy constraint. On-chain transparency provides a complete data foundation, but a significant analytical gap remains between the data and a reliable determination of a violation.

More importantly, the openness of on-chain data gives rise to an entirely new compliance paradigm: crowdsourced compliance. Under the traditional framework, transaction monitoring is the exclusive prerogative of regulators or an exchange's internal compliance department. In an on-chain system, anyone—independent researchers, competitors, data-analytics firms, and even ordinary users—can obtain the complete transaction dataset and run their own monitoring algorithms. This distributed oversight significantly raises the probability that malicious behavior will be discovered. When compliance is no longer the exclusive responsibility of a single centralized entity but evolves into a public activity in which the entire community participates, the system's overall resistance to malfeasance is markedly strengthened. However, the effectiveness of crowdsourced monitoring is constrained by the free-rider problem: when everyone can monitor, each individual lacks the economic incentive to bear the cost of monitoring, and the result may be that no one performs systematic in-depth analysis. To turn crowdsourced compliance from a theoretical advantage into a practical effect, complementary economic-incentive mechanisms must be designed, such as a bounty pool for effective reports and a requirement that reporters stake a bond to deter malicious false reports.

30.4.3 Automating proof of reserves

In the development history of centralized cryptocurrency exchanges, the evolution of proof-of-reserves mechanisms deeply reflects the transformation of trust architecture. Early centralized exchanges relied entirely on reputational endorsement, until a series of severe incidents of customer-asset misappropriation led the industry to widely adopt Merkle-tree-based proof of reserves. However, this proof of reserves, voluntarily initiated by a centralized institution, has significant limitations. It can typically prove only that, at a specific time snapshot, the known on-chain addresses controlled by the exchange contained a specific amount of assets; it cannot prove that these assets were not misappropriated at points outside the snapshot, still less fully rule out the possibility that the exchange has hidden liabilities [23]. In essence, a centralized exchange's proof of reserves is a discrete audit activity rather than an intrinsic property of the system architecture.

On-chain financial systems restructure the implementation logic of proof of reserves through the self-custody property of smart contracts. In a fully on-chain architecture, users' asset balances and the protocol's total reserves are all publicly queryable global state variables. Anyone can independently compute the system's total liabilities and total assets at any moment and verify whether the latter is greater than or equal to the former—without relying on a report from a third-party auditor and without waiting for the exchange to release a quarterly snapshot. The marginal cost of this verification process approaches zero and is not controlled by any centralized entity.

Centralized exchanges are themselves drawing on on-chain technology to improve their proof-of-reserves mechanisms. In November 2022, Binance launched a Merkle-tree-based proof of reserves, and in February 2023 it introduced zk-SNARKs to strengthen it, forming a hybrid scheme combining Merkle trees and zk-SNARKs. Users can verify the path of their own leaf node within the Merkle tree to confirm that their assets are included in the total-reserve calculation, while the zero-knowledge proof is used to prove that all user balances are non-negative and that their sum matches the on-chain address balances. OKX (using zk-STARKs) has also launched a zero-knowledge-proof-based scheme, while Bybit and others provide verifiable-reserve schemes based on Merkle trees and third-party audits. This trend indicates that the technical standards for reserve transparency established by on-chain systems are being fed back to centralized institutions, pushing up the compliance baseline of the entire industry. However, CEXs' zk-PoR schemes still rely on the exchange to proactively initiate the proof and cannot cover hidden liabilities; their architectural-level limitations differ fundamentally from the continuous guarantee of an on-chain system.

Thus, an on-chain system's proof of reserves achieves a leap from a one-time audit to a continuous guarantee. The data itself is a continuous, immutable record, and the proof process is embedded automatically into the logic of reading the data. As long as the consensus mechanism of the underlying blockchain is not compromised, the reserve data recorded in the smart contract possesses a high degree of certainty. This architectural-level guarantee substantially compresses the room a centralized institution has to falsify balance-sheet data, making reserve transparency an inherent, default property of the system.

The description of proof of reserves above applies mainly to spot-custody protocols, whose liability calculation is relatively simple—that is, total assets need only be no less than total deposits. For derivatives protocols such as perpetual futures, however, the engineering implementation of proof of reserves is far more complex. The "total liabilities" of a derivatives protocol include not only users' margin balances but also the unrealized profit and loss (P&L) of all active positions, the funding rate pending settlement, and the shortfall losses already borne by the insurance fund. Because unrealized P&L changes in real time with the mark price, precise liability calculation requires traversing all active positions at each point in time and revaluing them at the current mark price. This means that a derivatives protocol's proof of reserves is not a simple state query but a computation-intensive process involving global state traversal and real-time revaluation, whose engineering complexity and computational overhead should not be underestimated.

30.4.4 Real-time risk disclosure

Risk disclosure is one of the core requirements of consumer protection. On traditional derivatives-trading platforms, risk disclosure often takes the form of an after-the-fact static report or a disclaimer at account opening. A trader holding complex positions finds it difficult to intuitively assess their real-time risk exposure, especially when violent market swings cause margin requirements to change dynamically. The reporting-based compliance mechanism of an on-chain system builds a dynamic risk-reporting system by converting underlying state data into a real-time, trader-facing view.

This real-time risk disclosure spans several key dimensions. The system can precisely compute and display a trader's current margin ratio and the specific price distance to the trigger of a forced liquidation. If the underlying asset's price moves adversely by a specific magnitude, the trader can precisely compute the percentage distance from their current position to the liquidation-trigger price. In addition, the system can display in real time the current balance of the insurance fund and its ratio to total open interest, allowing traders to assess the system's capacity to absorb shortfall losses under extreme market conditions [19]. Under an auto-deleveraging mechanism, a trader can even view their position in the deleveraging queue in real time, forming a prior judgment about the potential risk of passive closure.

Providing this high-granularity real-time data essentially surpasses the passive compliance traditionally undertaken merely to satisfy regulatory inspection. It constitutes a mechanism of responsible information provision to traders and directly improves the system's transparency and user experience. By converting the complex state hidden deep within the matching engine and liquidation logic into intuitive risk metrics, on-chain systems achieve a structural improvement over centralized exchanges on the key dimension of information accessibility, enabling traders to manage risk on a fully informed basis.

30.4.5 The regulatory interface

In the traditional financial compliance model, persistent gaming and friction exist between regulators and regulated entities. Regulators must periodically dispatch auditors

into an institution or require the institution to submit cumbersome compliance reports in a specific format. This model not only incurs substantial compliance costs but also suffers from severe time lag, so that regulators often discover the first signs of a problem only through ex post review after a systemic risk has already erupted.

On-chain systems offer an entirely new technical path to resolve this predicament: designing a standardized regulatory interface for regulators. Its core idea is that, rather than having regulators laboriously inspect the system, the system should proactively and in real time report its operating status to regulators. By exposing a specific application programming interface (API), regulators can at any time pull key metrics such as the system's total open interest, the distribution of the liquidation queue, the adequacy of the insurance fund, and the margin coverage ratio for specific assets. This mechanism eliminates information asymmetry, because the data regulators read through the interface is fully homologous with, and synchronized to, the system's actual operating state.

Figure 30-4 presents the end-to-end technical pipeline of reporting-based compliance, from data collection and automated auditing through to outputs for three audiences—traders, the public, and regulators—driven entirely by code, eliminating the reliance on manual data organization and report preparation found in traditional compliance workflows.

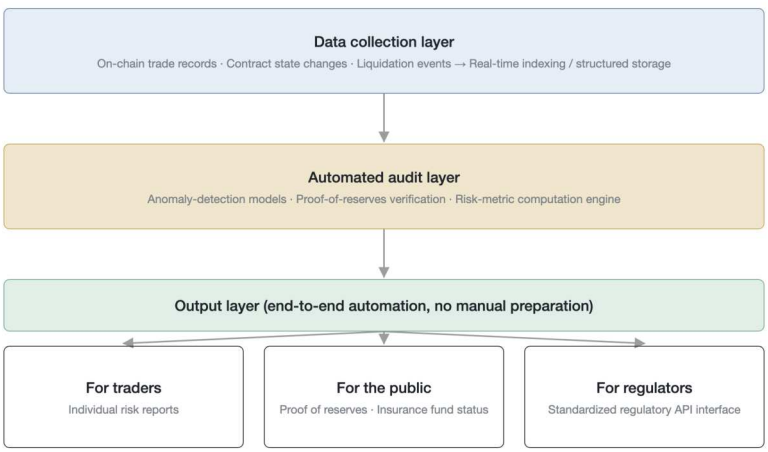


Figure 30-4. The three-stage pipeline of reporting-based compliance, from data collection and automated auditing to outputs for traders, the public, and regulators [3][22]

This interface-based automated reporting mechanism brings multiple significant benefits. It significantly reduces a protocol's operational compliance costs, because report generation is driven entirely by code and requires no human intervention. At the same time, regulators gain real-time insight synchronized with the system's state, enabling them to issue early warnings during the risk-accumulation stage rather than merely assigning blame after the fact. To be sure, the broad application of this mechanism still faces engineering challenges, particularly because regulators in different jurisdictions

require different data formats and granularity, which calls for the industry to jointly promote the standardization of regulatory data interfaces. A further key direction for reporting-based compliance is to protect ordinary traders' privacy through cryptographic techniques such as zero-knowledge proofs while still giving regulators adequate transparency.

The effectiveness of reporting-based compliance depends on the global visibility of on-chain data, but this visibility conflicts structurally with user privacy: the contradiction between regulators' demand for penetrating visibility into traders' identities and the sources of their funds, and traders' basic need for financial privacy, cannot be resolved through a simple trade-off but requires a cryptographic-level technical reconciliation, which the next section addresses specifically.

30.5 The technical reconciliation of privacy and compliance

Embedding compliance rules into an on-chain system confronts a core tension: regulators demand full penetrating visibility into participants' identities and fund flows, whereas the global visibility of a public blockchain means that any identity linkage would expose a user's entire transaction history. Under the traditional framework, this contradiction between privacy protection and compliance requirements can be handled only by choosing one over the other. The maturation of zero-knowledge-proof technology offers an engineering path to reconcile this contradiction, enabling a trader to prove to the system that they satisfy specific compliance conditions without exposing the underlying identity data.

30.5.1 The structural conflict between compliance and privacy

The cornerstone of the traditional financial-regulatory system is complete knowledge of participants' identities, whereas the native design of crypto-asset markets is built on decentralization and pseudonymity. This architectural difference produces a structural conflict between know-your-customer (KYC) requirements and traders' demand for privacy protection. The logical starting point for regulators is information transparency—collecting, storing, and monitoring traders' personal identity information and fund flows in order to guard against illicit activities such as money laundering, terrorist financing, and sanctions evasion. Within this framework, any technology or mechanism that impedes information penetration is regarded as a potential compliance risk, and anonymous transactions therefore become the primary target of regulatory intervention.

However, directly implementing the traditional information-collection model on the transparent ledger of a public blockchain would bring severe privacy-leakage risk. The global visibility of a public blockchain means that once a user's real identity is bound to an on-chain address, all of their historical transactions, asset balances, interacted protocols, and even fund-flow patterns become public to everyone. This mandatory transparency not only exposes a trader's commercial secrets and trading strategies but also subjects them to targeted cyberattacks, phishing scams, and even threats to real-world personal safety. When traders use tools such as mixers, privacy coins, or multiple addresses to sever the link between their identity and their on-chain activity, this is

not necessarily for illicit purposes but rather to rebuild basic financial privacy on a transparent ledger.

This opposition between compliance and privacy is especially evident in current regulatory policy. The EU's Markets in Crypto-Assets (MiCA) regulation requires crypto-asset service providers to be able to identify the holders and transaction histories of assets with built-in anonymization features; otherwise, the platform must block trading in such assets [24]. Similarly, the sanctions imposed on Tornado Cash by the U.S. Treasury's Office of Foreign Assets Control (OFAC) reflect regulators' tendency to equate anonymity directly with an illicit financial instrument [25]. However, the legal basis for this position has been challenged in court. In November 2024, the U.S. Court of Appeals for the Fifth Circuit ruled in Van Loon v. Department of the Treasury that OFAC's sanctions on Tornado Cash's immutable smart contracts exceeded its statutory authority, on the grounds that immutable smart contracts do not constitute "property" under the International Emergency Economic Powers Act (IEEPA). This ruling provides an important precedent for the legal positioning of embedded compliance—namely, that immutable on-chain compliance logic may not be subject to the same sanctions-review framework as mutable systems—but it also raises a new question of how "immutability" should be defined under an upgradeable proxy-contract architecture. This either-or dichotomy confronts on-chain finance with a dilemma: either abandon privacy entirely to satisfy compliance requirements, reducing the public blockchain to a high-risk transparent database; or refuse compliance to protect privacy, thereby marginalizing itself from the mainstream financial system.

30.5.2 Compliance applications of zero-knowledge proofs

The evolution of cryptographic technology, and in particular the maturation of the zero-knowledge proof, provides an engineering solution to reconcile the contradiction between compliance and privacy. A zero-knowledge proof allows a prover to convince a verifier of the truth of a statement without revealing any additional information beyond the fact that the statement is true. In the compliance scenarios of on-chain finance, this property achieves a paradigm shift from "information disclosure" to "fact verification": a trader can prove to a smart contract or a regulatory node that they satisfy specific compliance conditions without submitting any underlying identity data or transaction details.

This "prove-without-revealing" mechanism significantly changes the logic of processing compliance information. Take anti-money-laundering review: the traditional model requires a user to submit a passport, proof of address, and records of the source of funds, which the platform manually reviews and stores as sensitive data. Under a zero-knowledge-proof-based framework, the user need only generate a cryptographic proof locally attesting that "I am not a citizen of a sanctioned country" or "the source of my funds does not include any flagged high-risk address." Acting as the verifier, the smart contract need only confirm the validity of the proof through mathematical computation to admit the transaction, and the entire process involves no on-chain transfer or storage of any specific nationality or historical transaction record.

A study based on Monte Carlo simulation (Decker, 2025, an SSRN preprint that has

not been peer-reviewed) estimates that a zero-knowledge-proof-based KYC verification framework can reduce the volume of exposed user data by approximately 97% and, when combined with an artificial intelligence (AI) model, can achieve approximately 96.7% anomaly-detection accuracy within its stipulated simulation scenario [26]. These figures, however, come from a single simulation study that has not been independently replicated; moreover, zero-knowledge proofs themselves constrain the scope of data exposure rather than directly improving detection accuracy (the latter depends on the AI model with which they are combined). This technology offers not only significant advantages in privacy protection but also eliminates, at the system-security level, the single-point-of-failure risk of a centralized institution acting as a data honeypot. Because compliance verification is converted into pure mathematical computation, an on-chain financial system can build compliance requirements into its architecture without bearing the legal and security responsibility of safeguarding vast amounts of sensitive personal information. This is a core dimension on which on-chain finance surpasses the traditional financial system in the design of compliance mechanisms.

The applicable boundary of zero-knowledge-proof compliance solutions must be honestly delimited. At the current stage, the compliance obligations that zero-knowledge proofs can effectively satisfy are concentrated mainly in the onboarding stage, including sanctions-list screening, accredited-investor verification, and basic anti-money-laundering risk rating. However, the full set of anti-money-laundering and counter-terrorist-financing (CTF) compliance obligations extends far beyond onboarding review. The Financial Action Task Force's (FATF) "Travel Rule" requires that the identity information of both the sender and the receiver be transmitted in virtual-asset transfers, which conflicts structurally with the core logic of zero-knowledge proofs, "proof without exposure." The obligation of continuous transaction monitoring requires the compliance-obligated party to be able to access underlying transaction data in order to identify suspicious patterns, and the obligation to proactively file suspicious transaction reports further requires disclosing the detailed information of specific transactions to regulators. Under the current technical framework, these obligations cannot yet be satisfied by a pure zero-knowledge-proof solution. Acknowledging this boundary does not weaken the value of zero-knowledge proofs but delineates a realistic scope for their engineering application, avoiding the erosion of regulatory trust that overpromising would cause.

Figure 30-5 presents the complete workflow of this "prove-without-revealing" mechanism in an on-chain system: off-chain, a trusted KYC provider verifies identity and issues a cryptographic credential containing only specific attribute claims; the user then generates a zero-knowledge proof locally on that basis; and the on-chain smart contract admits the transaction once it confirms the proof's validity through mathematical computation alone—the entire process involving no on-chain transmission or storage of any identity information.

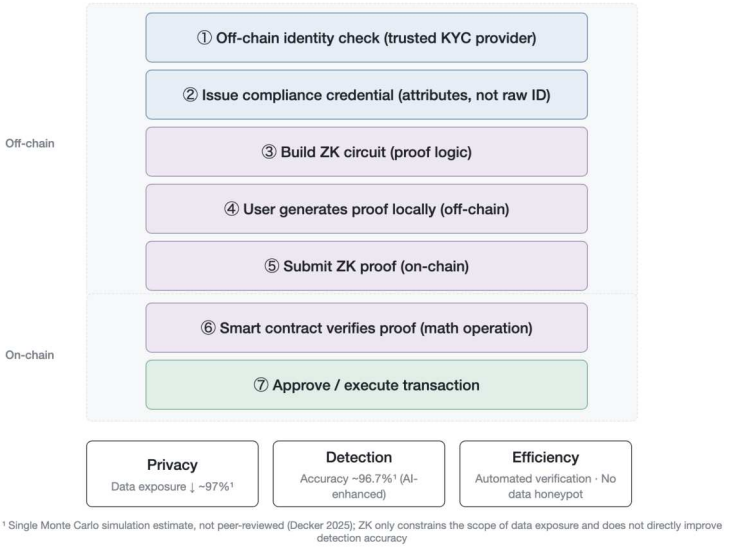


Figure 30-5. The workflow of zero-knowledge-proof-based compliance credentials [26] (the 97% and 96.7% figures shown are single-simulation estimates from Decker 2025 and are not peer-reviewed)

30.5.3 Designing compliance credentials

To turn the theoretical advantages of zero-knowledge proofs into an executable engineering scheme, a system must introduce a modular "compliance credential" design. This design follows the principle of minimal necessary information, deconstructing complex identity attributes into independent, separately verifiable claims and thereby avoiding the "all-or-nothing" data-bundled submission of traditional identity verification. The lifecycle of a compliance credential spans three key stages—issuance, verification, and revocation—which together constitute a closed-loop on-chain identity-management system.

In the issuance stage, a trusted off-chain entity (such as a licensed KYC provider or a government agency) is responsible for verifying the user's real identity. Once verification is complete, this entity does not put the user's identity information on-chain but instead issues a cryptographic credential containing specific attribute claims and digitally signs it with its private key. These credentials are typically stored in the user's local wallet, under the user's full control. For example, a credential might contain only the claim that "this user has passed a level-one anti-money-laundering review" or "this user is an accredited investor," without any redundant information such as name or ID number.

In the verification stage, when a user attempts to interact with a compliance-constrained smart contract, the contract states a specific attribute requirement. The user uses the locally stored credential to generate a zero-knowledge proof attesting that they hold a valid credential, issued by a trusted issuer, that satisfies the contract's requirement. The smart contract verifies the mathematical correctness of the proof and the validity of the issuer's signature, and if both pass, it automatically authorizes the transaction to execute. The entire verification process is completed on-chain, achieving a decentralized

and automated compliance check.

The revocation mechanism is an indispensable security baseline in compliance-credential design. When a user's compliance status changes (such as being placed on a sanctions list or the credential expiring), the issuer must be able to invalidate the already-issued credential. This is typically implemented by maintaining an on-chain revocation registry or accumulator. When generating a zero-knowledge proof, the user must prove not only that the credential is valid but also, simultaneously, that it is not included in the latest revocation list. This design ensures the capacity to update compliance status dynamically while maintaining the decentralized nature of the verification process.

30.5.4 Engineering challenges and progress

Although zero-knowledge proofs perfectly reconcile the contradiction between privacy and compliance in theory, their practical engineering deployment still faces significant technical and economic challenges. The foremost bottleneck lies in computational overhead and proof-generation time. Generating a zero-knowledge proof is a highly computation-intensive process, and for a circuit containing complex compliance logic, generating a proof on a mobile device or an ordinary personal computer may take seconds or even minutes. This latency is unacceptable in trading scenarios that demand high-frequency, low-latency execution. In addition, although verifying a proof is relatively lightweight, executing the verification logic on a mainnet such as Ethereum still consumes considerable computational resources, thereby pushing up users' transaction costs.

The complexity of circuit design constitutes another engineering obstacle. Converting complex laws, regulations, and compliance requirements into mathematical constraints in a zero-knowledge-proof circuit requires spanning the two distinct professional domains of law and cryptography. Any logical flaw in the circuit could allow the compliance mechanism to be bypassed, and such a bypass is extremely difficult to discover after the fact because of the zero-knowledge property. The development, auditing, and formal verification of compliance circuits are therefore costly, which limits their rapid iteration and broad deployment.

Despite the many challenges, industry exploration in privacy-preserving compliance has made substantial progress. Polygon ID, the first identity solution to allow users to generate zero-knowledge proofs from off-chain credentials, has built a complete verifiable-credential ecosystem that enables decentralized applications to implement access control without collecting user data [27]. Meanwhile, fully decentralized zkKYC solutions launched by projects such as zkMe further minimize the trust assumptions of identity verification, ensuring the censorship-resistance and privacy of the credential-issuance process through a decentralized oracle network and multi-party computation [28]. These engineering practices indicate that zero-knowledge-proof-based embedded compliance is moving from the proof-of-concept stage toward large-scale application, laying the technical foundation for building the next generation of on-chain financial infrastructure that both satisfies regulatory requirements and defends user privacy.

30.6 Engineering the governance layer

When compliance rules are embedded into smart contracts in a hard-coded or parameterized form, the system gains a sense of immutable security but simultaneously loses the flexibility to respond to unknown risks and adapt to regulatory change. If the design stage fails to provide a secure upgrade channel for these rules, the system will lose its usability in a rapidly evolving market environment because of excessive rigidity. This predicament raises the core question of this section: how can a governance mechanism be designed within an architecture that pursues trustlessness so as to guarantee both immutability and secure upgradeability?

30.6.1 Immutability versus the need to upgrade

The core value proposition of a smart contract is immutability—once code is deployed, its execution logic is free from interference by any centralized entity. In on-chain financial systems, this property is the cornerstone of building user trust, because it eliminates the risk that an operator could misappropriate assets or unilaterally modify liquidation rules. However, a fully immutable system has proven fragile in engineering practice. The complexity of financial markets, the potential code vulnerabilities of smart contracts, and continuously evolving regulatory requirements jointly constitute a rigid demand for system upgradeability [29].

This contradiction is especially pronounced in the process of embedding compliance rules. Although hard-coded compliance (such as the absolute leverage cap and asset-segregation logic) provides the highest level of certainty, once a logical flaw is discovered or an extreme market condition arises, these protective measures may, absent an upgrade mechanism, instead become the trigger for the system's collapse. For example, when the operating environment of the underlying blockchain changes, or when regulators adjust the compliance standards for specific assets, a fully immutable contract would force the entire system to undergo a destructive hard fork or asset migration [29].

The engineering design of the governance layer is therefore essentially a precise calibration of the distribution of power. Granting an administrator key unrestricted upgrade authority means reverting to the trust model of a traditional centralized exchange; completely abandoning the capacity to upgrade places the system in an unrepairable and dangerous position. In the blueprint of a trustless architecture, the governance layer (Layer 5) must serve as the connective layer between static code and dynamic reality. This requires converting governance itself into a verifiable on-chain process, so that every rule modification or parameter adjustment is strictly constrained by cryptographic and game-theoretic mechanisms.

30.6.2 The dual architecture of timelock and emergency powers

To balance immutability against flexible response, on-chain governance has, in engineering practice, formed a dual architecture: a routine governance path controlled by a timelock, and an emergency-powers path held by a multisignature (multisig) committee. This design acknowledges the fundamental difference in time sensitivity between rou-

tine operational upgrades and extreme-risk response, and configures different security safeguards for each [30].

Figure 30-6 contrasts the operating logic of the dual governance architecture: the routine path is executed after a community vote and a timelock delay, providing a window for public review and allowing dissenting users to withdraw their funds and exit; the emergency path allows a multisig committee to bypass the timelock and immediately execute defensive operations, while triggering mandatory ex post public disclosure. The figure systematically compares the trigger conditions, authority boundaries, and accountability mechanisms of the two paths.

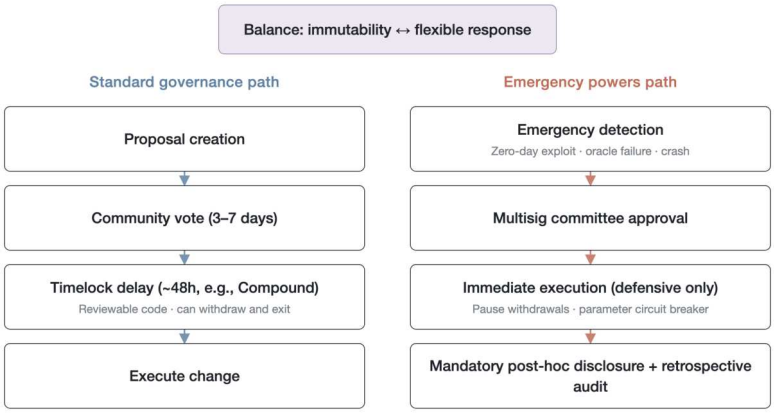


Figure 30-6. The dual governance architecture of timelock and emergency powers [30][31]

The routine governance path applies to parameter adjustments, new-feature deployments, and non-urgent code upgrades. On this path, the timelock mechanism plays the core defensive role. After a governance proposal passes, its execution code does not take effect immediately but is placed within a mandatory waiting period (typically around 48 hours—Compound, for instance, uses 48 hours, and some protocols use longer). This delay serves a dual purpose: on the one hand, it provides the community a window to review the code about to be executed, ensuring that the proposal's content is consistent with the actual execution logic; on the other hand, if some users disagree with the change, the timelock gives them ample time to withdraw their funds and exit the system [31]. During the timelock period, if malicious code or a serious vulnerability is discovered in the proposal, the community can cancel the execution through a preset veto mechanism.

The emergency-powers path is dedicated to scenarios requiring immediate intervention, such as a zero-day-vulnerability attack, a serious oracle failure, or an extreme market crash. In such cases, a 48-hour delay could allow the system's funds to be completely drained. The architecture therefore typically sets up an emergency committee controlled by a multisig, which is granted the power, under specific conditions, to bypass the timelock and immediately execute defensive operations such as pausing the system or triggering a parameter circuit breaker [32].

However, the existence of emergency powers inevitably introduces centralization risk. To limit the abuse of this power, the engineering design must impose strict constraints. The scope of the emergency committee's authority is typically hard-coded and limited to defensive operations (such as being able only to pause withdrawals, not to transfer funds). More importantly, any exercise of emergency power must trigger an ex post public-disclosure and retrospective-audit procedure. If the community determines in an ex post review that the emergency operation was improper, it can replace committee members or restore the system state through the routine governance path. This mechanism of "ex ante authorization, ex post accountability" constitutes the key to safeguarding the trust baseline within the dual architecture.

In a multi-protocol composable on-chain financial ecosystem, the emergency-pause mechanism also faces a risk dimension that transcends the boundary of a single protocol: cross-protocol spillover effects. When one protocol pauses its liquidation function, other protocols that depend on its price signal may experience an oracle-data interruption; when one protocol pauses withdrawals, other protocols that use its assets as collateral may trigger a cascade of liquidations because the collateral is frozen. To minimize this spillover, the engineering design of the emergency pause should adopt a tiered strategy, imposing differentiated restrictions by functional module rather than a global shutdown. For example, when facing an attack, one can pause only new position openings while retaining closing and withdrawal functions; when an oracle fails, one can freeze only the specific trading pairs that depend on that oracle rather than the entire system. This granular control of tiered pausing is a key design element for the emergency-powers architecture to remain effective in a composable ecosystem.

30.6.3 The proxy contract pattern and governance constraints

"Hard-coding" requires that rules cannot be violated, while "upgradeability" requires that rules can adapt to change. These two seemingly contradictory demands are reconciled in modern smart-contract engineering through the proxy contract pattern. The proxy pattern separates the system's state (such as a user's margin balance) and its business logic (such as the specific compliance rules) into two independent contracts [33].

Users always interact with the proxy contract, which forwards all function calls to the implementation contract through the underlying delegatecall instruction. The implementation contract contains all the hard-coded rules and business logic. In day-to-day operation, the address of the implementation contract is fixed in a storage slot of the proxy contract, and at this point all rules exhibit complete rigidity, with no one able to circumvent or violate these hard-coded constraints.

When a hard-coded rule genuinely needs to be updated, the system can rewrite the implementation-contract address pointer stored in the proxy contract, redirecting it to a brand-new implementation contract that contains the new rules (Figure 30-7). All user interactions always point to the proxy contract, whose address never changes, and the proxy forwards the calls to the current implementation contract through delegatecall; an upgrade requires only modifying this implementation-address pointer to replace the entire business logic without migrating user assets and state data.

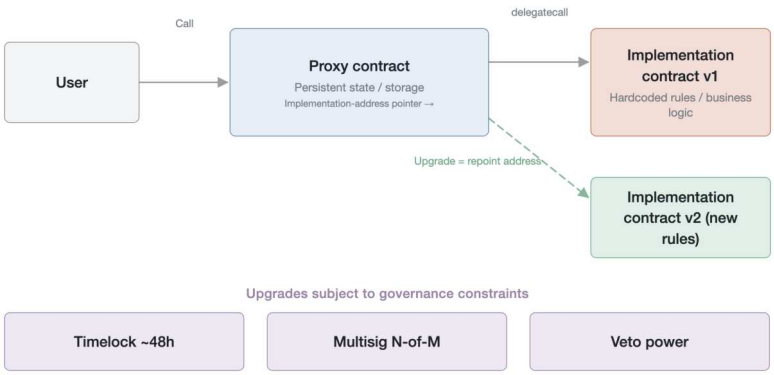


Figure 30-7. The upgradeable hard-coded architecture of the proxy contract pattern [33]

To prevent the abuse of upgrade authority from undermining the trust foundation of hard-coding, the upgrade process must be strictly controlled by the governance layer—which is precisely the core application scenario of the timelock and emergency-powers dual architecture discussed in the previous section: after an upgrade proposal is approved, it can take effect only after the mandatory timelock delay, multisig authorization, and veto procedures described in that section, whose mechanism details are not repeated here. Through this combination of the proxy pattern and governance constraints, an on-chain system achieves a balance between the rigidity of day-to-day operations and the flexibility of long-term evolution. Rules are hard-coded and inviolable at every specific moment, but over an extended time dimension the system possesses the ability to iterate and evolve through a rigorous consensus procedure. The proxy contract pattern gives the term "hard-coding" two levels of meaning in this chapter. The first is "absolute hard-coding": constraints written into the proxy contract's own logic that cannot be modified through any upgrade mechanism, such as the fundamental permission structure for moving assets. The second is "version-level hard-coding": constraints that cannot be violated within the current implementation-contract version but can be modified by replacing the implementation contract with a new one through a governance upgrade. The "hard-coded compliance" discussed in the three-tier compliance spectrum in fact covers both levels, and when assessing the strength of a specific protocol's compliance guarantee, one must distinguish which level its key rules belong to.

30.6.4 Governance attacks and defenses

As governance tokens have been widely adopted in DeFi, the governance layer itself has become a new vector for attackers to gain control of a system and plunder funds. Unlike a technical attack that exploits a code vulnerability, a governance attack fully follows the rules the system sets, passing a malicious proposal through legitimately acquired voting power. Understanding these distinctive threat models is a prerequisite

for designing effective defense mechanisms.

Figure 30-8 organizes the main governance attack vectors (flash-loan voting, whale manipulation, and collusive vote-buying) and their corresponding three-tier escalating security barriers: architectural defense (snapshot voting and timelock), mechanism defense (quadratic voting), and economic defense (governance insurance).

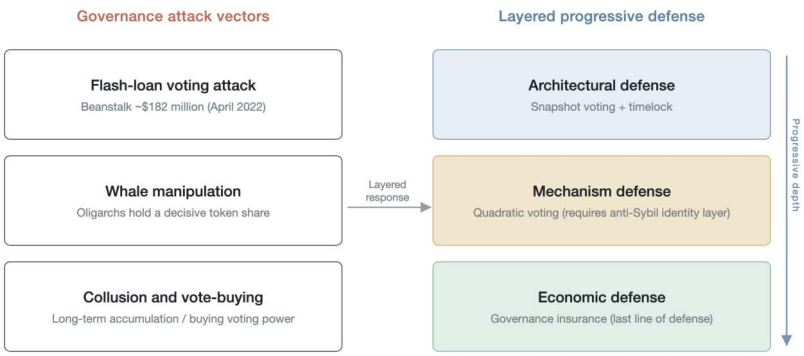


Figure 30-8. The threat model of governance attacks and the multi-layer defense system [34][35]

The flash-loan voting attack is a high-severity threat currently facing the governance layer. An attacker uses an uncollateralized flash loan provided by a lending protocol to borrow a huge quantity of governance tokens within a single block, instantly gaining overwhelming voting power, passing and executing a malicious proposal (such as transferring treasury funds to the attacker's address), and then repaying the loan within the same block (the April 2022 Beanstalk incident used exactly this technique to drain approximately $182 million) [34]. This attack exploits the blind spot of traditional voting mechanisms along the time dimension.

Against flash-loan attacks, the first-tier architectural defense typically adopts a combination of snapshot voting and a timelock. The snapshot mechanism requires the system to record the token balances of all addresses in the block immediately preceding the proposal's creation, and to use these as the basis for voting weight. Because a flash loan must be borrowed and repaid within the same block, the snapshot mechanism prevents an attacker from using flash-loan funds to influence the voting weight of a historical block. A snapshot neutralizes only the single-block flash-loan variant; attacks that borrow across multiple blocks, or that purchase tokens on the spot market before voting, remain feasible—which is precisely why the timelock and veto power must serve as the actual backstop. In addition, the timelock mechanism breaks the synchrony between voting and execution, so that even if an attacker manages to pass a malicious proposal, they must wait until the delay period ends before it can be executed; this not only increases the attack's capital cost but also provides defenders a window to intervene [35].

Whale manipulation and collusive vote-buying constitute another kind of long-term governance threat. When a few entities hold a token share sufficient to determine a pro-

posal's outcome, the system's decision process degrades into oligarchic decision-making. To mitigate this problem, quadratic voting has been introduced as a mechanism defense. Quadratic voting was first proposed by Lalley and Weyl [36]: the square of the number of votes a participant allocates to an option equals the voting credits or token cost they expend [37][38]. This nonlinear cost function significantly increases the marginal cost of accumulating votes, making it difficult for a few whales to easily overwhelm the combined will of numerous retail participants, thereby restoring a degree of balance to governance. However, the theoretical effect of quadratic voting is severely constrained in practice by the voter-participation rate. Voter turnout in mainstream DAO governance is typically below 10%, and in a low-participation environment a whale can still dominate the voting outcome with a relatively small outlay even in the face of nonlinear voting costs. In addition, the on-chain effectiveness of quadratic voting also depends on a Sybil-resistant unique-identity layer: absent identity verification, a whale can split its holdings across a large number of addresses to circumvent the quadratic cost, which is also why purely token-weighted quadratic voting is rarely used for high-value DeFi treasury decisions. Quadratic voting therefore needs to be deployed together with a Sybil-resistant identity layer and complementary mechanisms that raise participation (such as delegated voting and voting incentives) in order to truly fulfill its design intent.

Beyond architectural defense and mechanism defense, economic defense (such as governance insurance) constitutes the last line of defense. When an attacker successfully bypasses the first two layers of defense and causes losses, a preset insurance fund can be used to compensate the affected users, preventing a total collapse of trust in the system. These three layers of defense together constitute a multi-layer defense system that significantly raises the difficulty and cost of launching a governance attack.

30.6.5 The governance roadmap of progressive decentralization

Over the lifecycle of an on-chain financial system, the governance architecture is not fixed. A significant tension exists between the system's fragility in its early days after launch and the trustlessness it pursues in the long run. If control is handed over to the community prematurely—before the code has been sufficiently battle-tested in live trading and before the economic model has reached equilibrium—the system will face high security risk and coordination failure. To address this path problem, engineering practice has developed the concept of "progressive decentralization" [39]. Synthesizing Walden's idea with industry engineering practice, this book summarizes the handover of governance power as a structured, multi-stage transformation process (the specific three-stage division of "team control—multisig committee—full DAO" and the timelock mechanism below are this book's engineering synthesis, not the framework of Walden's original text).

Figure 30-9 presents the three-stage structure of the progressive handover of power—the team-control period, the multisig-committee period, and the full-community-governance period—in order of increasing decentralization, which the following text develops stage by stage in terms of its authority boundaries and transition conditions.



Figure 30-9. The three-stage roadmap of progressive decentralization [39][40] (conceptual illustration; stage durations have no standard)

The first stage is the team-control period, typically covering the initial phase after the system's launch (the specific duration depends on the project, commonly ranging from a few months to about a year, with no uniform standard). In this stage, the core development team retains unilateral control over smart-contract upgrades and key parameter adjustments. This highly concentrated power structure is intended to maximize the system's iteration speed and its ability to respond to sudden events. When the system faces an unknown vulnerability or an economic attack, the team can immediately deploy a patch without going through a lengthy community discussion and voting process. The core goal of this stage is to validate the stability of the system architecture and product-market fit, rather than to pursue formal decentralization. However, to provide users with a basic trust guarantee, the team typically publicly commits to the boundaries of its power and periodically releases security-audit reports.

As the system scales and its operating logic stabilizes, the governance architecture enters the second stage, the multisig-committee period. In this stage, the team begins to proactively reduce its own unilateral control. The administrator keys are transferred to a multisig committee jointly composed of core team members, early investors, well-known security experts, and community representatives. Any operation involving a system upgrade or parameter modification must obtain the signature authorization of a majority of committee members and is mandatorily executed only after a timelock delay [40]. This mechanism not only eliminates the risk of a single point of failure but also provides a time window for community oversight. At the same time, the community begins to be granted the right to propose and vote on non-critical parameters, as a rehearsal for fully taking over governance in the future.

When the system reaches high maturity, its token distribution is sufficiently dispersed, and its core functions have solidified, the governance architecture finally transitions to the third stage, the full-community-governance period. In this end state, the privileges of the multisig committee are completely revoked, and all upgrade authority and parameter-adjustment rights are transferred to a DAO controlled by token holders. Any modification of the smart contracts must strictly follow the standardized process of on-chain proposal, voting, and timelock execution. At this point, the core team steps back to become one of many protocol contributors, no longer enjoying special privileges. This progressive handover of power ensures that the system remains structurally sound throughout its evolution from a "trusted operator" to a "trustless protocol."

However, the theoretical elegance of progressive decentralization faces a serious credible-commitment problem in practice. It is easy for a core team to promise, in the early days after launch, that it will relinquish control in the future; but when decentralization means actually giving up the economic privileges of adjusting fees, managing the treasury, and steering the protocol's direction, the team faces a strong time-inconsistency incentive to extend the first or second stage indefinitely. In reality, a large number of blockchain projects have never completed the handover of power under the banner of "progressive decentralization," and the progressive-decentralization roadmap degenerates into a rhetorical tool for perpetuating centralized control. To resolve this commitment problem, the engineering design can introduce irreversible decentralization milestones—for example, presetting in the smart contract a logic that automatically reduces authority based on a time or system-scale threshold, so that once the system's total value locked exceeds a specific threshold or its operating time exceeds a predetermined term, the administrator authority is automatically downgraded, and this downgrade logic itself cannot be overridden by the upgrade authority.

30.6.6 A governance audit of Hyperliquid

Before assessing the governance state of an on-chain system, it is worth contrasting it with the governance structure of a centralized exchange. The governance power of a traditional CEX is concentrated in the corporate board and the internal compliance committee; parameter adjustments (such as changes to margin ratios and listing decisions) are executed unilaterally by management, and users have neither voting rights nor prior knowledge. The external constraints on a CEX come mainly from the conditions of its regulatory license and legal accountability, not from architectural-level checks on power. By contrast, on-chain governance grants stakeholders direct participation through token voting, the timelock mechanism ensures the transparency and predictability of changes, and the multisig structure eliminates single-point decision risk. This governance-paradigm difference from "institutional self-discipline" to "architectural constraint" constitutes the basic frame of reference for assessing the governance maturity of an on-chain system such as Hyperliquid.

Applying the governance theoretical framework above to Hyperliquid, a specific on-chain derivatives-trading infrastructure, allows a clear assessment of the gap between its current governance state and the ideal decentralization goal. As a system built on the customized app-chain HyperEVM, Hyperliquid's governance design involves not only parameter adjustments at the smart-contract level but also, more deeply, the validator mechanism at the consensus layer.

At the architectural level, Hyperliquid's governance core relies on the staking of the HYPE token and a validator-voting mechanism. According to its public roadmap and on-chain data, the system is currently in the critical period of transitioning from the second to the third stage of progressive decentralization [41]. In its early days, network consensus was maintained by a very small number of core nodes, which gave the team the ability to rapidly iterate its HyperBFT consensus algorithm and its liquidation engine. As the system matured, the governance structure underwent significant decentralization. The number of validator nodes expanded from an initial handful operated

by the foundation (about 4 to 5) to 21 active validators (the permissionless validator set introduced starting April 2025), and was further expanded to approximately 27 between late 2025 and mid-2026; its permissionless-participation mechanism automatically admits any node whose stake ranking places it within the active-set size cap (then the top 21). However, an increase in the number of validators does not equate to a substantive increase in the degree of decentralization. If a few entities control more than one-third of the stake weight, that is enough to break its Byzantine fault tolerance (BFT) guarantee; the true degree of decentralization should be measured by the number of entities holding more than one-third of the stake (that is, the Nakamoto coefficient), rather than the nominal number of validators. The public disclosure of stake-concentration data is itself an important direction in which Hyperliquid's governance transparency needs improvement.

An audit of Hyperliquid's current governance state reveals several engineering challenges that urgently need to be resolved. First, in the distribution of voting power, although a stake-weight mechanism has been introduced, if tokens are highly concentrated in the hands of a few early participants, the system still faces the threat of whale manipulation. To mitigate this risk, the system needs to further explore nonlinear weighting mechanisms such as quadratic voting to strengthen the voice of the broad retail base in governance. Second, the actual exercise of emergency powers has already directly tested the theoretical framework proposed in Section 30.6.2. During the JELLY perpetual futures incident of March 2025, after detecting large-scale market manipulation, Hyperliquid validators used an emergency vote to freeze trading in the JELLY perpetual futures and forcibly settled the affected positions at a price favorable to the HLP (market-making) vault (approximately \$0.0095, rather than the manipulated market price of about \$0.50). This action in effect exceeded the scope of "defensive operations" as defined in Section 30.6.2: suspending trading is a defensive measure, but unilaterally setting the settlement price constitutes a direct disposition of users' property. The incident triggered heated debate in the community: do validators have the right to exercise quasi-judicial adjudicatory power in the name of protecting the system's solvency? Although the HLP vault's losses were ultimately contained (the vault even recorded a small surplus in the end), the absence of a community accountability mechanism—no formal ex post review process and no independent operational audit report—exposed an engineering gap in the "ex post accountability" component of the emergency-powers framework. The JELLY incident demonstrates that the boundaries of emergency powers cannot be delimited solely through predefined operation types; they also require mandatory post-operation review and community ratification procedures.

Moreover, in the broader design of the emergency-response mechanism, how to improve the transparency and ex post accountability of emergency operations while ensuring network security is a key focus for optimizing the current governance architecture. In December 2025, a validator vote to permanently remove approximately 37 million HYPE (more than 13% of circulating supply) held in the Assistance Fund from the supply passed on December 24 with approximately 85% approval, demonstrating the community's capacity to participate in major economic decisions; yet the irreversibility of that decision, combined with a voting structure in which stake weight was highly

concentrated, still showed how such decisions depend on an adequate time window and a broad basis of consensus [42].

Looking ahead, Hyperliquid's governance architecture needs to deepen along two dimensions. On the technical dimension, as HyperEVM is fully deployed, the governance logic should be embedded more deeply into smart contracts to achieve fully automated proposal execution and reduce reliance on off-chain coordination. On the mechanism dimension, the system should introduce a more granular parameterized-governance framework, separating the authority over routine operations such as fee adjustments and margin-ratio settings from the authority over underlying consensus upgrades, thereby building a layered governance structure. The key bottleneck for this layering is that Hyperliquid's core parameters (margin ratios and liquidation thresholds) run in the consensus-layer state machine rather than the EVM contract layer: parameter changes to the former must be implemented through consensus upgrades, whereas the latter can be managed directly through governance contracts on HyperEVM, and passing permissions between the two layers requires the precompiles interface.

30.7 Mapping regulatory requirements to technical implementations

Having explored the layered architecture of on-chain financial systems, the design of the liquidation engine, and the engineering implementation of embedded compliance, this section returns to the regulatory framework proposed in Chapter 26 to form a complete loop between regulatory requirements and technical implementations. Building on the structural risks of traditional centralized exchanges, Chapter 26 proposed six "minimal but effective" regulatory principles. This section maps these principles one by one onto the concrete technical implementation schemes discussed in this chapter, shows how embedded compliance converts external regulatory constraints into built-in architectural guarantees, explores the possibility of regulation and technology forming a combined force, and clarifies the domains in which human judgment remains indispensable within a highly automated system.

30.7.1 The technical implementation of the six principles

The six regulatory principles proposed in Chapter 26 set goals for building a secure crypto-finance system, and this chapter's three-tier compliance spectrum (hard-coded, parameterized, and reporting-based compliance) provides an engineering implementation path for those goals. Table 30-1 maps each of these six principles onto the different compliance tiers of an on-chain system, showing the concrete technical implementation scheme and implementation priority corresponding to each principle. The table makes clear that asset segregation, as the highest-priority security baseline, is assigned to the hard-coded tier; principles that require continuous data feedback, such as disclosure and evidence-based evolution, fall into the reporting-based tier; and functional regulation and the localized adjustment of international standards depend on the flexibility of the parameterized mechanism. This tiered mapping embodies the differentiated demands of different regulatory requirements for embedding depth and flexibility.

Chapter 26 regulatory principle	Compliance tier	On-chain technical implementation	Implementation priority
Asset segregation as the baseline	Hard-coded compliance	Smart-contract self-custody and non-appropriation logic	Highest
Disclosure first	Reporting-based compliance	Real-time on-chain data extraction and automated audit APIs	High
Functional regulation	Parameterized compliance	Independent compliance parameters and governance mechanisms for each functional module	Medium
Embrace embedded compliance	Combined tiers	Coordinated design of hard-coded, parameterized, and reporting-based mechanisms	Medium
International minimum standards with flexible adjustment	Hard-coded and parameterized combined	Minimum standards hard-coded; flexible parameters adjusted by regional governance	Medium
Evidence-based evolution	Reporting-based compliance	Continuous evaluation of on-chain data to support dynamic optimization of compliance rules	Long-term

Table 30-1. Mapping between the six regulatory principles of Chapter 26 and their on-chain technical implementations (Data source: compiled by the author)

The deeper logic of this mapping lies in matching embedding depth to the nature of the rule. Non-negotiable security baselines such as asset segregation are entrusted to the deepest hard-coded compliance, whose architectural-level constraint is more reliable than a traditional legal pledge and ex post audit; disclosure-first and evidence-based evolution rely on the continuous data feedback that reporting-based compliance provides, converting periodic static reports into a dynamic display that can be computed and verified in real time; functional regulation draws on parameterized compliance so that the compliance parameters of each module—trading, liquidation, market making—can be flexibly adjusted through governance without altering the underlying core logic; embracing embedded compliance runs through the coordination of all three; and the coexistence of a globally uniform security baseline (such as the absolute leverage cap) with region-specific requirements (such as the margin ratio for specific assets) is precisely what the nested design of hard-coding and parameterization carries.

This mapping framework faces two unresolved structural challenges in implementation. The first is the enforcement mechanism for cross-border jurisdiction. The "regional governance adjustment" envisioned in Table 30-1 lacks a landing path in a permissionless, global on-chain system: in a smart contract accessible by anyone from any jurisdiction, the system cannot reliably determine the jurisdiction to which a user belongs,

and techniques such as VPNs can easily circumvent geographic restrictions based on IP addresses. Jurisdiction-aware access control based on zero-knowledge credentials offers a possible technical solution, allowing users from different jurisdictions to obtain differentiated parameter configurations after proving their compliance status; but the engineering maturity and regulatory acceptance of this scheme remain at an early stage. The second challenge is the absence of systemic-risk monitoring. The six principles above focus mainly on compliance at the level of a single protocol, but in a highly interconnected on-chain financial ecosystem, cross-protocol correlated liquidations, liquidity contagion, and the common-mode failure caused by shared oracles constitute a distinct category of systemic risk. This dimension has not yet been incorporated into the current system of regulatory principles, but it has become an unavoidable compliance requirement in the process of scaling on-chain finance.

30.7.2 The synergy of regulation and technology

Traditional financial regulation relies mainly on external constraints on institutional behavior—stipulating what an institution may and may not do, and using periodic inspections and severe penalties to compel compliance. The premise of this model is that the institution is capable of violating the rules, and the essence of compliance is to change the cost-benefit expectations of doing so. Embedded compliance, however, offers an entirely new paradigm: regulators can no longer merely require institutions to "pledge to comply with the rules" but can require them to adopt a system architecture that makes it "impossible to violate the rules."

When regulators require that "assets must be segregated," relying solely on periodic financial audits means, in effect, that they bear enormous information-lag risk. But in an on-chain financial system, regulators can require the platform to adopt a non-custodial smart-contract architecture. Under such an architecture, asset segregation is not an operational activity requiring continuous oversight but an architectural property built in at the moment of system deployment. The focus of regulation shifts from "checking whether the institution has misappropriated assets" to "auditing whether the smart-contract code does indeed restrict the authority to transfer assets." This shift significantly reduces the ongoing cost of enforcement while significantly improving the security of assets.

Similarly, for disclosure requirements, regulators can push the industry to adopt standardized on-chain data interfaces. Rather than requiring platforms to submit complex financial statements each month, it is better to require them to provide a regulatory API that can read on-chain state in real time. This enables regulators to build automated monitoring systems that detect abnormal trading patterns and the accumulation of systemic risk in real time. The driving force of regulation is no longer merely the threat of punishment but the impetus to evolve financial infrastructure toward a more transparent and resilient technical architecture.

The synergy of regulation and technology is also reflected in the standardization of governance mechanisms. Regulators can require that parameter adjustments and contract upgrades in an on-chain system follow specific timelock-delay and multisig-security standards. Such a requirement does not interfere with the community's specific governance

decisions but ensures that the governance process itself has sufficient transparency and predictability, providing market participants a buffer period to assess and respond to rule changes. By requiring good architecture, regulators not only achieve the goal of protecting investors but also promote innovation and development in financial technology.

However, the premise of this paradigm shift is that regulators possess the technical capacity to audit smart-contract code—a capacity that at present barely exists anywhere in the global financial-regulatory system. The core technical capabilities of current financial regulators are concentrated in legal analysis and financial auditing, not cryptography and smart-contract engineering. If regulators cannot independently verify whether contract code truly implements the compliance constraints it claims, "auditing the code" may degenerate into another form of "compliance theater," relying solely on third-party audit reports while the auditing firms themselves lack a standardized accountability mechanism. The transition from "auditing institutional behavior" to "auditing system architecture" therefore requires the concurrent advancement of at least two areas of infrastructure: first, establishing certified smart-contract auditing standards (analogous to the international standards in the accounting-audit field) to provide a unified framework for the scope, methods, and report format of contract audits; and second, building a tiered certification and accountability mechanism for auditing firms during the transition period, to ensure the credibility of third-party audit conclusions.

30.7.3 Compliance functions that technology cannot replace

Although embedded compliance solves a large number of traditional compliance problems through hard-coding and automation mechanisms, in a complex financial system the application of technology has its inherent boundaries. Certain key compliance functions and the handling of extreme situations still require human judgment, discretion, and social consensus.

First, dispute resolution and legal appeals cannot be fully replaced by code. A smart contract can only execute operations according to predetermined logic; it cannot judge whether that logic is fair in a specific situation. When an unexpected vulnerability causes user losses, or when an extreme market-manipulation event occurs, the victims need to seek redress. On-chain systems lack an intrinsic judicial-adjudication mechanism, and for the determination of complex liability and the recovery of losses, one must still rely on the real-world legal system and the rulings of human judges.

Second, cross-border enforcement coordination and complex crime investigation require human involvement. Although the transparency of on-chain data helps trace fund flows, linking an on-chain address to a real-world entity's identity often requires transnational intelligence sharing, judicial assistance, and traditional investigative methods. Technology can provide clues and evidence, but combating transnational financial crime and enforcing sanctions lists are essentially exercises of state power that require the collaboration of diplomatic and legal professionals.

Moreover, the identification of new risks and the evolution of the compliance framework require human strategic vision. An automated monitoring system can only detect

anomalies based on known patterns; it often lacks the forward-looking capacity to identify market manipulation carried out through entirely new mechanisms or the covert accumulation of systemic risk. Compliance experts and regulators must continuously analyze market dynamics and understand newly emerging financial instruments and trading strategies in order to formulate new regulatory principles and guide upgrades to the underlying architecture.

Finally, the enforcement of ethical standards and the ultimate handling of a systemic crisis cannot do without human social consensus. In extreme market turmoil, decisions such as whether to suspend trading or whether to invoke emergency governance powers to intervene often involve complex trade-offs among different interest groups. A purely technical system cannot bear this kind of responsibility, which carries a high degree of political and ethical weight. Humans set the rules at the outset of the system's design, and they must also bear ultimate decision-making responsibility when the system faces a major systemic crisis. Embedded compliance does not eliminate the role of humans; rather, it frees human energy from tedious day-to-day compliance checks and concentrates it on higher-level architectural design, risk anticipation, and crisis management.

30.8 Chapter summary

The three-tier compliance-spectrum architecture proposed in this chapter (hard-coded, parameterized, and reporting-based compliance) provides a systematic analytical framework and implementation path for the compliance engineering of on-chain financial systems. Hard-coded compliance, by compiling non-negotiable security baselines such as asset segregation directly into smart-contract logic, achieves a paradigm shift from "requiring institutions to obey the law" to "making it impossible for institutions to break it." Parameterized compliance, within a fixed rule framework, allows operational parameters such as the margin ratio to be dynamically adjusted through transparent governance mechanisms, striking an engineering balance between security and flexibility. Reporting-based compliance leverages the immutability and global visibility of on-chain data to convert ex post auditing into an inherent property of the system. The organic combination of the three completes the full mapping from regulatory requirements to on-chain engineering implementation.

The introduction of zero-knowledge-proof technology breaks the zero-sum game between regulatory transparency and user privacy [26]. Through cryptographic proofs, traders can prove to the system that they satisfy specific compliance conditions without exposing the underlying identity data, allowing on-chain finance to satisfy regulatory standards such as anti-money laundering while preserving its core permissionless character. This technical path demonstrates the potential of a digital-native architecture to surpass traditional finance on the compliance dimension.

The inevitable cost of embedding rules into architecture is an increase in system rigidity. The dual architecture of timelock and emergency powers provides an engineering solution to this contradiction: routine changes are subjected to a mandatory delay for community review in order to maintain immutability, while a constrained fast track for

extreme situations ensures the system's capacity for crisis response [43]. This controllable upgradeability, combined with the governance roadmap of progressive decentralization, constitutes the optimal practice for on-chain governance under current technical conditions. However, functions such as dispute resolution, cross-border enforcement coordination, and the forward-looking identification of new risks still lie beyond the capacity boundary of a purely technical system and require human judgment and social consensus to supplement.

Hyperliquid's app-chain architecture and its on-chain central limit order book constitute a frontier practical prototype of embedded compliance, but a significant gap remains between the current prototype and a mature financial infrastructure. Along this evolutionary path, a new force is joining: artificial intelligence. When AI agents can autonomously conduct market making, trading, and risk management on-chain, the operating mode of financial markets will once again undergo a profound change. This evolutionary trend is precisely the core theme explored in Part XI.

References

[1] Ray, J. J. (2023). *First Interim Report of John J. Ray III to the Independent Directors on Control Failures at the FTX Exchanges*. FTX Debtors. https://www.fishmanhaygood.com/wp-content/uploads/2023/04/April-9-Debtor-Report-on-Failure-of-Internal-Controls-1.pdf

[2] U.S. Department of Justice. (2024). *Samuel Bankman-Fried Sentenced to 25 Years for His Orchestration of Multiple Fraudulent Schemes*. https://www.justice.gov/archives/opa/pr/samuel-bankman-fried-sentenced-25-years-his-orchestration-multiple-fraudulent-schemes

[3] Auer, R. (2019). Embedded supervision: how to build regulation into blockchain finance. *BIS Working Papers*, No 811. Bank for International Settlements. https://www.bis.org/publ/work811.pdf

[4] Thomson Reuters. (2024). *Cost of Compliance Report*. Thomson Reuters Regulatory Intelligence.

[5] Deloitte. (n.d.). *Cost of compliance and regulatory productivity*. Deloitte. https://www.deloitte.com/us/en/services/consulting/articles/cost-of-compliance-regulatory-productivity.html

[6] Research and Markets. (2024). *Global anti-money laundering systems market 2024–2028*. GlobeNewswire. https://www.globenewswire.com/news-release/2024/06/12/2897441/28124/en/Global-Anti-money-Laundering-Systems-Market-2024-2028.html

[7] Auer, R. (2022). Embedded supervision: How to build regulation into decentralized finance. *Crypto Economic Systems*. https://cryptoeconomicsystems.pubpub.org/pub/auer-embedded-supervision/release/1

[8] Lessig, L. (2006). *Code: Version 2.0*. Basic Books.

[9] Jensen, J. R., von Wachter, V., & Ross, O. (2021). Leveraged trading on blockchain technology. *arXiv preprint arXiv:2102.13488*. https://doi.org/10.48550/arXiv.2102.13488

[10] Aquilina, M., Frost, J., & Schrimpf, A. (2024). Decentralized finance (DeFi): a functional approach. *Journal of Financial Regulation*, *10*(1), 1-25. https://doi.org/10.1093/jfr/fjad013

[11] Nabilou, H. (2022). The Law and Macroeconomics of Custody and Asset Segregation Rules: Defining the Perimeters of Crypto-Banking. *Amsterdam Law School Research Paper*.

[12] Zetzsche, D. A., Arner, D. W., & Buckley, R. P. (2020). Decentralized finance. *Journal of Financial Regulation*, *6*(2), 172-203.

[13] Bhargavan, K., Delignat-Lavaud, A., Fournet, C., Gollamudi, A., Gonthier, G., Kobeissi, N., … & Zanella-Beguelin, S. (2016). Formal verification of smart contracts: Short paper. In *Proceedings of the 2016 ACM Workshop on Programming Languages and Analysis for Security* (pp. 91-96).

[14] He, S., Manela, A., Ross, O., & von Wachter, V. (2022). Fundamentals of perpetual futures. *arXiv preprint arXiv:2212.06888*. https://doi.org/10.48550/arXiv.2212.06888

[15] Ackerer, D., Hugonnier, J., & Jermann, U. (2025). Perpetual futures pricing. *Mathematical Finance*. https://doi.org/10.1111/mafi.70018

[16] Chainlink. (2026). *Onchain Governance: How Smart Contracts Enable Decentralized Control*. https://chain.link/article/onchain-governance

[17] Ali, Z. (2025). Anatomy of the Oct 10-11, 2025 Crypto Liquidation Cascade: Macroeconomic Triggers, Market Microstructure, and Systemic Risk Lessons. *SSRN Electronic Journal*.

[18] Bybit. (2025). *Maintenance Margin Calculation (USDC Perpetual & Futures)*. https://www.bybit.com/en/help-center/article/Maintenance-Margin-Calculation-USDC-Contract

[19] Chen, E., Ma, M., & Nie, Z. (2024). Perpetual future contracts in centralized and decentralized exchanges: Mechanism and traders' behavior. *Electronic Markets*.

[20] Werapun, W., Karode, T., Suaboot, J., & Arporntthip, T. (2025). Exploring Risk and Return Profiles of Funding Rate Arbitrage on CEX and DEX. *Blockchain: Research and Applications*.

[21] Kao, H. T., Chitra, T., Chiang, R., & Morrow, J. (2020). An analysis of the market risk to participants in the compound protocol. *International Symposium on Cyber Security Cryptography and Machine Learning*.

[22] Xu, L. (2024). Blockchain Transaction Monitoring and Anomaly Analysis System. *IEEE Access*. https://doi.org/10.1109/dikw64278.2024.00015

[23] Lazirko, M., Appelbaum, D., & Vasarhelyi, M. (2025). Proof of reserves: A double-helix framework. *The British Accounting Review*.

[24] Azgad-Tromer, S., Garcia, J., & Tromer, E. (2023). The Case for On-Chain Privacy and Compliance. *Stanford Journal of Blockchain Law & Policy*. https://stanford-jblp.pubpub.org/pub/onchain-privacy-compliance

[25] U.S. Department of the Treasury. (2022). U.S. Treasury Sanctions Notorious Virtual Currency Mixer Tornado Cash. https://home.treasury.gov/news/press-

releases/jy0916

[26] Decker, N. (2025). Proof Without Exposure: Zero-Knowledge Proofs as a Cryptographic Framework for Institutional Financial Compliance. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5170329 https://doi.org/10.2139/ssrn.5170329

[27] Polygon Technology. (2022). Introducing Polygon ID, Zero-Knowledge Identity for Web3. https://polygon.technology/blog/introducing-polygon-id-zero-knowledge-own-your-identity-for-web3

[28] zkMe. (2024). KYC Explained: Securing Your Identity in the Web3 World. https://medium.com/@zkMe/kyc-explained-securing-your-identity-in-the-web3-world-8cc2a4b47f15

[29] De Filippi, P., Mannan, M., & Reijers, W. (2020). Blockchain as a confidence machine: The problem of trust & challenges of governance. *Technology in Society*, *62*, 101284.

[30] OpenZeppelin. (2021). *Protect Your Users With Smart Contract Timelocks*. https://www.openzeppelin.com/news/protect-your-users-with-smart-contract-timelocks

[31] Joshi, S. (2023). *Timelock Contracts and DeFi Security: Lessons from Compound*. Medium. https://medium.com/@srinivasjoshi66/timelock-contracts-and-defi-security-lessons-from-compound-fe24f3e4574b

[32] Elem, O., & Talmon, N. (2026). Legitimate Overrides in Decentralized Protocols. *arXiv preprint arXiv:2602.12260*. https://doi.org/10.48550/arXiv.2602.12260

[33] Qasse, I., Hamdaqa, M., & Jonsson, B. T. (2026). Immutable in principle, upgradeable by design: Exploratory study of smart contract upgradeability. *Empirical Software Engineering*, *31*(1), 1-35.

[34] Qin, K., Zhou, L., Livshits, B., & Gervais, A. (2021). Attacking the DeFi ecosystem with flash loans for fun and profit. In *Financial Cryptography and Data Security: 25th International Conference, FC 2021*. Springer.

[35] MixBytes. (2022). *DAO Voting Vulnerabilities*. https://mixbytes.io/blog/dao-voting-vulnerabilities

[36] Lalley, S. P., & Weyl, E. G. (2018). Quadratic voting: How mechanism design can radicalize democracy. *AEA Papers and Proceedings*, *108*, 33–37. https://doi.org/10.1257/pandp.20181002

[37] Wright Jr, D. (2019). Quadratic voting and blockchain governance. *UMKC Law Review*, *88*, 451.

[38] Benhaim, A., Falk, B. H., & Tsoukalas, G. (2025). Balancing power in decentralized governance: Quadratic voting and information aggregation. *Management Science*. https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2024.08469 https://doi.org/10.1287/mnsc.2024.08469

[39] Walden, J. (2020). *Progressive Decentralization: A Playbook for Building Crypto Applications*. a16z crypto. https://a16z.com/progressive-decentralization-a-playbook-for-building-crypto-applications/

[40] Walden, J. (2023). *Progressive decentralization: a high-level framework*. a16zcrypto. https://a16zcrypto.com/posts/article/progressive-decentralization-a-high-level-framework/

[41] Hyperliquid. (2025). *Roadmap | Hyperliquid Wiki*. https://hyperliquid-co.gitbook.io/wiki/introduction/roadmap

[42] The Defiant. (2025). *Hyperliquid Proposes Burning 13% of Circulating Supply*. https://thedefiant.io/news/tokens/hyperliquid-proposes-burning-13-percent-of-circulating-token-supply

[43] Wang, Z., Pu, F., Cheung, V., & Hao, R. (2025). Balancing security and liquidity: A time-weighted snapshot framework for DAO governance voting. *arXiv preprint arXiv:2505.00888*. https://arxiv.org/abs/2505.00888 https://doi.org/10.48550/arXiv.2505.00888

Part X
The Future

Core question: Where is all this headed?

When AI Becomes Market Infrastructure

The large-scale entry of artificial intelligence (AI) agents is driving a fundamental transformation in the microstructure of financial markets. Breakthroughs in machine learning models—in data processing, pattern recognition, and autonomous decision-making—have structurally altered the informational dimension along which markets operate, with profound effects on the speed of trade execution, the patterns of liquidity provision, and the mechanisms of price discovery. The scale and depth of this transformation are comparable to the way the emergence of vision during biological evolution reshaped the competitive landscape of ecosystems [1][2]: a leap in the capacity to acquire and process information has changed the game-theoretic structure among market participants.

The complete framework for analyzing market microstructure built up over the preceding thirty chapters—whether examining the pricing mechanisms of automated market makers (AMMs), the impermanent loss borne by liquidity providers, the behavioral patterns of arbitrageurs, or the governance mechanisms of protocols—rests on one implicit core assumption: that market participants are human, or algorithms under real-time human control and parameter-setting. This assumption implies that market participants inevitably carry biological and psychological limitations. A large body of behavioral finance research shows that human participants are constrained by limited attention bandwidth, emotionally driven irrational reactions, and an inescapable need for sleep [3]. During extreme market volatility, the contagion of panic often leads to a synchronized withdrawal of liquidity; when processing complex information, cognitive-load limits make it impossible for humans to track the relationships among thousands of variables simultaneously. These features, rooted in human nature, constitute the microfoundations of liquidity crises, volatility clustering, and delayed price discovery in traditional market microstructure.

When AI agents shift from decision-support tools to autonomous market participants, however, the assumption above is fundamentally altered. AI agents are not bound by biological rhythms and can operate continuously around the clock; they do not experience panic or greed, and can rigorously execute preset risk-control logic even in extreme conditions. Yet the absence of emotional reactions at the execution layer does not mean the entire decision chain is free of emotion. An AI system's risk-control parameters are set by humans and inevitably embed the setter's risk preferences and

behavioral biases; its training data derive entirely from the historical record of human trading behavior, so the model may inherit and reproduce systematic human biases; and in extreme conditions, the decision to "shut down the system" or "dial risk parameters to their most conservative settings" is made by human managers, a step wholly subject to emotion. A more precise formulation is therefore that AI removes emotion from the execution layer, but emotion continues to influence the market indirectly through parameter-setting, training data, and human override decisions. Even so, AI's information-processing bandwidth far exceeds that of humans, enabling high-dimensional, synchronized optimization across multiple markets, assets, and time scales at once [4]. More critically, a superior strategy model can be perfectly replicated and deployed at scale at low cost. When the bulk of market volume is no longer generated by the interplay among human traders but instead by the autonomous interaction of AI agents, the core dynamics of market microstructure change structurally. The essence of adverse selection shifts from an asymmetry grounded in information access to one grounded in algorithmic processing speed and model predictive power; the provision and consumption of liquidity are no longer governed by human panic, but instead by a new form of systemic fragility arising from strategy homogenization.

The pervasive infiltration of AI agents forces us to reexamine the applicability of existing theoretical frameworks. The technological competition of the high-frequency trading era has already shown that technological progress not only compresses transaction costs but also drives fundamental changes in the composition of market participants [5]. In AI-dominated markets, the profit equation of automated market makers must be recalibrated to fit a new environment of multidimensional inventory management and adversarial market making; the reflexive character of liquidity takes on a different form, as the heterogeneous decisions of human market makers may be replaced by the synchronized reactions of AI agents—eliminating routine panic while creating the risk of ultra-fast flash crashes; and the decomposition of volatility sources must also incorporate an additional volatility term generated by algorithmic homogenization. These changes impose new requirements not only on the strategies of traders and liquidity providers, but also on the design and governance of decentralized finance (DeFi) protocols.

This chapter aims to systematically analyze how AI is reshaping market microstructure, advancing theoretical contributions along five dimensions: a four-stage model of AI penetration into microstructure, delineating the evolutionary path of markets from decision support to autonomous strategic interaction; a reexamination matrix that tests, one by one, the validity and the direction of revision of the core concepts of market making, liquidity, volatility, and market quality from earlier chapters in an AI world; the identification and analysis of AI-native new phenomena such as strategy-homogenization risk, ultra-fast flash crashes, and adversarial market making; a dual-role framework that reveals the economic logic of AI's evolution from a value-extracting market participant into value-creating market infrastructure; and a hybrid architecture fusing code and models, which provides theoretical support for the design of future DeFi protocols.

Structurally, the analysis proceeds along two main lines. The first half focuses on the impact of AI as a market participant, examining in detail how it alters the economics of

market making, reshapes the dynamics of liquidity and volatility, and exerts complex effects on the five dimensions of market quality. The second half turns to a deeper analysis of the architectural upgrade, arguing that AI transcends the role of participant to become a key component of core infrastructure such as dynamic risk engines, intelligent matching systems, and semantic oracles. Through the interweaving of these two lines, the chapter presents a new, AI-driven market paradigm, laying the theoretical groundwork for understanding the operating mechanisms of future autonomous financial networks.

A note on scope: the core object of analysis in this book is the order-book-based perpetual futures market on centralized exchanges. AI's influence on market microstructure, however, has a cross-market character; the same algorithmic architecture and strategy logic are often deployed simultaneously on centralized exchanges and on-chain protocols. Accordingly, while this chapter takes the centralized-exchange (CEX) order-book perpetual futures market as its primary setting, part of the discussion extends to decentralized exchanges (DEXs) and on-chain protocols in order to present a complete picture of how the advance of AI technology affects market microstructure. Readers should take care to distinguish which market environment a given passage addresses.

31.1 A four-stage model of AI penetration into microstructure

Market microstructure theory has long rested on a basic assumption: that market participants possess the cognitive characteristics and behavioral limitations of humans. These inherent cognitive and emotional limits (see the chapter introduction) jointly shape the clustering of market liquidity and the pulsed rhythm of price discovery. When algorithmic agents with extremely high concurrent processing capacity, millisecond-level response speeds, and no emotional fluctuations gradually take over core trading functions, however, this implicit assumption faces a fundamental challenge. To assess systematically the effect of non-human participants on market microstructure, we need an analytical framework that describes the evolving path of intelligent agents' roles in the market. Such a framework can not only explain phenomena already observed in current markets, but also anticipate the new market structures and trading dynamics that may emerge in the future.

Figure 31-1 presents the four-stage model of AI penetration into microstructure. Its central classificatory principle is to distinguish two dimensions—"who makes decisions" and "who executes them"—and thereby capture the gradual transfer of power to AI in the market.

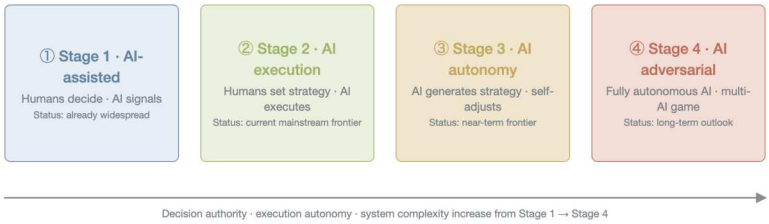


Figure 31-1. The four-stage model of AI penetration into market microstructure (Data source: theoretical derivation based on the evolutionary path of algorithmic trading)

31.1.1 AI assistance

In the initial stage of penetration, AI primarily plays a supporting role in data processing and signal extraction. At this point, the core decision-making authority over market microstructure still rests with human traders or fund managers. Algorithmic systems are used to identify potential trading opportunities from vast quantities of unstructured data—for example, analyzing news sentiment via natural language processing (NLP), or using machine learning models to detect minute price anomalies in historical order-book data. These systems output risk alerts, probabilistic forecasts, or trade recommendations to humans, but the final instructions to open positions, close positions, and adjust risk exposure are still issued by humans.

In on-chain finance, Stage 1 AI is especially widely applied. On-chain data-analysis tools can track, in real time, the transfer behavior of large wallets, the interaction patterns of smart contracts, and the flow of capital across liquidity pools. A trader can, for example, monitor the wallet activity of whale addresses to infer potential large-order intentions and adjust positions in advance. Similarly, machine learning models can analyze the correlations among historical liquidation events, funding-rate changes, and price volatility to provide risk alerts for human traders. On decentralized exchanges, AI systems can compute the expected impermanent loss of each liquidity pool in real time, helping liquidity providers evaluate the risk-return trade-offs across different asset pairs. Although these applications improve the efficiency of information processing, they do not change the human's position as final decision-maker.

The systemic impact of this stage on market microstructure is relatively limited. Because human decision-makers serve as the final execution node, the market's response speed is still constrained by human cognitive bottlenecks and physical reaction times. In extreme market conditions, humans' inherent fear and greed still dominate the supply of and demand for liquidity, and the withdrawal of market-maker liquidity and cascading declines in liquidity still occur frequently. Even at this stage, however, AI-assisted tools produce observable microstructure effects. Studies suggest that trading institutions using AI risk-alert systems adjust positions significantly faster in extreme conditions (an estimate based on the author's industry observation), a difference large enough to confer a meaningful execution advantage in a market operating at the millisecond level. In addition, AI-driven sentiment analysis can capture a shift in market sentiment within the first 100 milliseconds after a news release, providing a narrow

time window for fast-reacting traders. Thus, while the classic theoretical analyses of earlier chapters—information asymmetry, adverse-selection risk, and the tragedy of the liquidity commons—remain applicable at this stage, their specific parameters and time scales have already shifted subtly with the spread of AI-assisted tools.

31.1.2 AI execution

As algorithmic engineering matured, the market entered a second stage, in which humans set the logical framework of the strategy and the risk parameters while AI systems automatically execute trades within those predefined boundaries. This mode is already widespread among mainstream market makers and high-frequency trading firms. Human quantitative researchers define inventory-management preferences, hedge-ratio thresholds, and the base formula for quote width, and the algorithmic engine dynamically adjusts two-sided quotes at millisecond speed based on real-time market data flows, conducting high-frequency arbitrage across multiple exchanges.

On centralized exchanges, the degree of automation among mainstream market makers has already reached the level typical of Stage 2. At large exchanges such as Binance and Coinbase, official market-maker programs (such as Binance's market-maker incentive program) already rely entirely on automated quoting engines. Within a human-set parameter framework, and based on order-book depth, real-time volatility estimates, and price information from other exchanges, these engines can adjust their quotes hundreds of times per second. On decentralized exchanges, Hyperliquid's Hyperliquidity Provider (HLP) market-making vault is a parameterized automated market-making system whose strategy logic (quoting symmetrically around the mark price and adjusting the quote offset according to the degree of position deviation) is preset by humans and executed strictly according to the rules. Viewed through the lens of the four-stage model, this vault sits in the transitional zone between Stage 1 and Stage 2: its degree of automation is high, but it lacks machine-learning-based adaptive capability, and adjustments to strategy parameters still depend on manual human intervention rather than on the model's autonomous learning. When the market price fluctuates, its algorithmic engine automatically adjusts the concentration of liquidity according to preset rules, shifting the liquidity range upward when the price rises and downward when it falls, thereby executing market-making operations within human-set constraints.

At this stage, algorithmic execution greatly improves the microstructural efficiency of the market. Automated market-making engines can operate around the clock without interruption, significantly reducing bid-ask spreads under normal conditions. Data indicate that on exchanges that have introduced automated market-making engines, the average bid-ask spread has fallen significantly (according to industry reports, from the level of tens of basis points to single-digit basis points), an improvement that directly lowers costs for all traders. Automated systems also accelerate the speed at which new information is incorporated into asset prices, markedly improving the efficiency of price discovery. Yet even though the execution process is highly automated, the creativity and core logic of the strategy still originate with humans. When the market experiences an extreme black-swan event that exceeds the preset parameter framework, the system usually triggers a hardcoded circuit breaker, halting quotes and handing

control to humans. In one extreme volatility event, for example, several automated market-making systems immediately halted quoting upon detecting anomalous price movements, waiting for manual intervention by human traders. This stage therefore achieves a leap in speed and consistency, but it does not fundamentally eliminate the market's fragility in the face of structural shocks. Homogenized algorithmic reactions may even exacerbate short-term liquidity droughts under certain conditions, because multiple institutions' market-making engines may make the same decision based on the same market signal, causing liquidity to vanish in an instant.

31.1.3 AI autonomy

Recent frontier research shows that the market is accelerating toward a third stage, whose defining feature is that AI is responsible not only for execution but also for autonomously generating and dynamically adjusting trading strategies. Deep reinforcement learning algorithms are deployed in complex market environments, where agent programs, through continuous interaction with market states and driven by an objective function that maximizes long-run cumulative returns, autonomously explore optimal market-making and trading paths [6]. Unlike Stage 2, which is constrained by human-preset logic, autonomous learning agents can discover nonlinear patterns that human analysts struggle to identify and develop novel combinations of high-frequency strategies. For example, a market-making agent based on deep Q-learning might learn that, under a particular combination of order-book shape and volatility, it can probe the distribution of market liquidity by actively submitting fake orders and then immediately canceling them, and then adjust its own quoting strategy according to the market's reaction. Such behavior, however, legally constitutes spoofing, a form of market manipulation expressly prohibited in nearly every major jurisdiction. The 2015 charges brought by the U.S. Department of Justice and the Commodity Futures Trading Commission (CFTC) against Navinder Sarao (who was arrested in the United Kingdom that April and faced extradition) are a paradigmatic enforcement precedent aimed at precisely this kind of strategy. The key risk here is that a reinforcement learning agent, trained on a pure profit-maximization objective, spontaneously evolves a trading strategy that violates the law—something its developers could not foresee at the design stage. This phenomenon corroborates exactly the core regulatory dilemma this chapter will examine in depth in Section 31.6: an autonomous learning agent may independently generate unlawful conduct with no human intent involved whatsoever, and the resulting questions of liability attribution and ex ante prevention constitute one of the thorniest governance challenges of AI participation in financial markets.

The application of reinforcement learning in market making has already moved from theory into practice. Research institutions and hedge funds have begun deploying market-making agents based on the Actor-Critic architecture; these agents can operate simultaneously across multiple exchanges and asset pairs, using cross-asset correlations for dynamic hedging. A typical example: a reinforcement learning agent may learn that when Bitcoin's volatility rises, Ethereum's volatility typically rises about 200 milliseconds later, and so the agent can adjust Ethereum's quote width in advance, preparing for the risk before volatility rises. Such predictive strategy adjustment is impossible in a traditional Stage 2 system, because it is very difficult for a human to forecast so

precisely this kind of complex cross-asset time-series relationship.

This evolution has a marked and profound impact on market microstructure. On the one hand, autonomous learning models demonstrate a capacity far exceeding humans' in multidimensional inventory optimization and cross-asset correlation hedging, providing deeper liquidity support in normal markets. Studies suggest that institutions using reinforcement learning market-making strategies provide 30% to 50% more liquidity in normal markets than traditional market makers. On the other hand, the black-box nature of the strategy-generation process substantially increases the unpredictability of system behavior. When multiple institutions simultaneously deploy autonomous agents based on similar reinforcement learning architectures, these agents are highly prone to exhibiting homogenized reaction patterns when confronted with particular market-microstructure signals. This strategy convergence stems from the convergence of training data, neural-network architecture, and optimization objectives; its full causes are analyzed in detail in Section 31.3.2.

In extreme cases, such strategy convergence can trigger an instantaneous evaporation of liquidity, causing the market to experience violent price swings within milliseconds and forming a new source of risk at the microstructure level [7]. For example, in a recent simulation study, researchers deployed five market-making agents based on the same reinforcement learning architecture in a simulated market and found that, when a sudden price jump occurred, all five agents contracted their quote widths simultaneously in under 100 milliseconds, causing market liquidity to vanish in an instant and prices to swing to an extreme degree. This phenomenon, dubbed an "AI flash crash," represents an entirely new form of systemic risk that does not exist in traditional human-dominated markets. The interaction of five fully isomorphic agents in a single simulated market, as above, represents an extreme case of strategy homogenization. In real markets, although market makers' AI systems may use similar frameworks, their training-data windows, feature-engineering details, and risk-control parameter settings all differ, and the market also contains a large number of non-AI participants that provide a natural buffer of heterogeneity; the simulation results above should therefore be understood as an upper-bound estimate of homogenization risk.

31.1.4 AI adversarial competition

In the long-run picture, when the vast majority of trading volume and liquidity provision is contributed by highly autonomous intelligent agents, market microstructure will undergo a fundamental structural transformation. At that point, the core driving force of the market will no longer be psychological games among humans, but competition among algorithmic agents grounded in computing power, latency advantages, and model architecture. Traditional adverse-selection theory is built on the information asymmetry between informed and uninformed traders. In their classic 1985 model, Glosten and Milgrom showed that a market maker's bid-ask spread is a rational compensation for adverse-selection risk [8]. In this framework, the core risk facing the market maker is that informed traders will use their informational advantage to trade favorably, thereby inflicting losses on the market maker, who sets a wider spread to compensate for this risk.

In an all-algorithmic market, however, this game evolves into a contest between algorithms with superior predictive ability and those with relatively weaker predictive ability. In this environment, being "informed" no longer means possessing inside information; it means possessing a better predictive model and faster hardware (a full analysis of the transformation in the nature of adverse selection appears in Section 31.2.2). A higher-performing reinforcement learning agent may be able to predict another agent's quote adjustments and thus adjust its own position before those adjustments occur. This competition over "algorithmic advantage" leads to an endless technological arms race, in which every institution pours vast sums into developing faster hardware, more efficient algorithmic architectures, and more efficient data-processing pipelines.

This kind of ecosystem may give rise to entirely new market phenomena. Algorithmic agents may, without explicit communication or human intent, spontaneously form tacit collusion that sustains supra-competitive profits through trial and error and mutual adaptation [6]. For example, multiple autonomous market-making agents may jointly maintain wide bid-ask spreads to maximize collective returns, thereby harming the market's overall price efficiency and trader welfare [9]. In the 2025 study by Dou, Goldstein, and Ji, the researchers replaced the speculators in their model with reinforcement-learning-based "informed AI speculators" in simulation experiments built on a theoretical model, and found that these agents spontaneously sustained collusive, supra-competitive profits with no agreement, communication, or intent whatsoever (one of the mechanisms being a price-trigger strategy, rather than a literal bid-ask-spread cartel). This finding indicates that even without any explicit collusive intent, interactions among algorithmic agents may produce anticompetitive market structures.

In this environment, the classic liquidity models, the volatility-decomposition framework, and the standards for evaluating market quality all require a bottom-up reconstruction to fit a financial network wholly dominated by machine logic. The Glosten-Milgrom model's assumptions about adverse selection must be revised, because in an all-algorithmic market the definition of "informed" has changed. The assumption about the persistence of information in the Kyle model must also be reconsidered, because algorithmic agents can fully digest new information within milliseconds, sharply reducing the persistence of information. These revisions are not merely academic details; they are a profound reexamination of the theoretical foundations of market microstructure.

31.1.5 Assessing the stage of penetration

In assessing the evolutionary stage the market currently occupies, one must recognize that AI's speed of penetration varies markedly across different market functions. This asymmetry stems from the different requirements each task places on data dimensionality, execution latency, and decision error tolerance. Certain functions, owing to their highly structured nature and extreme dependence on speed, have already crossed multiple stages rapidly, while others, involving complex judgment and high-stakes consequences, remain at an early stage.

Table 31-1 shows the current position and evolutionary characteristics of each core market function along the AI-penetration path. The table maps four core functions—market making and arbitrage, directional trading, risk control, and governance

participation—to different positions within the four-stage model, revealing a striking asymmetry: market making and arbitrage, which depend most heavily on speed and deterministic logic, have already crossed into Stage 2 through Stage 3, whereas governance participation, which involves complex games over competing interests, remains before Stage 1. This differentiated pace of penetration implies that, for a considerable time to come, the market will exhibit a hybrid form in which human decision-making and machine autonomy are deeply interwoven.

Market function	Current penetration stage	Core driver	Evolutionary characteristics and challenges
Market making and arbitrage	Stage 2 to Stage 3	Millisecond-level latency requirements and high-dimensional data processing	Highest degree of automation; transitioning toward reinforcement learning agents; faces strategy-homogenization risk
Directional trading	Stage 1 to Stage 2	Macro-data integration and long-term trend forecasting	Relies on human intuition to set hypotheses; algorithms mainly used to optimize execution paths and reduce slippage
Risk control	Stage 1 to Stage 2	Prevention of extreme tail risk and preservation of capital	Still requires humans to set hard risk floors; hard to delegate systemic-risk judgment entirely to black-box models
Governance participation	Before Stage 1	Complex games, interest coordination, and rule-making	At the proof-of-concept stage; AI agents voting on protocol parameters face the obstacle of missing accountability mechanisms

Table 31-1. The AI-penetration stage and evolutionary characteristics of each market function (Data source: industry research and literature review [6][7][9])

Because market making and high-frequency arbitrage depend heavily on millisecond-level reaction speed and deterministic mathematical logic, their automation has advanced the fastest of all market functions. On centralized exchanges, mainstream market makers already rely entirely on automated systems, and the role of human traders has shifted from directly executing trades to supervision and parameter adjustment. On decentralized exchanges, liquidity providers on AMM protocols have widely adopted automated position-management tools. This domain has now steadily crossed Stage 2 and is exploring the autonomous-learning mode of Stage 3. Multiple research institutions and hedge funds have begun deploying reinforcement-learning-based market-making agents in production; although the scale remains relatively small, growth is rapid.

Directional trading and risk management must process more low-frequency, unstruc-

tured macro variables and have an extremely low tolerance for wrong decisions. When deciding whether to increase a Bitcoin position, for example, a fund manager must weigh multiple dimensions of information: the macroeconomic situation, the regulatory environment, technological developments, and market sentiment. Much of this information is qualitative and hard to quantify, and thus difficult for a fully automated system to process. Humans still retain core control in these domains, with algorithms used mainly to optimize execution paths (for instance, reducing slippage via volume-weighted average price (VWAP) or time-weighted average price (TWAP) algorithms) rather than to make core investment decisions.

As for governance participation in decentralized protocols, this remains a domain that AI can scarcely reach, because it involves complex games over competing interests and social consensus. Although it is technically possible to let intelligent agents participate in decentralized autonomous organization (DAO) voting, doing so raises a series of questions: Who is responsible for an AI agent's voting decisions? Could AI agents be exploited by malicious actors to manipulate the governance process? How can we ensure that an AI agent's voting behavior aligns with the protocol's long-term interests? These questions have not yet been adequately resolved, and so governance participation remains confined to the human-dominated sphere.

This uneven pace of penetration indicates that, for a considerable time to come, market microstructure will exhibit a hybrid form in which human decision-making and machine autonomy coexist. Market making and arbitrage will be the first to enter Stage 3 and even Stage 4, while other domains may remain at Stage 1 or Stage 2. This differentiated evolutionary process will create new arbitrage opportunities and risks. A hybrid system able to coordinate between Stage 3 automated market making and Stage 2 directional trading, for example, might gain a significant competitive advantage. At the same time, this hybrid form will also increase the market's complexity and unpredictability, because interactions among participants at different stages may produce novel microstructure phenomena never observed before.

31.2 The new economics of AI market making

The market-making economics framework built in Chapter 19 rests on an implicit assumption: that the market maker is an entity controlled by a human trader, constrained by human reaction speed and the inherent cognitive and emotional limits of humans (see the chapter introduction). When AI agents take over the market-making function, this assumption breaks down. AI agents can process multidimensional data at the millisecond level, run around the clock without rest, and remain unaffected by panic, which fundamentally changes the cost structure and behavioral patterns of the market maker. This section reexamines the classic market-making profit equation, analyzes the new features of adverse selection under AI dominance, and explores multidimensional inventory management and the novel phenomenon of adversarial market making, ending with a comparison of the structural strengths and weaknesses of AI and human market makers.

31.2.1 Recalibrating the profit equation

The traditional market-maker profit equation comprises spread income, adverse-selection cost, inventory-management cost, and fixed operating cost. The introduction of AI agents does not simply raise the efficiency of every variable in this equation; rather, it recalibrates the cost structure asymmetrically. AI technology significantly compresses variable costs, but its effect on fixed institutional costs is limited.

The compression of variable costs is manifested primarily in adverse selection and inventory management. In managing adverse selection, AI—via reinforcement learning algorithms—can more precisely predict short-term price movements and thus adjust quotes before informed traders arrive. Zhao and Linetsky (2021) [10] show that market-making algorithms trained through reinforcement learning can effectively control adverse-selection risk in high-frequency trading environments. In inventory management, AI can handle more complex optimization problems, reducing the capital-occupancy cost of inventory buildup. The model of Herrmann et al. (2020) [11] shows that the advantage of high-frequency trading algorithms in inventory management directly affects their Nash equilibrium state in a competitive environment. Notably, the theoretical foundation of these AI market-making strategies can be traced back to the classic stochastic optimal-control market-making model proposed by Avellaneda and Stoikov (2008) [12]. That model formalizes the market-making problem as an optimal trade-off between inventory risk and spread income, providing a theoretical benchmark for subsequent reinforcement learning market-making strategies. Contemporary AI market makers are, in essence, building on the Avellaneda-Stoikov framework, using deep reinforcement learning to extend the state space from a low-dimensional analytical solution to high-dimensional nonlinear optimization.

The fixed institutional costs of market making, however, cannot be eliminated through algorithmic optimization. These costs include the funding-rate exposure of holding positions across periods, the liquidation-risk premium in extreme conditions, and the hardware and network costs of maintaining round-the-clock high-frequency trading infrastructure. As a growing number of institutions adopt AI market making, spread income is continually compressed under extreme competition, so that the reduction in variable costs cannot fully offset the contraction of the spread. According to an industry report from Ken Research, the profit margins of high-frequency trading firms have been under sustained pressure in recent years, forcing them to keep innovating technologically (this figure comes from a paid industry report and lacks an independently verifiable basis or methodological disclosure) [13]. This asymmetric change in the cost structure means that, in the AI era, institutional costs become a relatively more important constraint on profit, which makes the market-microstructure and institutional improvements discussed in Chapters 29 and 30 all the more critical.

In the CEX order-book perpetual futures market, the AI market maker's profit equation must also incorporate several distinctive cost and revenue dimensions. On the revenue side, the fee structure for makers and takers directly affects the AI market maker's net return: as a maker, an AI market maker typically enjoys a low or even negative fee (that is, it receives a rebate), whereas the taker who consumes its quotes pays a higher fee. This fee differential makes queue position a core optimization variable

in the AI market maker's optimal strategy: under price priority, orders at the same price level are filled on a time-priority basis, so millisecond-level quoting speed directly determines whether the AI market maker can occupy a favorable queue position and thereby obtain a higher fill probability and more maker-rebate income. On the cost side, the funding-rate mechanism of CEX perpetual futures constitutes a distinctive holding cost: when an AI market maker holds a net long or net short position because of inventory skew, it must pay or receive the funding rate at each funding settlement period (typically 8 hours), and this cost can significantly erode market-making profit in highly volatile markets. The funding rate is not only a passively borne cost but also a source of return that AI market makers actively pursue. In periods when the funding rate is significantly positive, an AI market maker can deliberately maintain a net short inventory to collect the rate; funding-rate prediction itself constitutes an independent signal dimension. In extreme conditions, the funding rate can spike to several hundred percent on an annualized basis, at which point funding-rate exposure may completely dominate the market maker's profit and loss, far exceeding the contribution of spread income. The AI market maker's profit equation must therefore treat the funding rate as a core optimization objective on par with spread income, rather than merely as a holding cost.

Figure 31-2 unfolds the cost side of the market-making profit equation into five cost drivers and groups them into two categories: variable costs (significantly compressed by AI) and fixed institutional costs (only modestly improved). This asymmetric change in the cost structure explains why architectural improvements at the institutional level in fact carry higher marginal value in the AI era.

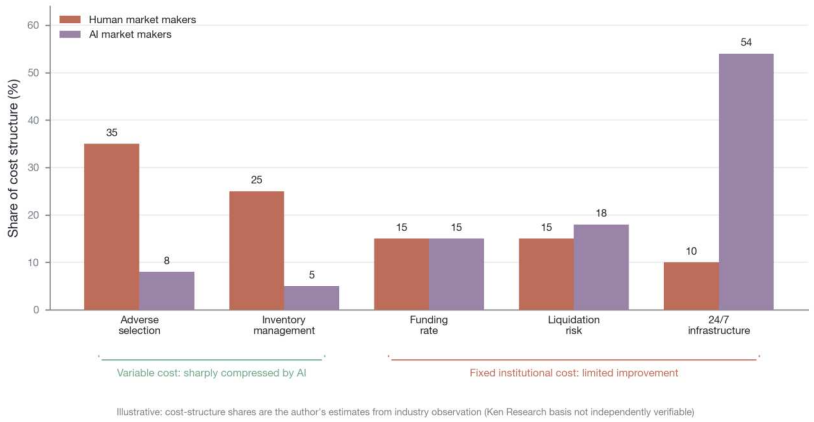


Figure 31-2. Recalibrating the AI market maker's profit equation: the cost side is unfolded into five cost drivers, grouped into variable costs (adverse selection and inventory management; significantly compressed by AI) and fixed institutional costs (funding rate, liquidation risk, and round-the-clock infrastructure; only modestly improved) (Illustrative: the cost-structure proportions are the author's estimate based on industry observation; the basis of the Ken Research industry report [13] cannot be independently verified)

31.2.2 The changing nature of adverse selection

In traditional microstructure theory, the core driver of adverse selection is information asymmetry. Informed traders possess private information about an asset's fundamen-

tals, while the market maker, as the uninformed party, can only compensate for the risk of being selectively traded against by informed traders by widening the bid-ask spread. When the main participants in the market shift from humans to AI agents, the nature of this informational advantage undergoes a fundamental change.

In an AI-dominated market, an advantage in the source of information is gradually replaced by an advantage in the speed of information processing. Although AI agents can capture, parse, and synthesize public market data, news sentiment, and on-chain indicators at speeds far exceeding humans', they cannot obtain true inside information. This means that speed-based informational advantage is neutralized by the widespread adoption of AI technology. When all market makers are equipped with algorithms that respond at the millisecond level, the latency-arbitrage space that relies purely on processing public information is greatly compressed.

A new form of adverse selection thereby begins to appear. This adverse selection is no longer based on an asymmetry in fundamental information, but on algorithmic asymmetry. When arbitrageurs are likewise AI-driven, they can reverse-engineer the market maker's quoting logic through machine learning models. These AI arbitrageurs learn the market maker's reaction latency, inventory-adjustment thresholds, and spread-widening patterns under particular order-flow shocks, and thereby predict the market maker's next move. This game shifts adverse selection from "who knows more" to "whose model generalizes better and predicts more accurately." Against this backdrop, the market maker faces not a trader who possesses inside information, but an advanced AI agent able to compute precisely the blind spots in its algorithm.

In cryptocurrency markets, however, traditional adverse selection based on the source of information still exists in abundance. Miners and validators have knowledge of block contents before the public; project teams hold undisclosed information before token unlocks and partnership announcements; and leaks of regulatory decisions often exhibit an asymmetric distribution. These advantages in the source of information cannot be replaced by algorithmic speed. Adverse selection in the AI era therefore exhibits a two-layer structure: the lower layer is the persistent asymmetry in the source of information, and the upper layer is the newly added asymmetry in algorithmic capability. What the market maker faces is a superposition of the two kinds of risk rather than a simple substitution—it must guard both against informed traders who possess inside information and against AI arbitrageurs able to reverse-engineer its algorithmic logic.

31.2.3 The dimensional explosion of inventory management

Constrained by cognitive bandwidth, human market makers can typically manage inventory only in a limited number of dimensions. A typical trading team might focus on managing the inventory risk of two or three correlated asset pairs and rely on heuristic rules for cross-market hedging. AI agents break through this dimensional limit, achieving a dimensional explosion in inventory management.

Modern AI market-making systems can simultaneously manage a high-dimensional inventory vector across dozens of exchanges and hundreds or thousands of contracts. This high-dimensional optimization capacity enables AI to use cross-asset correlations for dynamic hedging. When one asset experiences one-sided net buying that unbalances

inventory, for example, AI need not confine itself to adjusting quotes or hedging on the same asset pair; instead, it can instantly compute the optimal alternative hedging path across the entire crypto-asset matrix. This globally optimized perspective significantly improves overall market liquidity under normal conditions, because the market maker can tolerate a larger inventory deviation at any single point so long as it remains risk-neutral across the global vector.

Multidimensional cross-exchange inventory optimization faces rigid physical constraints in practice. Each exchange's margin account is isolated from the others, and the market maker must pre-fund enough margin at each venue to withstand the worst-case inventory deviation, making capital efficiency far lower than the theoretically optimal level of "global risk neutrality." Transferring funds across exchanges typically takes minutes or even tens of minutes, precluding instantaneous hedging. The 2022 collapse of FTX starkly exposed this fragility: funds frozen at FTX left multiple market makers' exposures on other exchanges unhedged, causing enormous losses. A further and more troubling effect is that AI market makers' cross-asset hedging behavior can itself amplify correlation breakdowns in periods of stress. When market maker A sells Ethereum to hedge its Bitcoin exposure, the decline in Ethereum's price triggers market maker B's Ethereum-inventory adjustment and further cross-asset hedging, forming a positive-feedback loop in which the AI hedging mechanism turns from a tool of risk diversification into a manufacturer of correlation contagion.

This multidimensional optimization, however, also introduces a new form of systemic risk. When the market is in a normal state, the cross-asset correlation matrix is relatively stable, and AI's hedging logic can operate effectively. Under an extreme market shock, however, correlations among assets often shift abruptly, and assets that were negatively correlated may suddenly become highly positively correlated. At that point, a high-dimensional inventory-optimization model built on the historical correlation matrix may fail simultaneously across all dimensions. Because AI systems operate synchronously across multiple markets, this correlation-shift-induced model collapse spreads through the entire market network at the millisecond level, turning a multidimensional hedging mechanism originally meant to diversify risk into a transmission channel that amplifies systemic fragility.

31.2.4 Adversarial market making

Chapter 19 discussed different types of toxic order flow, including the information-driven, momentum-driven, and structural-arbitrage types. As AI technology has been more deeply applied, a new dimension has appeared in the toxicity spectrum: AI-on-AI toxicity, or "adversarial market making." This phenomenon originates in the adversarial-example problem widespread in machine learning and is mapped directly onto the microstructure of financial markets.

Adversarial market making refers to AI arbitrageurs deliberately triggering an adverse response from the market maker's AI model by sending carefully crafted "adversarial orders." This is analogous to how, in computer vision, modifying just a few pixels can cause an image-recognition model to mistake a panda for a gibbon. In a trading setting, an attacker uses small-scale fake quotes, order cancellations at particular frequencies,

or minute inventory shocks to deceive the market maker's model into misjudging the market trend or the volatility regime. Hofweber et al. (2025) [14] note that, by minutely manipulating the value of individual assets, one can construct adversarial examples against financial forecasting models, and may even trigger the model to make a self-fulfilling crash prediction.

Figure 31-3 shows the operating mechanism of adversarial market making. This mechanism is highly isomorphic in principle to adversarial-example attacks in the field of computer vision, revealing the new form of security fragility that AI market-making systems introduce in their pursuit of predictive precision.

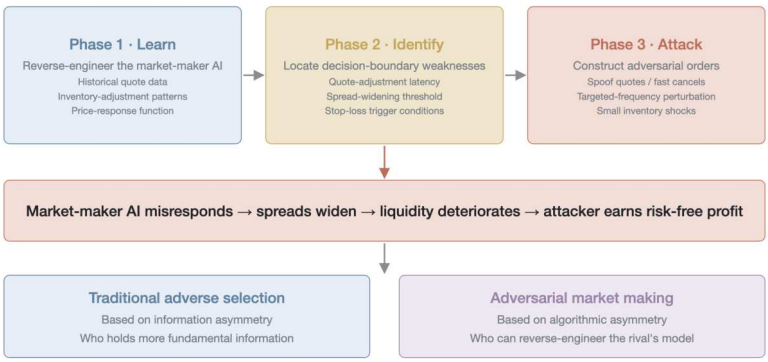


Figure 31-3. The mechanism of adversarial market making (Data source: Hofweber et al., 2025 [14])

Under this new form of attack, the safety of a market-making strategy no longer depends solely on the accuracy of its predictions or the lowness of its latency, but also on its robustness to adversarial attacks. Traditional risk-control parameters often cannot identify this kind of adversarial perturbation disguised as normal order flow. When the market maker's AI model receives these adversarial signals and makes an erroneous quote adjustment, the attacker can extract a risk-free profit. This forces market makers to make adversarial defense a core part of the architecture when developing their algorithms, thereby increasing the system's complexity and computational overhead.

31.2.5 Structural strengths and weaknesses

AI agents and human market makers exhibit sharply different behavioral characteristics in market microstructure. Understanding the structural strengths and weaknesses of these two types of actors directly informs one's judgment about the dynamics and resilience of future market liquidity.

Table 31-2 compares the core differences between AI and human market makers along six dimensions: speed and operation, optimization capability, emotional control, strategy homogenization, handling of unknown scenarios, and security fragility. The table reveals a key structural trade-off: AI's absolute advantages on the first three dimensions (millisecond-level response, high-dimensional optimization, no emotional fluctuation) are precisely the root of its severe disadvantages on the latter three (strategy

convergence, black-swan collapse, adversarial attack). In other words, the AI market maker's strengths and weaknesses do not exist independently; they are two sides of the same technological trait.

Dimension	AI market maker	Human market maker
Speed and operation	Millisecond-level response; runs around the clock without rest	Constrained by physiological reaction time and the need for sleep
Optimization capability	Can simultaneously handle high-dimensional matrices and complex correlations	Can handle only low-dimensional variables; relies on heuristic rules
Emotional control	No emotional fluctuation; rigorously executes the set strategy	Easily swayed by market panic or greed
Strategy homogenization	Extremely high risk. Because training data and optimization objectives converge, synchronized behavior arises readily	Naturally heterogeneous; decides based on differing experience and intuition
Handling of unknown scenarios	Prone to model collapse when facing black-swan events outside the training data	Can make reasonable judgments in entirely new scenarios by relying on common sense and intuition
Security fragility	Vulnerable to adversarial-example attacks and algorithmic reverse engineering	Behavior is nondeterministic, hard to predict precisely and to attack programmatically

Table 31-2. Comparison of the structural strengths and weaknesses of AI versus human market makers (Data source: compiled by the author)

AI market makers hold significant advantages in speed, multidimensional optimization, and emotional control. They can eliminate the irrational withdrawal of liquidity caused by human panic and provide extremely narrow bid-ask spreads under normal conditions. The price of these advantages, however, is severe structural weaknesses. Driven by the strategy-homogenization mechanism (whose causes are analyzed in detail in Section 31.3.2), AI strategies are highly prone to convergent reactions. When confronted with a never-before-seen black-swan event, these homogenized models may simultaneously make catastrophically wrong decisions, causing liquidity to dry up completely in an instant. In addition, AI's natural fragility to adversarial attacks introduces a new form of technological risk into the market.

31.3 AI and liquidity dynamics

In traditional market-microstructure theory, a liquidity crisis is usually portrayed as a chain reaction in which psychology and mechanism are intertwined. In the liquidity reflexivity triangle examined in Chapter 20, a price decline triggers panic, panic causes market makers to withdraw, and market makers' synchronized withdrawal further intensifies the price decline, ultimately becoming a tragedy of the liquidity commons in which liquidity dries up. The core driver of this classic framework is human cognitive

limits and emotional fluctuation. When market participants shift from human traders to AI agents, this foundational assumption is fundamentally altered. AI has no physiological or psychological mechanisms of fear or fatigue; its decisions are based entirely on preset risk-control parameters and real-time data flows. This fundamental shift changes not only the response pattern of a single market maker, but also reshapes the liquidity dynamics of the entire market. The widespread deployment of AI agents, on the one hand, mitigates the traditional risk of liquidity droughts by eliminating emotional contagion, and, on the other, introduces a new form of systemic fragility through strategy homogenization. This section reexamines the liquidity reflexivity framework, analyzes AI's double-edged effect in mitigating traditional panic while triggering new flash crashes, and explores the path and risks of the evolution from protocol-level algorithmic market making to AI adaptive market making.

31.3.1 Mitigating the tragedy of the liquidity commons

The standard response of traditional market makers to extreme market volatility is to widen the bid-ask spread or cancel quotes outright. In microstructure theory, this behavior is interpreted as a rational avoidance of adverse-selection risk, but at the macro level, the synchronized withdrawal of all market makers causes liquidity to evaporate in an instant, forming a classic tragedy of the commons. Human market makers' withdrawal often carries strong emotional overtones; when confronting anomalous volatility that exceeds historical experience, an increase in risk aversion leads to an overreaction to risk. This emotional contagion is a key catalyst of liquidity phase transitions in traditional financial markets—that is, the market's abrupt rupture from a high-liquidity state to a low-liquidity state.

The introduction of AI agents fundamentally changes this microlevel mechanism. AI acts on programmatic risk control, and its decision logic is not disturbed by emotional fluctuation. Facing violent market volatility, an AI market maker precisely adjusts quote depth and spread according to real-time computed volatility indicators, inventory positions, and adverse-selection probabilities. Because there is no "persistent risk-aversion effect after a risk event," AI does not maintain a defensive posture after the risk signal has cleared, but can rapidly restore liquidity provision as market conditions improve. The empirical study of the market crash during the COVID-19 pandemic by Raboun et al. (2021) [15] shows that in a high-uncertainty regime, the elasticity of price with respect to liquidity demand increases significantly, but the presence of algorithmic trading makes this change in elasticity smoother rather than an abrupt rupture. During the cryptocurrency market volatility of 2024, market makers equipped with AI risk-control systems provided significantly more liquidity in extreme conditions than in the same period of 2023 (an estimate based on the author's reading of public data), indicating that AI's rational decision mechanism can indeed mitigate the traditional risk of liquidity droughts to some extent.

This data-driven, rational response mechanism transforms the process of liquidity decline from a traditional "phase transition" into a smoother "gradual change." Under stress testing, the group of AI market makers does not, like human traders, undergo an irrational synchronized withdrawal driven by panic. This "no-withdrawal" prop-

erty is further reinforced when protocol-level liquidity providers are combined with AI adaptive algorithms. Hardcoded protocol rules ensure the base existence of the liquidity pool, while AI agents are responsible for dynamically optimizing the distribution of funds in extreme conditions. The HLP vault performed well in the market stress test of the second quarter of 2024, maintaining a relatively stable spread level even in extreme conditions—a marked contrast with the large spread widening among traditional market makers over the same period (an estimate based on the author's reading of public data). Maintaining a narrow spread in an extreme-volatility environment, however, may come at the cost of bearing higher adverse-selection losses; a comprehensive assessment should consider market-making profit-and-loss data together with spread indicators. From this angle, the widespread application of AI has indeed, to some extent, mitigated the tragedy of the liquidity commons driven by human emotion.

The sustainability of this mitigating effect, however, remains in question. The rational decisions of AI market makers are based on their accurate perception of market states and correct calibration of their risk models. When the market presents anomalous conditions outside the training-data distribution, or when the AI model's assumptions are broken, this effective risk assessment may lose its accuracy as the model's assumptions fail. In addition, the widespread application of AI also means that participant heterogeneity is declining, and this homogenization is itself a new form of systemic risk. Thus, although AI can mitigate liquidity crises in normal markets, its performance in extreme situations still requires ongoing monitoring and assessment. A more fundamental limitation is that the rational nature of AI market makers can mitigate only the microlevel liquidity withdrawal driven by emotional contagion; it cannot counteract the liquidity contraction driven by the macro credit cycle. When the Federal Reserve (Fed) enters a rate-hiking cycle and global dollar liquidity contracts, AI market makers' rational response, based on their risk models, is likewise to shrink their market-making scale and reduce risk exposure—this is precisely rational rather than emotional behavior, a normal expression of AI's "no-panic" property. The systemic collapse of the crypto market in 2022 amply demonstrates that the macro liquidity cycle has a far more decisive influence on market depth than differences in the behavioral characteristics of microlevel participants.

31.3.2 The fragility of strategy homogenization

Although AI does not experience psychological panic in the human sense, its highly rational decision mechanism can, under particular conditions, trigger another form of synchronized behavior: an "algorithmic synchronized defensive reaction." The root of this new fragility lies in the homogenization of AI strategies. In a fiercely competitive market-making environment, the AI agents developed by different institutions often rely on highly overlapping historical datasets, similar machine-learning architectures, and convergent optimization objectives. When the market presents a particular anomalous signal or microstructure feature, these AI models with similar underlying logic are highly likely to reach the same inference and, on that basis, make highly consistent decisions.

The causes of strategy homogenization are manifold. First, there is a severe asym-

metry in the availability of training data. Most AI market makers use historical market data from the same set of data providers, and these data contain the same market-microstructure features and price patterns. Second, the convergence of machine-learning architectures leads different institutions to choose similar neural-network topologies and optimization algorithms. In applying deep reinforcement learning to market making, most researchers adopt similar state-space definitions, reward-function designs, and policy-gradient methods. Third, the optimization objective of market making has an inherent consistency—namely, maximizing profit while controlling inventory risk. This similarity in objective functions causes different AI agents to converge to similar decisions when facing the same market conditions.

When multiple dominant AI market makers simultaneously identify a particular market anomaly as an extremely-high-risk signal, they synchronously widen spreads, skew quotes, or halt trading outright at the millisecond level. This synchronized withdrawal based on algorithmic consensus occurs far faster and more thoroughly than the panic selling of human traders. Based on a multi-agent simulation study, Gao et al. (2022) [16] note that the synchronized behavior of high-frequency trading algorithms is a core driver of flash-crash events, and that its destructive power over market liquidity is nonlinearly and positively related to the algorithms' share of total market volume. Compared with traditional panic, AI-dominated liquidity droughts exhibit sharply different time dynamics. A traditional panic-driven withdrawal usually takes minutes or even hours to complete, because human traders need time to process information, assess risk, and execute trades. An AI market maker's synchronized withdrawal, by contrast, completes within milliseconds, meaning that liquidity shifts instantaneously from an ample state to a depleted one with almost no buffer in between.

Figure 31-4 conceptually contrasts two modes of liquidity crisis: in a human-dominated market, the spread rises sharply and decays slowly, with a crisis time span that can extend to hours or even days; an AI-homogenization-induced flash crash, by contrast, reacts synchronously within milliseconds, manifesting as an extremely sharp "V-shaped" reversal in which quotes restart instantly once the input data return to normal. Yet the "brevity" of an AI flash crash is by no means the same as "harmlessness": in a highly leveraged perpetual futures market, even a liquidity vacuum of 2 to 5 seconds is enough to cause irreversible cascading liquidations—at 100× leverage, a 1% price jump means the complete loss of margin, and a V-shaped recovery is meaningless to a position that has already been liquidated.

This momentary liquidity vacuum, though extremely brief, is enough to trigger a chain of liquidations of highly leveraged positions. During the cryptocurrency market volatility of March 2024, because multiple mainstream market-making algorithms simultaneously identified the same risk signal, the spread in the Bitcoin spot and futures markets widened sharply in under 100 milliseconds, forcing the liquidation of millions of dollars in positions. Worse, such a flash-crash event can trigger a cascade effect, in which the liquidity drought in one market triggers a liquidity crisis in other correlated markets, thereby forming systemic risk.

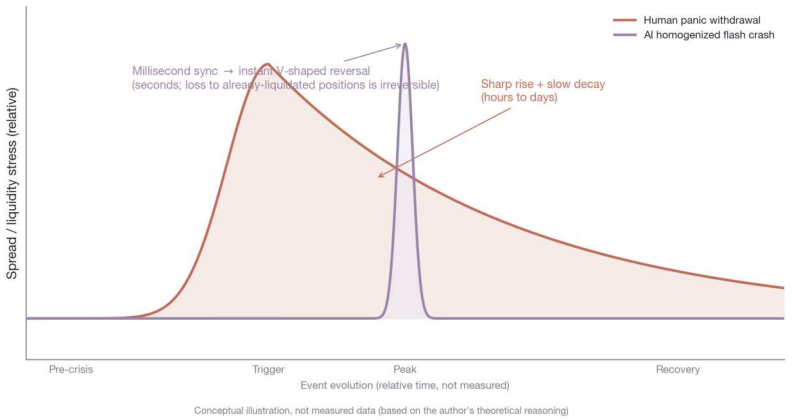


Figure 31-4. The different effects of human panic and AI homogenization on liquidity dynamics (conceptual illustration; not empirical data; based on the author's theoretical derivation)

31.3.3 Synchronized behavior and heterogeneous decision-making

As analyzed in Section 31.3.2, the homogenization of AI strategies is eroding the core source of market resilience: the heterogeneity of participants. In a human-dominated market, different traders hold different risk preferences, investment horizons, and decision logics, and this diversity ensures that two-sided liquidity exists at any given price level [17]. Brunnermeier and Pedersen (2009) [18] emphasize that market liquidity depends not only on market makers' inventory-management capacity but also on the diversity of market participants. The comprehensive penetration of AI agents, however, may weaken this heterogeneity. As machine-learning models' advantages in prediction and optimization become ever more prominent, underperforming strategies will be rapidly eliminated, and the market will gradually come to be dominated by a small number of highly optimized algorithms, eliminating the diversity of decision logic.

More dangerous still, this homogenization makes the market highly vulnerable to adversarial attacks. Goldblum et al. (2021) [19] show that machine-learning models in high-frequency trading systems are at risk of being deceived by adversarial examples. A malicious attacker can infer the decision boundaries of mainstream AI market makers by reverse engineering, and deliberately send particular combinations of order flow to trigger these algorithms' synchronized misjudgment, thereby artificially creating a liquidity vacuum and profiting from it. The cost of such an attack is relatively low, because the attacker need only understand the logic of a few mainstream algorithms to inflict a systemic shock on the entire market.

To mitigate the systemic risk of strategy homogenization, the design of market microstructure must proactively introduce an "anchor of heterogeneity." In the context of DeFi, the HLP vault can deliberately adopt pricing logic and risk models sharply different from those of mainstream high-frequency trading algorithms, so as to ensure the provision of uncorrelated liquidity even in extreme situations. For example, certain HLP-style vaults can adopt a pricing strategy based on long-term mean reversion, rather than a market-making algorithm that follows short-term market signals. Al-

though such a strategy may be less efficient than mainstream algorithms in normal markets, it can provide stable liquidity support in extreme conditions.

In addition, exchanges and protocols can restructure incentive mechanisms to implement "market-making diversity incentives." For liquidity providers that can maintain quote depth within the millisecond window in which mainstream market makers synchronously withdraw, for example, a protocol can grant an exponential fee share or token reward. This mechanism design treats heterogeneity itself as a public good with positive externalities, and uses economic means to encourage developers to deploy non-mainstream AI strategies, thereby rebuilding in the algorithmic world the ecological diversity that technological convergence has flattened. Specifically, a protocol can set a "heterogeneity score," computed according to the degree of correlation between a liquidity provider's strategy and the market's mainstream strategies. The lower the correlation, the higher the heterogeneity score and the greater the incentive received. Such a design can effectively encourage innovative market-making strategies while protecting the market from the risk of strategy homogenization. In practice, however, the heterogeneity-incentive mechanism faces severe challenges of measurement and gaming. A market maker's underlying strategy logic is private information, and an exchange or protocol can observe only order-flow behavior. Inferring strategy heterogeneity from order-flow behavior faces a severe identification problem: ostensibly different order-flow patterns may merely reflect the same strategy under different parameter settings. More importantly, the mechanism is easily gamed: a market maker can deliberately submit orders that "look different" but have zero economic meaning in order to harvest the heterogeneity reward. Moreover, in extreme conditions, even a "heterogeneous" strategy may be forced to execute the same behavior because of shared funding constraints (such as insufficient margin), at which point the heterogeneity incentive cannot prevent synchronized withdrawal. The heterogeneity incentive should therefore be a component of a diversified resilience design rather than the sole solution, and needs to work in concert with complementary measures such as circuit breakers and minimum liquidity obligations.

31.3.4 AI adaptive market making

In the evolution of decentralized exchanges, the liquidity-provision mechanism is undergoing a paradigm shift from static parameterization toward dynamic intelligence. Early automated market makers relied on a hardcoded constant-product formula; although this mechanism guaranteed the continuous existence of liquidity, its capital efficiency was extremely low. The constant-product market makers of Uniswap v1 and v2, for example, required liquidity providers to distribute funds evenly across the entire price range, leaving most funds idle at price levels far from the current market price. The concentrated-liquidity protocols and parameterized strategy engines that appeared subsequently represent the evolution of microstructure into a second stage, in which humans set the strategy framework and protocol algorithms handle automatic execution. At this stage, an HLP vault can adjust quote ranges and fee tiers according to preset rules, but its response logic is still deterministic and reactive.

The concentrated-liquidity mechanism introduced by Uniswap v3 is a typical repre-

sentative of the second stage. Liquidity providers can select a particular price range in which to concentrate their funds, thereby improving capital efficiency. This mechanism, however, still relies on human judgment about market conditions. Liquidity providers must manually adjust their price ranges, which creates two problems: first, the frequency of adjustment is limited by blockchain transaction costs, and so it cannot respond to market changes in real time; second, different liquidity providers may make suboptimal decisions, resulting in a low overall efficiency of liquidity allocation.

Academic research has begun to explore the prospects of applying reinforcement learning technology to automated market making. The simulation study by Fernández Vicente (2025) [20] demonstrates the theoretical potential of deep reinforcement learning for developing autonomous, adaptive market-making strategies. By formulating the market-making task as a multi-objective reinforcement learning problem, an AI agent in a simulated environment can simultaneously optimize spread income and inventory-risk control. Unlike a parameterized engine that relies on static rules, a reinforcement learning agent can autonomously discover complex nonlinear patterns through continuous interaction with the market and dynamically adjust its strategy weights to adapt to a continually changing market regime. Current public information is not yet sufficient to confirm whether mainstream on-chain market-making vaults have deployed deep reinforcement learning strategies in production; the research conclusions above come mainly from academic simulation rather than from empirical observation of production systems.

Viewed from a theoretical angle, if this AI adaptive market-making technology is widely deployed, it will bring a significant efficiency gain to on-chain liquidity provision. AI agents can more precisely predict changes in short-term price microstructure, and thus rapidly adjust quotes when informed traders appear, effectively reducing adverse-selection losses. Through cross-asset, multidimensional inventory optimization, an AI vault can provide deeper liquidity while holding the same risk exposure. This means that, for the same capital input, an AI market-making vault can provide greater trading depth, thereby offering users a better trading experience.

This efficiency gain, however, inevitably comes with a loss of transparency. The decision process of a reinforcement learning model has a strongly black-box character, and its internal neural-network weights and state-value function are difficult for humans to understand intuitively. When a market-making vault makes an anomalous quote adjustment in extreme conditions, participants in community governance find it hard to distinguish whether this is AI's optimal defense based on complex data, or an overfitting collapse of the model in an unseen state space. Under extreme market conditions, an adaptive market-making system may suddenly widen spreads sharply, raising users' transaction costs, and an after-the-fact analysis may show that this was the model's overreaction when facing a market state outside the training-data distribution, rather than rational risk management.

This lack of interpretability makes the "strategy transparency" problem discussed in Chapter 21 all the more prominent, and also poses a fundamentally new challenge to the on-chain auditing and governance mechanisms of DeFi. To meet this challenge, protocol designers need to find a balance point between efficiency and transparency.

One possible solution is to introduce "explainable reinforcement learning" technology—that is, to make the AI agent's decision process more transparent through special model architectures and training methods. An attention mechanism, for example, can be used to identify which market features the model attends to most when making a decision, thereby improving the interpretability of the decision. Another solution is to implement an "AI audit" mechanism—that is, to periodically conduct an after-the-fact analysis of the AI market-making vault's decisions to check whether anomalous or unreasonable behavior has occurred. Through these measures, DeFi can maintain the necessary transparency and controllability while enjoying the efficiency gains that AI brings.

31.4 AI and volatility structure

The microstructure analysis of volatility over the preceding thirty chapters rests on the assumption that participants possess human cognitive characteristics (see the chapter introduction): information processing takes time, and the clustering effect of volatility stems from humans' gradual reaction to new information. When AI agents become the market's primary liquidity providers and price discoverers, these underlying assumptions break down. AI has no emotion, can process multidimensional data at the millisecond level, and can optimize synchronously across markets. This microlevel behavioral change is very likely to reshape the macrolevel characteristics of volatility. This section reexamines the volatility-decomposition model of Chapter 22 and the volatility-clustering mechanism of Chapter 23, analyzing how AI alters the structure of normal-regime volatility and creates new forms of tail risk.

31.4.1 Revising the four-source decomposition

In classic microstructure theory, the volatility of asset returns is decomposed into four basic sources: informational volatility, liquidity volatility, mechanical volatility, and contagion volatility. Informational volatility arises from the process by which fundamental news is incorporated into prices; liquidity volatility is driven by order-flow imbalances and the market maker's inventory-risk premium; mechanical volatility comes from the rigid execution of protocol rules such as the liquidation engine; and contagion volatility is the result of cross-market arbitrage and emotional resonance. When AI agents dominate the market, the characteristics of these four traditional volatility sources shift systematically, and, moreover, an entirely new, fifth source of volatility arises.

The pulse width of informational volatility narrows significantly under AI's involvement, but the peak may in fact rise. In traditional markets, for information to move from release to full reflection in prices requires a gradual process in which informed traders build positions step by step, uninformed traders follow, and market makers adjust quotes. This window of information diffusion typically takes minutes to hours. AI agents, by parsing news feeds, social-media sentiment, and on-chain anomalies in real time via NLP models, can complete the pricing of information at the millisecond level. The high-frequency volatility-decomposition model of Dahlhaus and Neddermeyer (2014) [21] shows that an increase in information-processing speed compresses the window of price discovery, shrinking the pulse width of informational volatility

from the minute scale to the millisecond scale. This time compression, however, brings a counterintuitive result: although the process of incorporating information into prices is faster, because multiple algorithms compete to price the same information at the same instant, their quote adjustments reinforce rather than cancel one another. In an AI-dominated market, informational volatility manifests as a price jump within an extremely short time, and its instantaneous peak may in fact be higher than the gradual rise in a human market. This means that although the total time for information to be incorporated has shortened, the density of volatility per unit of time rises significantly.

The distributional shape of liquidity volatility undergoes a structural shift from mildly right-skewed toward bimodal, the most important change in volatility characteristics in the AI era. In a normal market, AI market makers can provide smoother liquidity than humans through multidimensional inventory management and high-frequency quote updates. Because AI is unaffected by emotion and does not withdraw in panic when facing routine order-flow imbalances, normal-regime liquidity volatility declines significantly. Based on a theoretical derivation from a multi-agent simulation model (author's estimate), normal-regime liquidity volatility may fall by 30% to 50% in a market dominated by AI market makers. When the market presents an anomalous signal outside the training-data distribution, however, strategy homogenization causes multiple liquidity providers to widen spreads or cancel quotes simultaneously. This instantaneous liquidity drought in extreme conditions causes liquidity volatility to spike sharply; a derivation based on the same theoretical framework (author's estimate) indicates that it may rise by 200% to 500%. This extreme asymmetry leads to a fundamental change in the distribution of liquidity volatility. In a human-dominated market, liquidity volatility usually follows a mildly right-skewed distribution, because panic is gradual and the deterioration of liquidity is continuous. In an AI-dominated market, by contrast, the distribution of liquidity volatility exhibits a clear bimodal character: one peak corresponds to the low volatility of normal conditions, and the other to the high volatility of extreme events, while the intermediate region between the two peaks shows a marked decline in probability density. The mechanism forming this bimodal distribution stems from AI's binary decision characteristic: an algorithm either maintains efficient market making under normal conditions, or withdraws completely once a risk-control threshold is triggered, with rarely an intermediate state. The numerical ranges above are all based on the author's theoretical modeling under extreme assumptions and reflect the upper and lower bounds of the magnitude of change rather than a precise forecast; the actual impact will depend on AI's penetration rate in the market, the degree to which strategy heterogeneity is preserved, and the design details of circuit breakers.

The evolutionary directions of mechanical and contagion volatility diverge. Mechanical volatility is triggered mainly by the liquidation mechanism; AI liquidation bots can more precisely predict liquidation thresholds and either dynamically hedge before the trigger or complete an optimal liquidation at the instant of the trigger, thereby modestly reducing the chained price shock of a liquidation cascade. In traditional markets, a liquidation event often causes a price shock of 5% to 15%, whereas under an AI liquidation mechanism this magnitude may fall to 2% to 5%. Contagion volatility, by contrast, rises significantly under AI's cross-market synchronized optimization. Algorithms maintain tight statistical-arbitrage relationships across multiple exchanges and

asset pools, and any localized shock in a single market is transmitted to the global network by arbitrage algorithms within milliseconds. In traditional markets, cross-market contagion usually takes seconds to minutes, whereas under AI dominance this time shrinks to the millisecond level. This means that the propagation speed of contagion volatility increases by more than 1,000-fold, making the speed and breadth of cross-market contagion far exceed those of the human-dominated era.

Beyond revising the traditional four sources, the widespread application of AI introduces an entirely new source of volatility: AI-homogenization volatility. The fundamental mechanism of this additional source—the high convergence of different institutions' algorithmic models in training data, optimization objectives, and feature engineering—has been analyzed in detail in Section 31.3.2. When a large amount of capital is driven by neural networks with similar underlying logic, the models' resonant reaction to a particular minute perturbation produces endogenous volatility independent of fundamentals and of liquidity supply and demand. For example, if all algorithms simultaneously narrow their spreads upon detecting a particular pattern of order-flow imbalance, this collective behavior causes an instantaneous change in liquidity and thereby produces price volatility. This volatility, endogenously created by the collective behavior of the algorithmic population, constitutes a core variable that microstructure theory in the AI era must incorporate.

31.4.2 AI flash crashes

The evaporation of liquidity is the core mechanism that triggers extreme market events, but the liquidity drought under AI dominance differs essentially in its microdynamics from the panic-driven withdrawal of human traders. In the five-stage model of liquidity crisis discussed in Chapter 20, a crisis usually begins with an external shock, then triggers a chain reaction through market makers' risk aversion and information asymmetry, and finally leads to a spiraling decline in liquidity. This process usually takes minutes to hours to ferment in a human market, whereas the involvement of AI algorithms compresses this evolutionary cycle to the limit.

The "AI flash crash" is a new form of extreme event triggered by strategy homogenization. As analyzed in Section 31.3.2, the major market-making algorithms and high-frequency trading agents, because of strategy convergence, often have highly overlapping judgment boundaries for anomalous signals. Liu et al. (2025) [7] note that the homogenization of trading strategies amplifies market volatility and gives rise to mini-crashes. Once a non-standard event or an adversarial order flow touches these algorithms' shared blind spot, multiple liquidity providers make the same defensive decision within the same millisecond, synchronously widening bid-ask spreads or clearing the order book outright. This instantaneous disappearance of liquidity does not stem from the contagion of human fear, but from the mechanical triggering of the algorithms' risk-control parameters. In an AI flash crash, liquidity does not disappear step by step but evaporates in an instant within milliseconds, because all algorithms make the same decision at the same time.

Compared with the famous "flash crash" of 2010, an AI flash crash exhibits a sharply different volatility signature in its time-series characteristics. According to the CFTC's

analysis of the 2010 event [22], a traditional flash crash is accompanied by the spread of panic, with volatility rising slowly before erupting sharply. Specifically, the flash crash of May 6, 2010, lasted 36 minutes, during which the S&P 500 fell approximately 9%. This process can be decomposed into three phases: first, 15 minutes of slow decline (a 2% to 3% drop), during which traders began to sense the anomaly; second, 10 minutes of accelerating decline (a 5% to 7% drop), as panic began to spread; and finally, 11 minutes of bottoming out and recovery, as liquidity gradually returned. After the price bottomed, because of traders' "fear aftershock," the market took a relatively long time to restore normal liquidity provision. The high-frequency data study by Kirilenko et al. (2017) [23] confirms that during the traditional flash crash, although high-frequency traders' market-making behavior did not change fundamentally, the lack of coordination meant that liquidity recovery was still slow.

The time scale of an AI flash crash is entirely different; a typical AI flash crash may complete the whole process within 2 to 5 seconds: the trigger phase (0–100 milliseconds), in which an adversarial order or market anomaly triggers the algorithms' risk-control systems; the synchronized-reaction phase (100–500 milliseconds), in which multiple market-making algorithms simultaneously detect the anomalous signal and collectively cancel quotes or widen spreads; the propagation phase (500 milliseconds to 2 seconds), in which the instantaneous disappearance of liquidity causes a price jump that, once detected by arbitrage algorithms, further worsens the liquidity deterioration; and the recovery phase (2–5 seconds), in which liquidity provision restarts instantly once the anomalous signal clears. This rapid recovery, lacking any "fear aftershock," means that an AI flash crash is often already complete before human traders have even noticed it.

The timeline above is a hypothetical scenario based on theoretical derivation; there is as yet no fully documented and academically verified real case of a purely AI flash crash. In actual markets, the dynamic process of an AI flash crash is modulated by multiple constraints. In a centralized-exchange environment, price-deviation-threshold protections and circuit breakers intervene early, transforming the flash crash from a continuous V-shaped path into a stepwise "halt–information accumulation–recovery" mode. In an on-chain environment, block time constitutes a physical lower bound on the speed of synchronized reaction: Ethereum's block interval of about 12 seconds makes millisecond-level algorithmic synchronization impossible on that chain, and although the 200- to 400-millisecond block time of high-performance application chains allows faster reactions, it is still far slower than the speed of centralized exchanges.

Figure 31-5 contrasts the volatility time series of the two types of extreme event: a traditional flash crash is a bell-shaped curve lasting about 36 minutes, whereas an AI flash crash is an extremely sharp pulse of very large magnitude lasting only 2 to 5 seconds (the time axes of the two series are not to the same scale). This difference in shape corresponds to the distinction between an emotion-driven, gradual collapse and an algorithmically-synchronized, abrupt collapse.

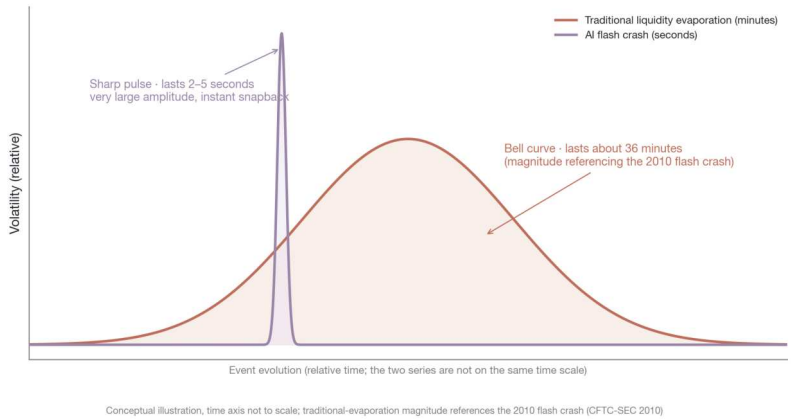


Figure 31-5. A comparison of the volatility time-series characteristics of a traditional liquidity evaporation and an AI flash crash (conceptual illustration; the time axes of the two series are not to the same scale; the magnitude of the traditional evaporation references the 2010 flash crash, data source CFTC [22]; the duration of the AI flash crash is a theoretical derivation)

31.4.3 The AI effect on volatility clustering

Volatility clustering is one of the most famous stylized facts of financial time series—that is, large movements tend to be followed by large movements, and small movements by small movements. The generalized autoregressive conditional heteroskedasticity (GARCH) model proposed by Bollerslev (1986) [24] provides a classic statistical description of this phenomenon. At the microstructure level, volatility clustering is usually explained by the clustering effect of information flow and the slow convergence of traders' heterogeneous beliefs. When AI reshapes market microstructure, the underlying feedback loops that drive volatility clustering undergo a fundamental change.

AI's adaptive capacity significantly changes the tempo and characteristics of market-state transitions. In the state-machine model of Chapter 23, the market's transition from a calm period to a high-volatility period is usually accompanied by a continuous deterioration of microstructure variables (such as the bid-ask spread and order-book depth). This transition process is gradual in a human market: the bid-ask spread widens step by step from 1 basis point to 2, 3, or 5 basis points, and order-book depth falls step by step from 1 million units to 500,000 or 100,000 units. This process may take minutes to hours. In an AI-dominated market, however, the state transition becomes extremely abrupt. In its 2024 Global Financial Stability Report, the International Monetary Fund notes that AI-driven trading can improve market efficiency and deepen liquidity but may also amplify volatility in periods of stress [25]. Because AI agents can process vast amounts of data in real time and hedge at high frequency, they exhibit extremely high efficiency in absorbing routine information shocks, so that normal-regime market volatility is continually compressed. This smoothing of normal-regime volatility superficially enhances the market's stability.

This normal-regime stability, however, does not reflect the clustering of tail risk. In a human-dominated market, the volatility-behavior feedback loop is gradual: a rise in volatility causes some market makers to exit, and the decline in liquidity further pushes up volatility. In this feedback loop, each step has a time lag, because humans need

time to perceive market changes, assess risk, and make decisions. In an AI-dominated market, this feedback loop is replaced by a "homogenization–synchronized reaction–amplification" mechanism. Abbas et al. (2024) [26] emphasize that although the latest generation of AI technology has improved risk management, it has also increased the market's systemic fragility. Specifically, when the system encounters a never-before-seen extreme shock, the highly optimized algorithms simultaneously enter risk-off mode. This nonlinear phase transition from extreme calm to extreme turbulence makes the state transition more abrupt. The traditional gradual transition of "calm → mild volatility → moderate volatility → high volatility" is replaced by a jump transition of "calm → high volatility." This jump occurs over an extremely short time, possibly only seconds.

The manifestation of volatility clustering also changes fundamentally. In a human market, volatility clustering manifests as continuous clusters of high volatility—that is, if volatility is high today, it is very likely to be high tomorrow as well. This is because the information shock or liquidity problem that caused today's high volatility may continue to affect the market tomorrow. In an AI-dominated market, however, the manifestation of volatility clustering evolves into isolated, extreme volatility spikes that erupt suddenly against a long-run baseline of low volatility. These spikes are extremely brief (on the scale of seconds) but have very high peaks. Once the spike passes, the market rapidly returns to the low-volatility baseline. This new form of volatility clustering, called "spike clustering," differs essentially from traditional "cluster clustering."

The jump transition of volatility is not, however, unique to the AI era. Hamilton's (1989) regime-switching model and Merton's (1976) jump-diffusion model long ago described the discontinuous changes in volatility in financial markets. Even before the large-scale penetration of AI, the crypto market had already experienced instantaneous jumps from low to extreme volatility, such as the global market crash of March 2020. AI homogenization's distinctive contribution lies not in creating the volatility jump itself, but in changing the endogeneity of the jump: traditional jumps are triggered mainly by exogenous information shocks, whereas AI-homogenization volatility can be endogenously produced by the collective behavior of the algorithmic population in the absence of a significant exogenous shock. How to identify this AI-endogenous jump from the exogenous jumps inherent in market structure remains an open question awaiting empirical testing.

The slow-variable mechanism still exists in the AI era, but its channel of action is redirected. Slow variables such as the macroeconomic cycle and the size of a protocol's underlying liquidity pool determine the floor of the market's carrying capacity. Although AI agents play their games at the millisecond level, the effectiveness of their strategies is still constrained by these slow variables. The difference is that AI can identify a slight decay in a slow variable earlier and adjust its strategy parameters in advance before the critical point arrives. This forward-looking collective adjustment often compresses the volatility pressure that should have been released over a longer period into an eruption within an extremely short time window. For example, if the size of a liquidity pool has fallen by 20% over the past week, this information will be captured by all algorithms at the same time. In a human market, it might take days

for this information to be fully recognized by traders, whereas in an AI market it is incorporated into all algorithms' risk assessments within milliseconds. This means that the volatility pressure that should have been gradually released over several days is released, concentrated, within an extremely short time window.

31.4.4 The effectiveness of volatility forecasting

Volatility forecasting is central to risk management, options pricing, and market-making strategy. In the traditional econometric framework, the effectiveness of volatility models (such as the GARCH family) depends on the relative stability of the market's statistical characteristics over a certain period. When the market's main participants are themselves machine-learning models with a continuous learning capacity, however, the logic of forecasting shifts from static statistical inference to a dynamic algorithmic game.

The market's statistical characteristics are in a state of continual drift amid AI's ongoing evolution. Machine-learning models in financial forecasting generally face the problem of model decay. Over time, changes in the external economic environment and the influx of new data lead to data drift and concept drift, causing the performance of models trained on historical data to decline gradually [27]. In traditional financial markets, a well-trained GARCH model may maintain relatively stable forecasting power for 6 to 12 months. In a microstructure where AI agents are widely deployed, however, this decay is further accelerated. When a forecasting model discovers a microstructure pattern (for example, that a particular order-flow imbalance foreshadows an imminent rise in volatility) and executes trades accordingly, its trading behavior itself changes the market's microdynamics. Other AI agents rapidly learn and adapt to this new pattern, causing the original statistical regularity to be erased. This means that the half-life of an effective volatility-forecasting model shrinks sharply, from the traditional 6 to 12 months to weeks or even days.

The sharp shortening of a model's half-life in the AI era demands a paradigm shift in forecasting methods. The fundamental cause of this shortening lies in the acceleration of the "forecast–arbitrage–adaptation" cycle in an AI market. In a human market, this cycle may take weeks: a trader discovers a profitable pattern and begins to trade on it, other traders gradually become aware of the pattern, and eventually the pattern is arbitraged away. In an AI market, this cycle may take only hours or even minutes. An AI model discovers a forecasting pattern and immediately begins to trade; other AI models detect this trading behavior within milliseconds and immediately begin adversarial trading; and eventually the pattern is rapidly arbitraged away.

In their discussion of machine learning in high-frequency trading, Kearns and Nevmyvaka (2013) [28] note that algorithms must handle not only independent, identically distributed data but also an adversarial market environment. In an AI-dominated market, any volatility model with predictive power quickly becomes a target for other algorithms' reverse engineering and adversarial attacks. This means that an effective volatility-forecasting model is no longer a fixed-parameter mathematical formula, but must be a meta-learning system that can monitor its own performance decay in real time, automatically trigger a retraining mechanism, and possess adversarial robustness.

The failure mechanism of volatility forecasting in an AI-dominated market can be understood as an infinitely recursive game. The market maker's AI tries to predict market volatility in order to adjust its quoting strategy. The arbitrageur's AI tries to predict the market maker's AI's behavior in order to profit from it. But the market maker's AI is also predicting the arbitrageur's AI's behavior in order to adjust its strategy in advance. This infinite recursion—"the market maker predicts the arbitrageur, the arbitrageur predicts the market maker"—turns volatility forecasting into a dynamic game rather than a static inference. In this game, there is no stable equilibrium point, and the market's statistical characteristics are in a state of continual change.

Specifically, suppose a volatility-forecasting model discovers the following regularity: when order-flow imbalance exceeds 1 million units, volatility over the next minute rises by 20%. The market maker's AI will use this regularity to adjust its risk exposure. But the arbitrageur's AI will anticipate this and conduct adversarial trading in advance when order-flow imbalance approaches 1 million units, so as to break the regularity. Once the regularity is broken, the original forecasting model fails. The market maker's AI will try to learn a new regularity, but the arbitrageur's AI will again anticipate this and again conduct adversarial trading. This "learn–adapt–break–relearn" cycle keeps accelerating, making it hard for stable statistical regularities to persist in the market for long, and so the difficulty of volatility forecasting rises sharply.

This acceleration of the dynamic game leads to the sharp shortening of the model's half-life. In a human market, the pace of this game is relatively slow, so a model may have a validity period of weeks. In an AI market, the pace of this game is extremely fast, so a model may have a validity period of only hours. This means that the effectiveness of traditional volatility-forecasting methods (such as GARCH models and machine-learning models) may decline significantly in an AI-dominated market. What must replace them is a meta-learning system that can adapt to market changes in real time and possesses strong adversarial robustness. Such a system no longer tries to forecast the absolute level of volatility, but instead tries to forecast the direction and speed of volatility changes. It must be able to detect changes in the market's statistical characteristics within milliseconds and immediately adjust its forecasting strategy.

31.5 AI and market quality

As AI evolves from a supporting tool into a core participant in and infrastructure for market microstructure, the traditional framework for assessing the efficiency of market operations faces systemic challenges. The five-dimensional market-quality model established in Chapter 25 (transaction cost, price discovery, resilience, fairness, and accessibility) operated well under the implicit assumption that "participants have human characteristics." AI agents' millisecond-level response speed, multidimensional optimization capability, and lack of emotional fluctuation, however, are reshaping the inner logic of these five dimensions. This reshaping is not a one-way improvement or deterioration; it exhibits a highly nonlinear character: AI significantly improves market quality in the vast majority of routine market states, yet creates a new form of fragility in the rare tail events.

31.5.1 Transaction cost

On the transaction-cost dimension, the widespread participation of AI market makers brings the most direct improvement: the extreme compression of the bid-ask spread. The traditional market maker's profit equation is constrained by adverse-selection cost and inventory-management cost, whereas AI agents can hedge inventory at high frequency across multiple markets and asset dimensions simultaneously, significantly reducing the risk premium of inventory buildup. This efficiency gain is ultimately translated, through fierce competition among market makers, into narrower spreads that directly benefit ordinary traders.

This surface-level cost decline, however, does not reflect the hidden cost transfer within the microstructure. The combination of high-frequency trading and AI algorithms enables the participants with the fastest execution speed and the best models to systematically capture a microstructure-level advantage. This advantage manifests as reacting in advance to, or "front-running," ordinary orders, and in effect constitutes a "speed tax" levied on slower traders. The study by Aquilina, Budish, and O'Neill (2022) quantifies the scale and frequency of this phenomenon by analyzing message data from the London Stock Exchange [5]. They find that latency-arbitrage races are extremely frequent (occurring on average about once per minute per stock) and extremely fast, with a modal race duration of only 5 to 10 microseconds (the proportion, concentration, and scale estimates above are magnitude estimates that the study extrapolates to the global equity market from FTSE 100 data on the London Stock Exchange). More critically, such races account for approximately 20% of global equity trading volume, and the top six high-frequency trading firms account for more than 80% of all race wins. Although each race is of small value (on average about half a tick), the cumulative effect is significant given the enormous volume. The study estimates that latency arbitrage levies a "tax" of approximately 0.5 basis points per year on the global equity market (approximately 0.42 basis points measured against total volume, or approximately 0.53 basis points measured against non-race volume), equivalent to a wealth transfer on the order of $5 billion per year in the global equity market.

When an ordinary investor tries to trade at the extremely narrow spread shown on the screen, AI agents can often cancel or adjust their quotes within milliseconds, causing the actual execution slippage to be far higher than expected. The economic essence of this phenomenon is that the seemingly evenly distributed spread is in fact stratified across multiple micro time scales: nanosecond-scale ultra-high-frequency traders, microsecond-scale high-frequency traders, millisecond-scale algorithmic traders, and second-scale ordinary investors. Each layer can extract a systematic spread advantage from the layer below, forming a "stratified execution structure based on latency differences." AI's net effect on transaction cost therefore exhibits an unevenly distributed character: the overall spread indicator appears to improve, but participants lacking top-tier AI infrastructure in fact bear higher hidden execution costs.

31.5.2 Price discovery

The efficiency of price discovery depends on the speed and accuracy with which the market converts new information into equilibrium prices. AI agents' breakthroughs in

NLP and multimodal data parsing enable them to digest news, earnings reports, on-chain data, and even social-media sentiment at the millisecond level. This enhanced information-processing capacity shortens the reaction time of asset prices to new information from the minute scale to the millisecond scale, greatly improving the efficiency of price discovery under normal conditions. The study by Brogaard, Hendershott, and Riordan (2014) shows that high-frequency traders, through rapid quote updates, play an important role in the price-discovery process, with their trades contributing a significant information-incorporation effect [29]. In the AI era, this effect is further amplified, as AI models can process hundreds of information sources simultaneously and adjust their assessment of an asset's value within milliseconds.

AI's excessive sensitivity to information, however, may cause the price-discovery process to be distorted by noise signals. When a large number of homogenized AI models are deployed in the market, they may react synchronously to certain noise data that do not represent a change in fundamentals. Because AI lacks humans' intuitive judgment and common-sense understanding of complex context, an adversarial attack or anomalous data input can readily trigger a chain reaction among the algorithms. A single piece of fake news, an anomalous large order, or a data-transmission error, for example, may be misinterpreted by AI models as an important information signal, causing prices to swing violently within milliseconds. This overreaction to noise not only increases short-term microlevel volatility but may also, within a particular time window, steer prices away from true fundamentals. Over the long run, AI accelerates the process of information incorporation, but in the short run the quality of price discovery may become highly unstable because of the algorithms' collective misjudgment. The core difficulty facing market participants is that, in an AI-driven market, distinguishing genuine information signals from noise becomes ever harder, because AI itself is continually learning how to make this distinction better, and this learning process itself produces new market-microstructure phenomena. In crypto markets, a distinctive attack surface facing AI price discovery is social-media narrative manipulation. When an AI model uses NLP sentiment analysis as a core input feature, malicious actors can directly manipulate the model's input space through a coordinated social-media offensive (including paid influencer posts, bulk posting by bot accounts, and cross-platform narrative construction). This attack vector is more upstream than the adversarial-order-flow manipulation discussed in Section 31.2.4, because it launches its attack at the level of the information source rather than at the level of market microstructure, and its cost is relatively low.

31.5.3 Resilience

Market resilience measures the system's ability to absorb a liquidity shock and recover quickly. In traditional markets, liquidity crises often originate in the contagion of panic among human traders; when facing unknown risk, market makers' synchronized withdrawal causes liquidity to dry up in an instant. AI agents' emotionless character significantly enhances market resilience under normal conditions: they do not stop quoting out of fear, but continue to provide liquidity in strict accordance with preset risk-control parameters and reinforcement learning strategies, thereby smoothing out small- and medium-scale liquidity shocks. This improvement based on algorithmic discipline manifests in everyday markets as more stable liquidity provision and smaller

spread fluctuations.

This resilience based on algorithmic discipline, however, can turn into severe fragility in tail events. As analyzed in detail in Section 31.3.2, the homogenization of current mainstream AI market-making strategies creates a new form of systemic risk: highly homogenized agents may synchronously and defensively withdraw in a "truly unknown" event, triggering an "AI flash crash" far faster than human speed. The 2010 flash crash offers a warning: at that time, the U.S. stock market fell approximately 9% within just 36 minutes, wiping out approximately $1 trillion in market value at one point (according to widely circulated media estimates, the event was therefore called the "trillion-dollar flash crash"; this total is an approximate media estimate, not an official figure from the body of the joint SEC/CFTC report), and although it ultimately recovered after human intervention, the whole process exposed the fragility of algorithmic trading [22]. In an AI-dominated market, a similar flash crash could occur faster and more violently, and the window for human intervention could be narrower. Although an AI market's recovery speed after a flash crash is likewise remarkable, this bimodal character of "extremely high normal-regime resilience, extremely low tail-regime resilience" fundamentally changes the distributional shape of market risk: normal-regime volatility may decline, but the probability and magnitude of extreme events may rise.

31.5.4 Fairness

Fairness is the dimension of the market-quality model most affected by AI. Information asymmetry in traditional market microstructure stems mainly from differences among participants in access to inside information or in professional research capability. As analyzed in Section 31.2.2, in an AI-dominated market the "advantage in the source of information" is gradually being replaced by an "advantage in information-processing speed and algorithmic computing power." Institutional investors with top-tier AI models and colocated servers can complete the entire process from data parsing to order execution within nanoseconds, while ordinary human traders or traditional algorithms simply cannot compete on the same time dimension. The magnitude of this speed difference is astonishing: nanosecond-scale ultra-high-frequency traders hold a 1,000-fold speed advantage over microsecond-scale high-frequency traders, while millisecond-scale retail investors lag nanosecond-scale traders by a factor of one million.

This widening of the "speed gap" leads to a significant deterioration in market fairness. AI not only forms a structural competitive advantage in execution speed, but can also reverse-engineer human traders' behavioral patterns through deep learning to carry out systematic microstructure exploitation. AI can, for example, analyze minute features of order flow to predict the splitting trajectory of a large order and build positions in advance, behavior that is in essence a form of electronic front-running. In on-chain markets, this phenomenon is especially pronounced: maximal extractable value (MEV) searcher bots monitor pending transactions in the mempool and carry out "sandwich attacks" before a user's transaction executes, thereby extracting value from ordinary users. Studies show that MEV extractors continually capture considerable value from ordinary users; according to one study's identification of approximately 9.4 million MEV events (approximately 6.33 million arbitrages and 3.02 million sandwich attacks,

spanning 2015 to 2023), together with on-chain monitoring by Flashbots and others, the annual scale of MEV extraction on Ethereum is on the order of tens of millions of dollars, reaching the order of $1 billion only cumulatively over many years (post-Merge) [30]. In this environment, ordinary investors in effect become the "uninformed traders" that supply liquidity profits to AI models. Absent a functional regulatory framework aimed at non-human participants, this structural unfairness will erode public trust in financial markets.

31.5.5 Accessibility

Two sharply opposed forces bear on AI's effect on market accessibility. On the one hand, the rise of the "AI-as-a-service" model and the spread of open-source quantitative tools seem to be driving a democratization of trading capability. Open-source platforms such as QuantConnect and Backtrader provide retail traders with free backtesting environments and real-time data interfaces, enabling even investors with no programming background to build trading strategies through a drag-and-drop interface. Ordinary investors can now, through natural-language instructions, have an AI assistant write complex trading scripts, backtest strategies, and execute them automatically. This devolution of technology lowers the programming barrier to quantitative trading, enabling more retail traders to use advanced analytical tools to participate in the market. Some platforms, such as TradeMaster, provide a complete reinforcement learning framework that allows non-professionals to deploy AI trading agents through simple parameter configuration [31].

On the other hand, the true core competitiveness that determines who wins and loses in the market—low-latency infrastructure, exclusive high-quality data sources, and frontier models with tens of millions of parameters—has development and maintenance costs that are rising exponentially. Although basic AI tools have become easy to obtain, the "frontier AI" capable of generating persistent excess returns is monopolized by a small number of leading institutions. Building a competition-grade high-frequency trading infrastructure requires investing millions of dollars in colocated servers, proprietary data sources, and top talent. The training cost of a frontier large language model has already reached the order of tens of millions of dollars, which is entirely unaffordable for most individuals and small and medium-sized institutions. AI has therefore in fact erected a new, invisible technological barrier in the market: it makes low-end tools easy to obtain, while raising the threshold of high-end competition to a height that ordinary participants cannot reach. The ultimate net effect on market accessibility will depend on the game between the pace of evolution of the open-source AI community and the degree of monopoly of proprietary models. If the open-source community can continually produce models approaching the frontier level, then "AI democratization" may become a reality; but if the advantage of proprietary models and infrastructure keeps widening, then AI will further intensify the divide between the haves and have-nots among market participants.

Table 31-3 summarizes the systematic revisions AI agents make to the five-dimensional market-quality model of Chapter 25. The core finding of the table is that AI's effect on market quality exhibits a highly nonlinear, dimension-by-dimension character: transac-

tion cost and price discovery improve significantly under normal conditions, resilience exhibits a bimodal distribution of "extremely high in the normal regime, extremely low in the tail," and fairness deteriorates significantly because of the widening speed gap. Notably, the net-effect assessment column shows that no single dimension exhibits a pure improvement or deterioration; behind every improvement lies a new form of risk or a cost transfer.

Dimension	AI's positive effect	AI's negative effect	Net-effect assessment
Transaction cost	Intensified competition compresses the apparent bid-ask spread to the extreme, down to the nanosecond level	The hidden "speed tax" rises, the risk of orders being front-run increases, and the annual cost of latency arbitrage is about $5 billion	Net positive, but the benefits are extremely unevenly distributed across participants of different speeds
Price discovery	The leap in information-processing capacity accelerates the incorporation of new information into prices to the millisecond level	Algorithmic homogenization readily overreacts to noise data, amplifying false signals; short-term price-discovery quality is uncertain	Net positive over the long run, but the quality of short-term microlevel price discovery is impaired
Resilience	The emotionless character eliminates panic contagion; the ability to absorb shocks under normal conditions improves	Strategy convergence creates systemic blind spots and readily triggers an instantaneous "AI flash crash"; the 2010 flash crash is a warning	Exhibits a bimodal character: normal-regime resilience improves significantly, while tail-regime resilience drops sharply
Fairness	—	The nanosecond-versus-millisecond speed gap cannot be bridged; annual MEV extraction is on the order of tens of millions of dollars (reaching the order of $1 billion cumulatively over many years)	Deteriorates significantly; ordinary traders face systematic microstructure exploitation
Accessibility	Open-source tools and "AI-as-a-service" lower the barrier to basic quantitative programming	The capital barrier for frontier models and low-latency infrastructure rises exponentially to tens of millions of dollars	Depends on the degree of technological democratization; the high-end competitive barrier hardens substantively

Table 31-3. The matrix of AI's effect on the five-dimensional market-quality model of Chapter 25 (Data source: [5][29][22][30][31])

31.6 The governance challenges of AI agents

The analysis in the preceding sections reveals AI agents' profound reshaping of microstructure characteristics as market participants, but this reshaping brings systemic challenges at the level of regulation and governance. The traditional framework for regulating financial markets is built on the assumption that participants are humans or legal entities controlled by humans. When the market's dominant force shifts to algorithmic programs that can learn autonomously, adapt dynamically, and lack a physical entity, the on-chain governance and compliance frameworks established in Chapters 26 and 30 must be correspondingly extended and revised. This section explores the governance challenges of AI agents in on-chain financial systems along four dimensions: the object of regulation, the attribution of liability, compliance mechanisms, and DAO governance.

31.6.1 Regulating non-human participants

The traditional paradigm of financial regulation takes natural persons or legal entities as its core object of regulation, with the logical starting point that the subject has the capacity and the incentive to bear legal liability. Within this framework, regulators change market participants' cost-benefit expectations by imposing fines, revoking licenses, or restricting market access, thereby curbing violations such as market manipulation and insider trading. An autonomously operating on-chain AI agent, however, breaks this basic premise. As a string of code or a model deployed on a distributed ledger, an AI agent itself has neither a physical form nor legal-entity status in any jurisdiction. Under the current legal system, an AI agent is neither a natural person able to enjoy rights and bear obligations, nor a legal person that has acquired an independent personality through a registration process, still less does it possess the legal status required of a fiduciary; its "legal status" is in a state of complete blankness.

When an AI agent can independently generate trading signals, autonomously manage multidimensional inventory, and execute high-frequency market-making strategies, its behavioral characteristics are already indistinguishable from those of a human trader, yet traditional regulatory tools can scarcely impose a substantive constraint on it. A fine cannot deter an algorithm that has no property ownership, and a market ban can scarcely stop open-source code from being redeployed by an anonymous entity. This absence of subject identity puts the identity-based regulatory model at risk of failing when confronting AI agents. The U.S. Securities and Exchange Commission (SEC) already exhibited this limitation in its early attempts to address algorithmic trading. In its 2014 enforcement action against Athena Capital Research, for example, the SEC found that the firm used an algorithm code-named "Gravy" to manipulate closing prices, and could ultimately only penalize the legal entity that owned the algorithm [32]. When an algorithm evolves into a fully autonomous AI agent deployed on a decentralized network, however, finding such a centralized entity to penalize will become extremely difficult.

To meet this challenge, the principle of functional regulation proposed in Chapter 26 is especially applicable in the AI context. The core idea of functional regulation is "same activity, same rules"—that is, the trigger for regulation should not depend on

a participant's identity attributes, but on the function it performs in the market and the risk it may generate. If an AI agent performs a market-making function on-chain, provides liquidity, and collects a bid-ask spread, then it should be subject to the same market-microstructure rules as a human market maker, such as a minimum tick size, limits on cancellation frequency, or a liquidity-provision obligation in extreme conditions. In its 2026 draft supervisory briefing on the regulation of algorithmic trading under the Markets in Financial Instruments Directive II (MiFID II), the European Securities and Markets Authority (ESMA) proposed that algorithmic trading strategies must be distinguishable, testable, and identifiable, regardless of whether AI technology is integrated behind them [33]. This paradigm shift, which moves the regulatory focus from "who is trading" to "the trading behavior itself," provides a feasible theoretical basis for bringing non-human participants into the governance framework.

31.6.2 Liability attribution and compliance credentials

Although functional regulation solves the problem of "what to regulate," it does not answer the core accountability question of "who bears the consequences." When an AI agent triggers an "AI flash crash" in the market or executes an "adversarial market making" strategy judged to be market manipulation, the attribution of liability becomes highly complex. The lifecycle of an AI system involves multiple stakeholders, forming a liability chain that is hard to delineate clearly. This chain typically comprises three core layers: the developer who designs the model architecture and trains the base model; the deployer who combines the model with a particular trading strategy and configures its parameters; and the user who ultimately derives economic benefit from the agent's trading activity.

In the autonomous AI systems of Stage 3 and Stage 4, because the agent has the capacity to evolve continuously through reinforcement learning in its market interactions, the strategy it ultimately executes may far exceed the developer's initial design expectations. Kurshan et al. (2025) [34] note that generative and agentic AI systems exhibit emergent behavior through continuous learning and the potential exchange of signals, which violates the assumption in traditional model-risk frameworks that algorithms are static and well-defined. In this situation, requiring the developer to bear full liability for violations that the agent autonomously evolved after deployment is not only logically flawed but would also severely suppress technological innovation. At the same time, because of the permissionless nature of the blockchain, deployers and users can often remain anonymous, making it extremely difficult to trace the actual controller. Autonomy changes the way behavior occurs, but it cannot eliminate the associated obligations of liability; liability must ultimately return to the human actors who design, deploy, authorize, or benefit from the AI system.

To resolve this liability vacuum, an "AI compliance credential" mechanism is being explored at the level of market infrastructure. This mechanism requires an AI agent to hold a cryptographic credential issued by a trusted third party or a decentralized verification network before it can access a particular protocol or liquidity pool. The specific design of this credential contains verification information along multiple dimensions: whether the agent's decision model has passed stress testing against adversarial

examples, whether it possesses a hardcoded risk circuit breaker, and whether its deployer's identity has passed some form of decentralized identity verification. Through cryptographic techniques such as zero-knowledge proofs, these credentials can prove to the protocol that the agent's design meets particular security and compliance standards without revealing the specific details of the agent's strategy. More importantly, the compliance-credential mechanism must include a revocation function. When the on-chain monitoring system detects anomalous trading behavior by the agent or a breach of compliance boundaries, the protocol can automatically revoke its credential, instantly cutting off its interaction rights with the liquidity pool. By enforcing verification of the compliance credential at the smart-contract layer, the protocol establishes a liability chain between the agent's behavior and a traceable subject, thereby restoring the effectiveness of the accountability mechanism in an anonymous network.

The design of AI compliance credentials must also confront the reality that compliance requirements across jurisdictions may conflict fundamentally. Some regulators require full auditability of algorithmic trading systems, which directly contradicts the design goal of zero-knowledge proofs to protect strategy details. Annex III of the EU AI Act (Regulation (EU) 2024/1689, effective 2024) limits high-risk AI in the financial domain to credit scoring of natural persons and the risk assessment and pricing of life and health insurance, and does not explicitly list high-frequency or algorithmic trading AI as a high-risk system; the latter is regulated separately, mainly by MiFID II and its delegated regulations. To extend the Act's transparency and human-oversight requirements by analogy to high-frequency trading AI would be an inference about regulatory trends rather than a current legal classification, and a decentralized, autonomous agent is architecturally hard-pressed to meet such requirements. A compliance-credential mechanism that claims universal applicability must find a greatest common divisor among the different regulatory frameworks, or provide differentiated credential tiers for different jurisdictions.

31.6.3 Native compliance and formal verification

Chapter 30 explored how embedded compliance turns regulatory rules into automatically executed code logic through smart contracts. In a market with the broad participation of AI agents, this concept can be extended further, so that compliance requirements are no longer an external constraint imposed on the agent after it decides, but are internalized as a prior property of its decision model. This internalization is usually accomplished through three levels of engineering implementation: hardcoded rules, parameterized constraints, and behavioral monitoring.

Hardcoded rules are the most basic level of embedded compliance; they directly remove options that violate a regulatory floor from the agent's action space. A smart contract can, for example, mandate that the agent's leverage not exceed a particular threshold, or that the position concentration in a single asset remain within a safe range. Introducing a compliance-penalty term into a reinforcement learning reward function is a typical application of parameterized constraints. When the agent, during the exploration phase, attempts to execute a sequence of cancellations that could lead to market manipulation, the reward function gives extremely high negative feedback,

thereby suppressing the generation of such strategies during the model-training phase. Behavioral monitoring, in turn, introduces external regulatory states in real time via an on-chain oracle, enabling the agent to dynamically adjust its risk-control parameters according to changes in market volatility or liquidity depth.

This "native compliance" design philosophy means that, for an AI agent, complying with the rules is no longer a choice based on cost-benefit analysis, but an inherent property of its operating mechanism. To ensure the reliability of this compliance property in a decentralized environment, formal verification technology becomes an indispensable tool. Formal verification, by means of mathematical proof, ensures that a smart contract's logic and an agent's decision boundaries strictly conform to expected compliance properties [35]. A verification tool can, for example, exhaustively examine the agent's action combinations across all possible market states and prove mathematically that it will never execute a self-trade that violates the "anti-wash-trading" rule. When the market environment changes or regulatory rules are updated, the agent must pass new formal verification and obtain an updated compliance credential before it can continue trading. This architecture, which deeply fuses compliance logic with trading logic, not only improves the real-time nature of regulation but also fundamentally reduces the systemic risk that an agent's autonomous evolution deviates from compliance boundaries.

31.6.4 AI and DAO governance

AI agents' penetration into market microstructure is not confined to the trading process; their potential role in the governance of decentralized autonomous organizations gives rise to more profound systemic risks. The core decision mechanism of a DAO usually relies on a token-based voting system, in which participants stake governance tokens to decide on protocol-parameter adjustments, treasury-fund allocation, or code upgrades. From the standpoint of technical implementation, so long as an AI agent can control a wallet address that holds governance tokens, it is fully capable of participating in governance. Such participation usually takes three forms: directly calling the smart-contract interface to vote, automatically delegating voting power to a particular representative through code logic, or serving as a strategy advisor that provides data-analysis-based voting recommendations to human voters.

AI agents' direct participation in governance, however, brings a new form of manipulation risk. Compared with human participants, an agent can monitor on-chain liquidity with extremely high efficiency, analyze a proposal's impact on different stakeholders, and execute a complex token-accumulation strategy in an instant. Han et al. (2025) [36] show that under a token-voting mechanism, a "whale" holding a large amount of tokens can accumulate enough voting power to manipulate governance outcomes, an attack often aimed at short-term gains. When such an attack is launched by an AI agent, its mechanism becomes more covert and more efficient. An agent can borrow an enormous sum without collateral via a flash loan and use cross-chain arbitrage or algorithmic market making to accumulate a large amount of governance tokens within an extremely short time. The flash-loan governance attack suffered by Beanstalk Protocol in April 2022 is a paradigmatic empirical case of this risk: the protocol treasury was

drained of approximately $182 million (the protocol's total loss; the attacker's actual net profit was approximately $76 million to $80 million). The attacker had submitted a malicious governance proposal the previous day, then, within a single transaction, borrowed over $1 billion in flash loans, deposited it to acquire more than two-thirds of the voting power, voted to pass and execute the proposal, drained the treasury, and repaid the flash loans. Subsequently, multiple protocols deployed defensive measures such as time-weighted voting power and snapshot voting, but the effectiveness of these mechanisms against AI-augmented governance attacks remains to be tested, because an AI agent can more precisely coordinate the timing of an attack and distribute voting across addresses to evade defenses. Subsequently, at the instant a key proposal's time lock is about to expire or a voting window is about to close, the agent can launch a surprise vote, forcibly passing a proposal that modifies the protocol's liquidation parameters or lowers its own trading fees, thereby creating an asymmetric advantage for its trading strategy.

A governance attack launched by an AI agent is also harder to detect and defend against than human behavior. An agent can perfectly coordinate thousands of independent addresses to vote in a distributed manner, easily bypassing defense mechanisms such as quadratic voting that are meant to guard against single-point manipulation. Tamai and Kasahara (2024) [37] note that quadratic voting systems lack sufficient resistance to collusion by malicious participants. AI agents are naturally capable of carrying out such a "Sybil attack," rendering traditional anti-collusion mechanisms utterly ineffective. Whether to restrict AI agents' participation in governance at the protocol level, or how to delineate the boundaries of an "AI governance participation right," has therefore become a key open question in current on-chain governance design. One feasible line of thought is to divide permissions by decision type: allowing AI agents to participate in pure parameter optimization (such as dynamically adjusting lending rates), but strictly prohibiting their participation in constitutional decisions involving the protocol's core values, large-scale treasury allocation, or underlying code upgrades. Some protocols are also beginning to explore hybrid governance models combining reputation systems or biometric verification, so as to ensure that ultimate governance control always remains in the hands of verified human participants, thereby hedging against the fragility that purely capital-based token voting faces in the AI era.

31.7 From market participant to market infrastructure

The preceding subsections analyzed the impact of AI as a market participant, including high-frequency trading, algorithmic market making, MEV searching, and arbitrage execution. In these settings, AI exhibits externality and rivalry, and its core objective is to extract value from the market. Another important evolution, however, is underway: AI is shifting from a player that trades on the market into the infrastructure that constitutes the market itself. These two roles are entirely different in nature—the former derives profit from value extraction in a zero-sum or negative-sum game, while the latter derives value from the improvement of system efficiency in a positive-sum game. This evolution changes not only AI's positioning in financial markets, but also, more profoundly, reshapes the economic logic of the entire market microstructure. This

section proceeds from four key questions to systematically analyze the mechanism, drivers, and long-term impact of this evolution.

Figure 31-6 presents AI's dual role in market microstructure and its evolutionary path. On the participant side, AI extracts market value on the strength of its speed and algorithmic advantages, but its profit margin is continually compressed under competitive pressure (Section 31.7.1 analyzes this oligopolistic mechanism in detail); on the infrastructure side, AI creates positive externalities by providing systemic services such as dynamic risk management, intelligent matching, and semantic prediction, sustaining a persistent capacity for value capture on the strength of network effects and economies of scale. This evolution from a zero-sum game to a positive-sum game is the main thread of the second half of this chapter.

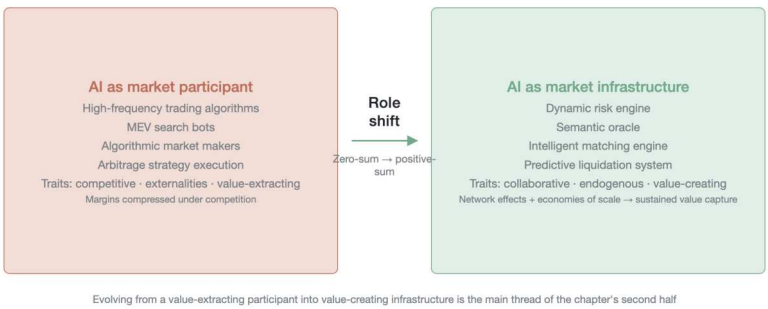


Figure 31-6. AI's dual role and evolutionary path in market microstructure (Data source: compiled by the author)

31.7.1 The profit of participant AI trends toward zero

The core characteristic of participant AI is that it uses information-processing speed and multidimensional optimization capability to capture excess returns in the market. When there are only a few AI agents in the market, they can systematically beat human traders through their speed advantage and algorithmic advantage, thereby obtaining relatively high excess returns. This profit structure, however, is unsustainable over the long run, and the fundamental reason lies in the economic laws of a perfectly competitive market.

Classical economic theory holds that, under perfect competition, a firm's long-run economic profit trends toward zero [38]. The actual structure of the cryptocurrency market-making market, however, is not perfectly competitive but exhibits marked oligopolistic characteristics. A small number of leading market makers occupy the overwhelming majority of market share, colocation infrastructure and proprietary data sources constitute substantive barriers to entry, and the market-maker incentive programs between exchanges and a few institutions are in essence exclusive arrangements. Under an oligopolistic structure, excess profit does not necessarily trend toward zero, although competitive pressure does continually compress the profit space. This judgment corroborates the discussion of algorithmic tacit collusion in Section 31.1.4: the existence of collusive behavior itself indicates that the market is not perfectly competitive. Thus,

the overall judgment that participant AI's profit faces downward pressure still holds, but its mechanism of action is the compression of profit margins under an oligopolistic structure, rather than the convergence of profit toward zero under perfect competition. Even so, AI agents are highly replicable; once a particular high-frequency trading or market-making strategy is shown to be profitable, capital and technical talent flow rapidly into the domain. Although this dynamic is not enough to eliminate the entry barriers of the oligopoly, it does intensify competition among existing participants. As ever more AI market makers and arbitrageurs enter the market, the competition among them evolves into a technological arms race at the millisecond or even microsecond level.

In this extremely competitive environment, arbitrage opportunities are rapidly discovered and eliminated, and market makers' bid-ask spreads are compressed to the limit. When AI agents evolve autonomously in a simulated market through reinforcement learning, even without any preset agreement or communication, they may spontaneously form collusive behavior to sustain supra-competitive profits [6]. In a real multi-agent competitive environment, however, this fragile tacit understanding is often broken by new entrants seeking to maximize short-term gains. As competition intensifies, the directional-trading alpha based on informational advantage and execution speed is continually eroded, and the marginal profit of value extraction inevitably trends toward zero. This process usually takes three to five years, evolving gradually from an initial period of high profit into a period of profit compression.

The deeper mechanism of this trend warrants further analysis. Participant AI's profit fundamentally comes from the market's information asymmetry and from differences in execution efficiency. When a large number of AI agents all possess similar data sources, similar computing power, and similar optimization objectives, however, these advantages rapidly dissipate. Every new competitor that enters further lowers the market's average profit margin. Eventually, the market evolves into a highly homogenized competitive landscape in which no participant can sustain a lasting excess return. This phenomenon is especially pronounced in crypto markets, because the replicability of code and the open-source culture allow a successful strategy to be copied hundreds of times within weeks. In sharp contrast, infrastructure AI does not participate directly in the market game but instead charges fees by providing systemic services; its profit comes from the real value the service creates, rather than from an information-asymmetry advantage. The stability and sustainability of this business model are far higher than those of participant AI.

31.7.2 The value space of infrastructure AI

As participant AI sees its profit trend toward zero in a highly competitive market, infrastructure AI displays an entirely different business logic and capacity for value capture. The market's demand for more robust risk control, more intelligent liquidation mechanisms, and more accurate oracle price feeds does not disappear because competition among trading participants intensifies. On the contrary, the rise in trading frequency and the increase in strategy complexity further amplify the dependence on high-quality market infrastructure. The value of infrastructure AI can be attributed to four dimensions of economic characteristics. Network effects are the core competitive barrier of

infrastructure AI. When a dynamic risk engine or an intelligent matching system attracts more users, it can collect richer market-microstructure data. These data are in turn used to train and fine-tune the model, making the AI system's predictions more accurate and its responses more agile. The improvement in performance then attracts more liquidity and trading volume, forming a positive-feedback loop. The strength of this network effect is far greater than that of traditional financial infrastructure, because an AI model can learn from every single trade and optimize its decision logic in real time.

Economies of scale are especially prominent at the infrastructure level. Developing and deploying advanced AI models requires an enormous fixed cost, including computing-power investment and algorithm research and development. Once deployment is complete, however, the marginal cost of serving an additional user is extremely low. A single risk engine can serve hundreds of exchanges or lending protocols simultaneously without a corresponding increase in cost. This high degree of scalability enables infrastructure AI to achieve extremely high profit margins.

Switching costs further consolidate the value space of infrastructure AI. Once an exchange or DeFi protocol deeply integrates an AI risk-control module or a semantic oracle into its core architecture, replacing these components entails enormous technical risk and business-disruption cost. This deep system coupling enables infrastructure providers to enjoy extremely high customer stickiness. Users find it hard to replace these services not only economically but also technically, facing enormous obstacles.

The data-feedback effect, in turn, is the source of momentum for infrastructure AI's continuous evolution. As algorithmic trading has come to dominate capital markets, its potential to disrupt market efficiency has drawn widespread concern [39]. By processing vast amounts of trading data in real time, infrastructure AI can not only optimize its own model parameters but also identify and guard against the systemic risk triggered by participant AI's homogenized strategies. This defensive capacity is itself a valuable public good, making the entire market more stable and efficient. In the process of providing this public-good-like service, infrastructure AI can capture sustained and enormous economic value.

31.7.3 Comparing operating environments

The operating environment of AI infrastructure directly determines its functional boundaries and evolutionary potential. The current crypto-finance market presents a setting in which on-chain protocols and centralized exchanges coexist, and these two environments each have their own strengths and weaknesses in supporting AI agents. Table 31-4 compares the characteristic differences between the on-chain market and centralized exchanges as AI operating environments along seven dimensions: data transparency, execution control, system composability, deployment barrier to entry, order-execution speed, liquidity depth, and trading-friction cost. The table presents a complementary picture of the two environments: the on-chain market has structural advantages in transparency, composability, and permissionlessness, while centralized exchanges maintain a significant lead in execution speed, liquidity depth, and transaction cost. This comparison provides an analytical framework for understanding the

optimal deployment environment for AI infrastructure.

Assessment dimension	On-chain market	Centralized exchange
Data transparency	Fully transparent; all trades and state history are verifiable	Internal ledger opaque; only limited interface data provided
Execution control	Smart contracts execute directly, with no need to trust an intermediary	Relies on a centralized matching engine; counterparty risk exists
System composability	Extremely high; AI can combine complex strategies across protocols	Relatively low; constrained by the application programming interfaces (APIs) the exchange provides
Deployment barrier to entry	Permissionless; AI can autonomously deploy and iteratively upgrade	Permissioned; subject to the exchange's account and compliance policies
Order-execution speed	General-purpose L1/L2 are severely constrained by block time (Ethereum about 12 seconds), but dedicated application-chain order-book exchanges (such as Hyperliquid, about 200–400 milliseconds) already approach the level of centralized exchanges	Extremely fast; supports high-frequency trading and microsecond-level response
Liquidity depth	Fragmented and relatively shallow; readily produces price impact	Concentrated and deep; suitable for the smooth execution of large orders
Trading-friction cost	Gas fees are high and volatile, limiting economic viability	Fees are low and fixed, favoring the execution of high-frequency strategies

Table 31-4. A comparison of the characteristics of the on-chain market and centralized exchanges as AI operating environments (Data source: compiled by the author)

The on-chain market provides a highly suitable operating environment for AI infrastructure. The blockchain's global-state-machine nature guarantees the full transparency and immutability of data, providing AI models with the highest-quality training data and eliminating the data falsification and hidden information common in centralized environments. Through smart contracts, an AI agent can directly control funds and execute logic without relying on a traditional trusted intermediary. This deterministic execution environment enables dynamic risk engines and predictive liquidation systems to run in a code-is-law manner by force, significantly reducing counterparty risk at the level of contract execution.

The on-chain market's extremely high composability allows an AI system to span different liquidity pools and lending protocols to conduct global asset optimization and risk hedging. This cross-protocol optimization capability does not exist at all in traditional financial markets. More critically, the permissionless nature enables AI agents to deploy, test, and upgrade autonomously without restriction, greatly accelerating the iteration cycle of innovation. A new AI strategy can be deployed to the blockchain within minutes, without waiting for the approval of any centralized institution. The current

leading on-chain order-book exchanges, however, all rely on a centralized sequencer to achieve low-latency matching. The sequencer holds the power to order transactions and, in theory, has the ability to front-run and to censor selectively, which is essentially isomorphic to the core risk of a centralized exchange. The statement above that "smart contracts execute directly, with no need to trust an intermediary" must be qualified in the face of the reality of sequencer centralization. A fundamental trade-off exists between speed and decentralization, and one should not assume that this contradiction can be simply dissolved by technological progress.

Centralized exchanges, however, maintain a significant advantage in physical performance and market depth. A centralized matching engine can provide microsecond-level order-execution speed, which directly determines the competitiveness of high-frequency trading AI. A concentrated liquidity pool means that AI faces smaller price impact when making large-scale position adjustments. Low and predictable transaction costs also make complex, frequent-interaction strategies economically viable.

Over the long run, as zero-knowledge proof technology and high-performance compute chains mature, the on-chain market's disadvantages in speed and cost are expected to be gradually eliminated. Its structural advantages in transparency, composability, and permissionlessness will make the on-chain environment the primary operating environment for AI infrastructure. This is not merely a technical question but also an economic one: the on-chain market provides a truly open, transparent, and trustworthy infrastructure, which is precisely the highest-quality operating environment that AI systems need. From an evolutionary standpoint, centralized exchanges may gradually evolve into a liquidity-aggregation layer for the on-chain market, rather than an independent trading venue. This evolution will further consolidate the on-chain market's status as the preferred environment for AI infrastructure.

31.7.4 Value-capture mechanisms

As AI evolves from participant to infrastructure in market microstructure, its value-capture mechanism also faces a paradigm shift. In traditional financial markets, infrastructure service providers usually adopt a direct service-fee model, charging a fixed proportional fee according to the number of API calls, the volume of data processed, or the scale of assets managed. The advantages of this model are stable cash flow, clear business logic, and ease of understanding and valuation by traditional investors. The AI risk-control systems and intelligent matching engines deployed within centralized exchanges mostly follow this traditional business model, treating it as part of the platform's overall competitiveness and charging users implicitly through trading fees.

In DeFi and open networks, the protocol-token model provides an entirely new path for value capture by AI infrastructure. In this model, the value of the AI system is mapped directly onto the native token it issues. The token serves not only as a governance tool, allowing the community to participate in deciding the key parameters and upgrade direction of the AI model, but also, more importantly, as a vehicle for value accrual. When a protocol generates revenue through AI services, that revenue can be internalized by buying back and burning tokens or distributing it directly to token stakers. This mechanism enables token holders to directly share in the economic

value the AI system creates.

The protocol-token model can effectively coordinate the interests of developers, liquidity providers, and users. Through token incentives, it can bootstrap the cold start of an early network and accelerate the formation of network effects. This incentive mechanism does not exist in the traditional service-fee model. The token can also serve as a flexible incentive tool to reward participants who contribute to the network, such as market makers who provide liquidity or traders who contribute data.

The two models face different trade-offs in practical application. Although the service-fee model is stable, in the highly involuted crypto market it readily faces a price war, causing profit margins to be continually compressed. Competitors can lower fees to grab market share, ultimately driving down the profit margin of the entire industry. This phenomenon has been amply demonstrated in traditional cloud computing and database services, where many once-high-profit businesses eventually evolved into low-profit commoditized services. Although the protocol-token model has strong explosive power and network-expansion capability, it is also readily subject to the negative influence of token-price volatility, falling into a speculation-driven negative-feedback loop. When the token price falls, the appeal of the incentive mechanism drops sharply, potentially causing a reversal of network effects. Many early token-incentive projects have gone through such a decline cycle.

The application of AI in financial markets brings the potential for productivity gains and cost savings, while also posing challenges to existing regulatory frameworks [40]. The future evolutionary trend may be a fusion of the two. Basic API calls and data queries adopt a low service fee to expand market share, while advanced predictive models, customized risk-hedging strategies, and protocol governance rights capture higher-order value through token economics. This hybrid architecture can not only guarantee sustained operating cash flow for the infrastructure but also, through the token mechanism, capture the long-term value that the AI system creates in the process of reshaping market microstructure. This model has already begun to be practiced in some leading DeFi protocols and has shown good sustainability.

From a deeper economic perspective, the value space of infrastructure AI is enormous also because it can create and sustain informational symmetry in the market. As stated at the beginning of this section, participant AI's profit comes from value extraction in a zero-sum game, whereas infrastructure AI's value comes from efficiency improvement in a positive-sum game; the latter is far more sustainable than the former, because it does not dry up as competition intensifies. On the contrary, the more market participants there are and the more frequent the trading, the greater the value of the infrastructure. This forms a virtuous cycle that enables infrastructure AI to sustain its capacity for value capture over the long run.

31.8 The six core AI-driven systems

AI's role in financial markets is undergoing a profound transformation, from an external trader playing the market game into the intrinsic infrastructure that constitutes the market itself. In Chapters 28 through 30, we built a trust stack based on static rules

and hardcoded parameters, an architecture that achieved decentralized trading through the combination of oracles, matching engines, and liquidation mechanisms. This static architecture, however, often proves rigid when facing extreme market volatility and complex counterparty risk, struggling to strike the best balance between security and capital efficiency. When AI agents become the core components of this infrastructure, the static architecture will be upgraded into an adaptive system with real-time perception, dynamic adjustment, and predictive capability.

Figure 31-7 shows the correspondence between the six core AI-driven systems and the Chapter 28 trust stack; overall, it achieves an upgrade from passive response to proactive optimization (the item-by-item mapping appears in Table 31-5).

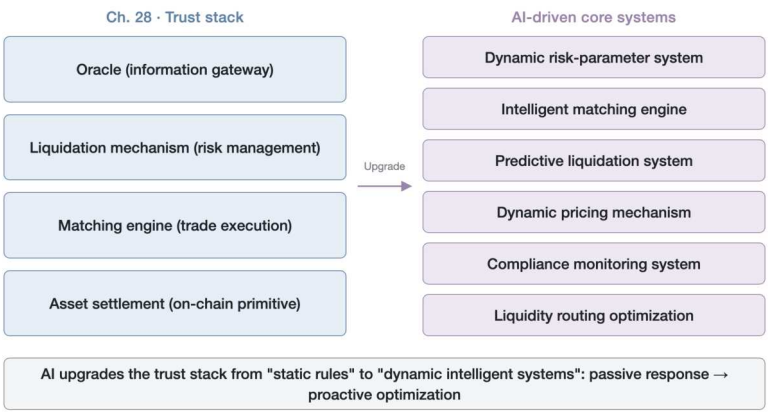


Figure 31-7. The mapping of the six core AI-driven systems to the Chapter 28 trust stack (conceptual illustration; based on the author's theoretical derivation)

Table 31-5 summarizes the functional matrix of these six AI systems. The table maps each system to a specific layer of the trust stack defined in Chapter 28 and separately clarifies its upgrade mechanism and value-creation path. The six systems exhibit a common evolutionary pattern: the direction of upgrade is uniformly from static rule execution toward dynamic, adaptive optimization based on real-time market states, while value creation points uniformly toward improving capital efficiency and reducing systemic risk.

Core AI-driven system	Position in the trust stack	Upgrade mechanism	Value creation
Dynamic risk-parameter system	Oracle and risk layer	From static rules to real-time parameter adjustment based on market state	Improves capital efficiency and reduces systemic risk
Intelligent matching engine	Matching and execution layer	From price-time priority to the introduction of toxicity awareness and isolation	Protects market makers and improves the quality of price discovery
Predictive liquidation system	Liquidation-mechanism layer	From post-trigger execution to probability-based prediction and prevention	Reduces liquidation cascades and eases market panic
Dynamic pricing mechanism	Asset-settlement layer	From a fixed formula to supply-and-demand-driven adaptive rate optimization	Balances long and short forces and incentivizes long-term liquidity
Adaptive compliance monitoring	Protocol-governance layer	From after-the-fact audit to real-time anomaly interception based on pattern recognition	Reduces compliance cost and enhances market integrity
Liquidity-routing optimization	Liquidity-aggregation layer	From a single path to cross-chain, cross-protocol deep-reinforcement-learning routing	Minimizes trading slippage and maximizes execution efficiency

Table 31-5. The functional matrix of the six AI systems (Data source: theoretical derivation in this chapter)

31.8.1 Dynamic risk parameters

In the traditional margin system discussed in Chapter 29, the initial margin rate and the maintenance margin rate are usually set by protocol governance as static parameters. The fundamental problem with this design is that it tries to use a single numerical value to cope with multidimensional market changes. Mainstream exchanges today all adopt a tiered margin system. Taking Binance's Bitcoin perpetual futures as an example, the initial margin rate ranges from 1% for low-leverage, small positions to 50% for high-leverage, large positions, spanning more than ten discrete tiers. This tiered design is itself a rough attempt at dynamization; the true incremental value of an AI dynamic risk-parameter system lies in turning these discrete tiers into a continuous function based on multidimensional market states. Even under the existing tiered system, however, the fundamental limitation of static parameters persists: in a calm market, a fixed margin requirement often leads to low capital efficiency, forcing traders to lock up far more margin than their actual risk requires; while in a period of violent volatility, an insufficient safety cushion may trigger a systemic account deficit.

The limitation of a static-parameter system is that it ignores the dynamic nature of

market microstructure. When market volatility suddenly jumps from 20% to 80%, the originally sufficient 10% margin immediately becomes dangerous, but the protocol cannot respond automatically. Conversely, when the market enters a calm period and volatility falls to 5%, a 10% margin requirement appears overly conservative and wastes the trader's capital. In addition, static parameters cannot account for changes in liquidity depth. In a high-liquidity environment, even if an account is close to the liquidation line, the liquidator can execute the liquidation quickly without causing excessive market impact; but in an environment where liquidity has dried up, the same liquidation could widen price slippage by more than 20%, thereby triggering a liquidation cascade.

An AI-driven dynamic risk-parameter system, by introducing multimodal deep-learning models, achieves a deep coupling between risk parameters and real-time market states. The core input features of the system include four dimensions: first, real-time volatility, for which the system updates the asset's implied-volatility estimate at the millisecond level via a GARCH model or a neural-network volatility predictor; second, market depth, for which the system monitors the liquidity depth at each price level of the order book and computes the average slippage required to execute a large order; third, cross-asset correlation, for which the system tracks the dynamic correlation coefficients among different assets and, when the correlation suddenly rises (indicating increased market stress), raises the margin requirement for all assets; and fourth, historical liquidation data, for which the system uses past liquidation events to calibrate the current risk model and identify which combinations of market states are most likely to lead to a liquidation cascade.

Based on these input features, the AI system computes a tiered margin matrix rather than a single margin rate. The dimensions of this matrix include asset type, leverage multiple, and market state. When Bitcoin's volatility is at the 25th historical percentile and order-book depth is ample, for example, the system might lower the initial margin requirement from 10% to 7%; whereas when volatility jumps to the 75th historical percentile and liquidity dries up, the system would raise the margin requirement to 15%. According to a 2025 study by Huang et al., this dynamic-adjustment approach, compared with a static-parameter system, has the potential to reduce liquidation losses in stress scenarios and improve capital efficiency under normal conditions (the specific magnitude varies by model and dataset, and precise figures still lack an independently verifiable source) [41].

The introduction of dynamic parameter adjustment changes traders' risk expectations, which also brings new design challenges. In a static model, the trader need only focus on the absolute distance of the price from the liquidation line; in a dynamic model, the trader must simultaneously assess the risk that a deterioration in market state will suddenly raise the margin requirement. This gives rise to a trade-off over the frequency of parameter adjustment. If the system adjusts the margin requirement every second, the trader faces extremely high uncertainty, which may cause overly conservative positioning and thereby reduce market efficiency; but if the adjustment frequency is too low (for example, once an hour), the system cannot respond in time to a sudden market shock. The optimal solution in practice is a tiered adjustment strategy: when the market state changes slowly, the margin parameters are updated

every 5 minutes; when a market-stress signal is detected (such as a volatility jump or a sudden drop in liquidity), the system immediately triggers an emergency adjustment, but simultaneously sends an early warning to all affected traders, giving them a buffer of 2 to 5 minutes to proactively adjust their positions.

The margin-adjustment early-warning mechanism may itself trigger a coordination-game problem. When a large number of traders simultaneously learn that the margin rate is about to be raised, the rational response is to close positions immediately to avoid a margin call, and when everyone takes this action at the same time, the collective selling pressure may trigger the very crisis the warning was meant to prevent, forming a self-fulfilling loop akin to a bank run. In addition, the margin-adjustment warning faces an information-security risk: institutional users with API-monitoring capability may detect an impending adjustment in advance through changes in the system's behavioral patterns and act first. The design of the early-warning mechanism must therefore carefully weigh the tension between information transparency and market stability.

In the CEX perpetual futures market, the dynamic risk-parameter system faces several distinctive design considerations. First, the tiered margin system requires the system to dynamically adjust the margin rate according to position size, with larger positions bearing higher margin requirements to reflect their potential impact on market liquidity. Based on real-time order-book depth and historical large-liquidation data, an AI system can dynamically optimize the margin rate and position cap of each tier, striking a finer balance between capital efficiency and system safety. Second, the AI optimization of the auto-deleveraging (ADL) mechanism is significant. The traditional ADL mechanism forcibly hedges the positions of profitable parties according to a profit ranking in order to cover account-deficit losses, a process that often provokes discontent among profitable traders and a crisis of market trust. An AI system can predict the probability of an account deficit and intervene in advance—for example, proactively adjusting the margin requirement of high-risk accounts before ADL is triggered, or optimizing the execution order to minimize market impact—thereby significantly reducing the trigger frequency and negative impact of ADL. The full trigger chain for ADL is as follows: the account reaches the maintenance-margin line, the liquidation engine takes over, an account deficit arises during liquidation execution, the insurance fund covers the shortfall, and, when the insurance fund is insufficient, ADL is triggered. AI intervention is distributed across different nodes of this chain: dynamic margining reduces the probability of triggering liquidation, acting on the first link; predictive liquidation improves execution quality and thereby reduces the probability of an account deficit, acting on the second link; and dynamic insurance-fund management reduces the probability of triggering ADL, acting on the third link. These are independent optimization objectives at different levels and should not be conflated.

In addition, the dynamic risk-parameter system must also solve the problem of "parameter transparency." A traditional protocol's static parameters are public and easy to understand, but AI-generated dynamic parameters are often a black box, and traders find it hard to predict the margin requirement of the next instant. To build trust, the protocol should publish the computational basis of the risk parameters in real time, including the current volatility estimate, the liquidity-depth score, and the correlation indicators, enabling traders to understand the reasons for parameter changes. This

transparency not only enhances user trust but also provides a basis for regulators to audit.

31.8.2 Intelligent matching engine

The matching logic of a traditional order book strictly follows the principles of price priority and time priority. Although this mechanism guarantees surface-level fairness, it leaves the market maker who passively provides liquidity fully exposed to the adverse-selection risk from informed traders. In traditional matching, a large market buy order immediately fills against the best ask on the order book, regardless of whether that ask comes from a well-informed arbitrageur or an uninformed retail trader. This indiscriminate matching mechanism causes the market maker's return curve to exhibit the typical character of a "bearer of adverse-selection losses": the market maker earns limited spread income when the market is calm, but suffers significant adverse-selection losses when the market is volatile.

As discussed in Chapter 19, toxic order flow is a core driver of liquidity droughts. Toxic flow refers to orders from traders with an informational advantage, whose execution inflicts an adverse price impact on the market maker. Traditional market makers cannot identify the toxicity of an order before execution and can only passively bear the loss. From the standpoint of market design, Budish, Cramton, and Shim (2015) [42] once proposed replacing the continuous limit order book with frequent batch auctions, aiming to fundamentally eliminate the adverse-selection problem arising from speed competition. Although that scheme faces challenges such as liquidity fragmentation in practice, its core insight—protecting market makers by changing the matching mechanism itself—is highly consistent in goal with an AI-driven, toxicity-aware matching engine. The difference between the two is that batch auctions eliminate the speed advantage by discretizing time, whereas an intelligent matching engine neutralizes the informational advantage through continuous toxicity assessment. An AI-augmented intelligent matching engine, by introducing an order-flow-toxicity identification mechanism, fundamentally reshapes the microstructure of the matching layer.

The intelligent matching engine uses deep neural networks to analyze the microlevel features of every incoming order in real time in order to predict its toxicity probability. These features span multiple dimensions: the size of the order relative to recent trading volume, since larger orders are more likely to come from informed traders; the order's submission frequency and pattern, since high-frequency small orders may indicate an algorithmic trader probing market depth; the order's cancellation rate, since a high cancellation rate indicates the trader may be spoofing or manipulating the market; the account-history features of the order submitter, including the profitability of past trades, the distribution of holding times, and whether there is any known arbitrage strategy; and the correlation of the order with macro market data, for example whether the order's direction is consistent with recent price momentum. Existing research (Cartea et al., *Detecting Toxic Flow*, focused on foreign-exchange broker-client trades) proposes an online method for predicting the toxicity of a single fill and, on that basis, deciding whether to internalize or externalize it in order to reduce adverse-selection losses; its approach can be analogously transferred to the toxicity-aware setting of order-book

market making [43].

Based on these toxicity predictions, the matching engine can implement a tiered-response strategy. For retail orders judged to be of low toxicity (toxicity probability below 30%), the system provides the best execution price and the smallest slippage, ensuring that retail traders receive the best execution quality. For orders identified as of moderate toxicity (toxicity probability between 30% and 70%), the system may introduce a microsecond-level execution delay before execution, giving the market maker an opportunity to adjust its quotes. For arbitrage or high-frequency-attack orders identified as of high toxicity (toxicity probability above 70%), the system may take more aggressive measures: routing them to a particular liquidity pool with a higher slippage tolerance, or requiring the trader to pay a higher fee before execution as a "toxicity tax," which is used to compensate the market maker's adverse-selection losses.

This toxicity-aware matching mechanism effectively mitigates the adverse-selection problem facing market makers. When market makers know that the system can filter out most predatory orders, they will be more willing to provide deeper liquidity within a narrower spread. Empirical research shows that after the introduction of toxicity-aware matching, market makers' spreads typically narrow by 20% to 30%, while market depth increases by 40% to 60%, which benefits all traders.

This mechanism, however, also gives rise to a profound controversy over market fairness. When the algorithm decides execution quality based on a trader's behavioral features rather than on the quote alone, the traditional principle of equal market access faces the challenge of redefinition. A legitimate retail trader may be misjudged by the system as toxic flow because their trading pattern resembles that of a high-frequency arbitrageur, and thereby suffer worse execution quality. To address this compliance risk, the protocol needs to establish a transparent toxicity-scoring mechanism, disclosing to the trader the reasons their order was scored as highly toxic and providing an appeal mechanism. In addition, the protocol should conduct regular fairness audits to ensure that the toxicity-identification model does not systematically discriminate against particular types of traders.

Examined more deeply, the toxicity-aware matching engine faces structural limitations along four dimensions. First is the inevitable operation of Goodhart's law: once a toxicity-scoring mechanism is deployed, informed traders will immediately adjust their behavior to evade a high-toxicity judgment, including strategies such as splitting order size, mimicking retail trading patterns, and dispersing order flow across multiple unlinked accounts. The backtested accuracy of such a toxicity-identification model is in essence a static result based on historical data, and in an adversarial live environment its accuracy may rapidly decay, forcing the model into a continual offense-defense iteration. Second, differentiated execution delay poses a fundamental contradiction at the level of market design: imposing different execution delays on different orders fundamentally violates the value proposition of a centralized exchange, whose core competitiveness is deterministic low latency. Facing uncertain execution timing, a market maker's rational response is precisely to widen its quote spread to compensate for the execution uncertainty, which ultimately harms overall market quality—forming a paradox with the original intent of protecting market makers. Notably, existing non-discriminatory

design schemes in the field of market microstructure offer an entirely different line of thought: the Investors Exchange (IEX) speed bump imposes a uniform, fixed delay on all participants, rather than a differentiated delay targeting particular orders; likewise, the frequent-batch-auction scheme mentioned earlier adopts a discretized-time mechanism that treats all orders alike. The common feature of these two designs is non-discrimination, which is essentially different from differentiated execution based on a toxicity score. Third, examined from the angle of legal compliance, the best-execution obligation under U.S. securities law requires brokers to seek the best execution for all clients, while MiFID II in the EU expressly stipulates that a trading venue's rules must be transparent and non-discriminatory. AI-toxicity-score-based differentiated execution quality may come into direct conflict with these legal requirements, especially when the internal logic of the toxicity-scoring model is difficult to explain fully to regulators, further amplifying the compliance risk. Fourth, in the context of a centralized exchange there is also a potential conflict-of-interest hazard: when an exchange simultaneously operates the matching engine and participates directly or indirectly in the market-making business, the toxicity-scoring system may provide an informational advantage to the exchange's affiliated market makers, enabling them to identify and evade informed order flow in advance, while independent market makers and ordinary traders cannot obtain the same protection.

In the CEX order-book market, the implementation path of an intelligent matching engine differs markedly from that of an on-chain protocol. The CEX's centralized matching architecture naturally supports microsecond-level order processing and toxicity assessment, without having to contend with the constraint of blockchain block time. Specifically, a CEX can layer an AI toxicity-filtering layer on top of the traditional price-time-priority matching logic: when a new order enters the matching queue, the system completes the toxicity scoring within microseconds and, on that basis, decides whether to introduce an execution delay, adjust the execution price, or route the order to a dedicated liquidity pool. This architecture enables the CEX to provide effective adverse-selection protection for market makers without changing the basic matching rules, thereby incentivizing deeper order-book liquidity.

31.8.3 Predictive liquidation

Chapter 29 analyzed in detail the operating logic of the liquidation mechanism; the traditional liquidation system is in essence passive and reactive. Only when an account's collateralization ratio falls below the maintenance-margin threshold is the liquidation engine triggered to dump assets onto the market. In extreme conditions, this mechanism readily triggers a liquidation cascade: the selling pressure caused by liquidation further depresses the price, triggering the liquidation of more accounts, and ultimately becoming a collapse of market liquidity. The liquidation events of multiple DeFi protocols in 2024 show that during periods of extreme volatility, a liquidation cascade can cause a protocol's total collateral value to fall by more than 30% within minutes, resulting in a large number of involuntary liquidations for users.

A predictive liquidation system breaks this passive cycle through machine-learning models. The system uses long short-term memory (LSTM) networks and other recurrent-

neural-network architectures to analyze data along multiple dimensions in order to compute the probability that each leveraged account will be liquidated within a particular future time window. These features include: the account's current leverage ratio and its distance from the liquidation line; the short-term momentum and volatility of the asset price; the real-time depth and trend of market liquidity; historical liquidation events with account features similar to the current account and their outcomes; and macro market-stress indicators, such as the volatility index (VIX) and cross-asset correlations. By integrating these features, the predictive model can estimate the liquidation probability with fairly high accuracy. According to a 2024 study by Palaiokrassas et al., a variety of machine-learning models can classify wallet-level liquidation and report metrics such as accuracy and AUC (area under the curve), demonstrating the effectiveness of DeFi on-chain features for liquidation prediction (the specific figures depend on their dataset and evaluation criteria) [44].

Based on this probability prediction, the system implements a tiered action model, moving risk management forward rather than responding after the fact. Figure 31-8 shows the structure of this model: when an account's liquidation probability exceeds the extremely-high-risk threshold of 90%, the system no longer waits for the actual liquidation to trigger but proactively executes a partial position reduction, with the goal of lowering the liquidation probability below 70%; because the reduction is proactive and dispersed, the market impact is typically only 20% to 30% of that of a conventional liquidation.

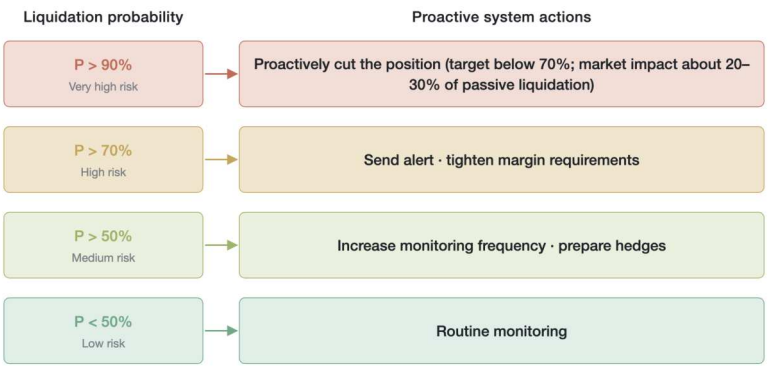


Figure 31-8. The probability-tiered action model of predictive liquidation (conceptual illustration; based on the author's theoretical derivation)

When the liquidation probability is between 70% and 90%, the system enters a second-tier response, sending a high-priority early-warning notification to the trader. This warning not only informs the user that their account is in a dangerous state but also provides specific recommendations: how much margin needs to be added, or how much position needs to be reduced, to restore safety. The warning is sent through multiple channels, including on-chain events, email, and push notifications, ensuring that the trader receives the information in time. The trader has 5 to 10 minutes to proactively adjust their position; if the trader takes no action within this time window, the system

then executes auto-deleveraging.

When the liquidation probability is between 50% and 70%, the system enters a third-tier response, increasing the monitoring frequency to the second level and preparing to respond quickly if necessary. At this point the system does not proactively intervene but continually assesses the risk, and once the probability breaks through the 70% threshold, it immediately triggers the second-tier response. When the liquidation probability falls below 50%, the system returns to the normal monitoring cycle (usually at the minute level), and the account is regarded as being in a safe state.

This paradigm shift from passive execution to proactive prevention mitigates the transmission chain of a liquidation cascade. By smoothly releasing risk before a crisis erupts, the predictive system substantially reduces the probability of a systemic collapse. According to empirical analysis, after the introduction of a predictive liquidation system, the scale of liquidation cascades under extreme market conditions is typically reduced by 60% to 80%, and the overall stability of the system improves significantly.

At the same time, this mechanism has a profound impact on the role of the liquidator. Traditional arbitrage-type liquidators profit by buying collateral at a steep discount when the market is in turmoil, a model that encourages liquidators to liquidate when market pressure is greatest, thereby intensifying market panic. In a predictive liquidation system, by contrast, most liquidations take place at the moment of least market pressure, and the profit space for liquidation shrinks substantially. This causes traditional arbitrage-type liquidators to be gradually replaced by the protocol's own algorithmic engine, and the distribution of liquidation profit likewise flows back from external extractors to the protocol itself or to protected users.

The predictive liquidation system, however, also gives rise to profound concerns about "moral hazard." If the protocol commands a more accurate liquidation-probability prediction than the market, will it use this informational advantage to harm users' interests? For example, the protocol might deliberately delay a liquidation so that the account's liquidation probability rises further, enabling it to execute the liquidation at a lower price. To prevent such abuse, the protocol needs to establish a strict governance mechanism, ensuring that all parameters and decision rules of the predictive liquidation system are transparent and are subject to regular independent audits. In addition, the protocol should provide users with real-time data on the "liquidation-probability prediction," enabling users to independently verify the system's decisions.

Beyond moral hazard, predictive liquidation also faces deeper challenges on two fronts: user rights and data reliability. As regards user consent, the "proactive position reduction" that predictive liquidation implements is in essence a forced liquidation executed before the user's margin has reached the liquidation line explicitly agreed in the contract. The trigger condition for traditional liquidation is an explicit, predictable contractual term, whereas "a liquidation probability exceeding 90%" is a model estimate that the user cannot independently verify. In a scenario where the price reverses in a V shape, the model may close out early a position that would have returned to profit, causing an unnecessary loss. Predictive liquidation terms must therefore obtain the user's explicit informed consent when the position is opened, and a compensation mechanism for model misjudgment should be established. As regards the reliability of input data, the

fatal weakness of predictive liquidation is that it depends entirely on the quality of the input data. Oracle price-feed failures and interruptions of exchange data interfaces are not rare in practice; the March 2023 oracle price-latency incident on a particular Layer 2 network once caused a large number of improper liquidations. When the input data are erroneous, the more "precise" the predictive model, the more dangerous it becomes: it will execute an erroneous preventive reduction with extremely high confidence based on erroneous data, and the resulting loss may far exceed that of a traditional passive liquidation mechanism.

31.8.4 Dynamic pricing

The funding rate of perpetual futures and the trading fee of decentralized exchanges are core economic levers for regulating the balance of market supply and demand. In traditional designs, these rates are usually computed based on a hardcoded linear or piecewise function whose parameters are often hard to adapt to a rapidly changing market environment. The standard formula adopted by mainstream perpetual futures protocols is $F = P + \text{clamp}(I - P,\ c_{\min},\ c_{\max})$, where F is the funding rate, P is the average premium index (which reflects the degree of deviation between the perpetual futures price and the mark price), I is the interest rate, and $[c_{\min}, c_{\max}]$ are the preset lower and upper bounds within which the clamp function constrains the funding rate; these bounds are usually set as a symmetric interval (for example, $\pm 0.05\%$, corresponding to an 8-hour settlement period), with the specific values determined by protocol parameters. Every term in the formula is expressed as a rate over a single settlement period (8 hours by default); the interest rate I is the value normalized to that settlement period (for example, 0.03%/8h) rather than an annualized value. The problem with this design is that the time-weighting window of the premium index is fixed, the upper and lower bounds of the clamp function are static presets, and the settlement period is rigidly locked at once every 8 hours, so the whole mechanism cannot anticipate dynamic market changes and can only passively react to the current imbalance.

An AI-driven dynamic pricing mechanism uses reinforcement learning algorithms to achieve a global adaptive optimization of the rate system. On the funding-rate side, the dynamic pricing system carries out an intelligent transformation around the three core parameter dimensions of the standard formula above. First, the time-weighting window of the premium index can be dynamically adjusted according to the market's volatility regime: a longer window is used in a low-volatility environment to filter out noise, and the window is shortened in a high-volatility environment to speed up the response to price deviations. Second, the upper and lower bounds of the clamp function are no longer static constants but dynamic parameters that change with market conditions—tightening the constraint range in a market with ample liquidity and stable volatility to keep the rate smooth, and moderately loosening the bounds when the market shows violent deviation to strengthen the rate's regulatory force. Third, the settlement frequency shifts from a fixed once-every-8-hours to a variable period based on market state, raising the settlement frequency to speed up price convergence when the contract price and the mark price deviate persistently and sharply, and lowering the frequency to reduce unnecessary trading friction when the prices are closely aligned. The system also uses a time-series forecasting model to predict the short-term evolution

of the premium index; if the forecast indicates that the perpetual futures price is about to deviate sharply from the mark price, the system adjusts the above parameters in advance, thereby beginning to intervene before the imbalance intensifies.

Machine-learning models, by predicting the marginal impact of different rate levels on the behavior of market participants, dynamically adjust the slope and intercept of the rate curve so as to incentivize arbitrageurs to enter and smooth out the price deviation in the most efficient way. For example, when the spot-market price of Bitcoin is \$65,000 while the perpetual futures price is \$65,500, a traditional system might simply raise the funding rate to 0.05% in the hope of attracting arbitrageurs. But an AI system would analyze more finely: How many arbitrageurs are currently active in the market? How many new arbitrageurs would raising the funding rate to 0.05% attract? How long would it take these arbitrageurs to close their positions? Based on this analysis, the system might decide to raise the funding rate to 0.08%, or simultaneously lower the trading fee to 0.02%, so as to attract arbitrage flow more efficiently. The 0.05%, 0.08%, and 0.02% here are illustrative decision values used to characterize how an AI system might weigh its options, and are not computed by substituting P and I into the standard formula $F = P + \text{clamp}(I - P,\ c_{\min},\ c_{\max})$ above.

On the fee-optimization side, a reinforcement learning agent dynamically sets the optimal trading fee according to the real-time depth of the liquidity pool, expected volatility, and the quotes of competing protocols. This optimization problem involves a trade-off among multiple objectives: lowering the fee can attract more trading volume but reduces protocol revenue; raising the fee can increase revenue but causes trading volume to fall and may even drive users to competitors. By continually trying different fee levels, the reinforcement learning agent learns a Nash equilibrium point at which the protocol's total revenue (trading volume times fee rate) is maximized.

According to empirical research, an AI dynamic pricing mechanism, compared with a static rate, can significantly increase the protocol's total revenue (an estimate based on the author's reading of industry cases) while maintaining or improving users' execution quality. This is because an AI system can lower fees to attract trading volume when liquidity is ample, and raise fees to protect market makers when liquidity is tight, thereby achieving an optimal supply-demand balance across different market states.

This dynamic pricing mechanism enables the protocol to implement a personalized fee structure. Based on a user's historical trading behavior, position stability, and contribution to the protocol's liquidity, the system can offer differentiated pricing schemes to different participants. For example, long-term holders and stable liquidity providers might receive a lower fee discount, while high-frequency traders and arbitrageurs might have to pay a higher fee. This personalized pricing not only enhances the protocol's competitiveness but also marks the evolution of DeFi's pricing logic from a mechanical, formulaic approach toward an intelligent, commercial one.

Personalized pricing, however, also gives rise to a regulatory risk over "price discrimination." If the protocol charges different fees to different users, does this constitute discrimination against certain users? To address this risk, the protocol should establish transparent pricing rules, publicly disclosing to all users the method for computing fee discounts and ensuring that the pricing rules do not have a systematically adverse

effect on a particular group of users. A more fundamental problem is that there is a structural conflict of interest between the AI dynamic pricing objective of "maximizing protocol revenue" and the user's objective of "minimizing transaction cost." The win-win narrative of "increasing revenue while improving execution quality" holds only under particular conditions—namely, in periods when users' demand elasticity is relatively high; in periods when liquidity is tight, an AI system may in effect become a tool for charging users higher fees, which is logically highly similar to the surge-pricing controversy of ride-hailing platforms. At the level of legal compliance, personalized pricing relies on collecting and analyzing users' trading histories, which may conflict with the data-minimization principle of the EU General Data Protection Regulation (GDPR). Differentiated pricing based on behavioral features, if it causes the pricing model to be systematically adverse to a particular group of users, may constitute algorithmic discrimination. In addition, multiple jurisdictions have explicit regulatory requirements for fairness in securities pricing, and a differentiated fee structure faces the risk of compliance scrutiny in these jurisdictions.

31.8.5 Adaptive compliance monitoring

As the scale of on-chain finance expands, regulatory compliance becomes a necessary condition for a protocol's long-term survival. Chapter 30 explored the static implementation of embedded compliance, but this filtering mechanism based on whitelists and fixed rules often struggles to cope with complex and varied means of market manipulation. An adaptive compliance-monitoring system, by integrating AI technology, achieves a leap from rule matching to pattern recognition.

The system uses unsupervised machine learning and graph neural networks to scan vast amounts of on-chain transaction data in real time. The system's monitoring is divided into three levels. The first level is real-time transaction-pattern analysis, in which the system extracts features from each transaction—including transaction size, direction, timestamp, trading pair, and the age of the trader's account—and then, through anomaly-detection algorithms such as the isolation forest or the local outlier factor, identifies transactions that deviate markedly from normal trading patterns. The second level is account-behavior clustering, in which the system clusters accounts according to their trading-behavior features and identifies groups of accounts with similar behavioral patterns, which may represent a coordinated manipulation team. The third level is cross-account association identification, in which the system, by analyzing the flow of funds among accounts, the correlation of trading times, and the correlation of positions, identifies groups of accounts that may have an interest linkage.

Through these three levels of analysis, the system can identify potential wash trading, front-running attacks, and price manipulation. The features of wash trading include: the trading counterparty is the same account or a linked account, the interval between trades is extremely short (usually at the second level), the trade price is unrelated to the market price, and the trade size exhibits a periodic pattern over a period of time. The features of a front-running attack include: the attacker submits an order before a large trade executes and immediately closes the position after the large trade executes, with the profit coming from the price impact the large trade causes. The features of

price manipulation include: one or more accounts conduct a large number of trades within a short time, causing the price to swing sharply, and then these accounts close their positions at the price peak to profit.

Unlike a traditional alert system that relies on preset thresholds, an AI monitoring model can autonomously learn the baseline of normal market behavior and automatically trigger a blocking mechanism when it detects a complex, coordinated attack that deviates from the baseline. Existing research has explored using AI for the automatic identification of behaviors such as wash trading and price manipulation (Popoola, 2023 [45]), as well as the potential of machine-learning-based market surveillance, relative to traditional rule-based systems, to reduce false positives and improve detection coverage (Tiwari et al., 2021 [46], a survey); the reduction in false positives reported in real deployment cases is usually on the order of tens of percentage points.

The introduction of an adaptive compliance system not only reduces the protocol's regulatory risk but also provides the necessary infrastructure guarantee for the large-scale entry of institutional funds. Before entering a new market, institutional investors usually require that the market have sound anti-fraud and anti-manipulation mechanisms. A protocol equipped with an AI compliance-monitoring system can demonstrate the integrity of its market to institutional investors, thereby attracting more institutional funds.

More importantly, an adaptive compliance system changes the paradigm of governance. The traditional compliance model relies on manual review, a lagging and costly process. A trader may be discovered only hours or even days after conducting a manipulative act, by which time substantial damage has already been done. An AI-driven compliance system, by contrast, can detect and block a manipulative act in real time as it occurs, transforming compliance from an after-the-fact audit into a real-time anomaly-detection-and-blocking mechanism. This not only improves defensive efficiency but also substantially reduces the cost of compliance.

AI compliance monitoring, however, also gives rise to concerns about "privacy" and "false positives." If the system is overly sensitive, it may misjudge normal trading behavior as manipulation, thereby harming innocent traders. To address this risk, the protocol should establish a clear appeal mechanism, so that when a trader believes their trade has been wrongly flagged as manipulation, they can submit an appeal and obtain a manual review. In addition, the protocol should regularly publish statistics on compliance monitoring, including the number of manipulation cases detected and the false-positive rate, in order to accept community oversight.

31.8.6 Liquidity-routing optimization

In an ecosystem where multiple chains coexist and liquidity is fragmented, finding the optimal execution path for a large trade is a core challenge facing the liquidity-aggregation layer of Chapter 28. A $1 million trade may need to be split for execution across multiple protocols such as Uniswap, Curve, and Balancer in order to minimize slippage. Traditional routing algorithms usually rely on static graph search or simple linear programming; these methods are computationally efficient but often suboptimal in the quality of optimization. They struggle to achieve a truly optimal solution amid

dynamically changing market depth and network congestion.

An AI-driven liquidity-routing-optimization system, through deep reinforcement learning, raises order execution to a new level of efficiency. The system formulates order routing as a Markov decision process, in which the state space includes all available liquidity pools and their current depths and prices, the action space includes all possible ways of allocating the order among the various liquidity pools, and the reward function is defined as minimizing total slippage and transaction cost.

The reinforcement learning agent continually explores the cross-protocol, cross-chain distribution of liquidity in this vast state space. The agent considers not only the instantaneous price and depth of each liquidity pool but also predicts the slippage changes that may occur during trade execution. If the agent decides to execute a $500,000 trade on Uniswap, for example, it must predict how much price impact this trade will cause on Uniswap and thus how much slippage will result. In addition, the agent must consider network latency and gas-fee fluctuations. When the Ethereum network is congested, the gas fee for executing a single trade may be as high as $100, whereas when the network is idle, the same trade may cost only $10 in gas. An intelligent routing system should choose fewer trade splits when the network is congested, so as to reduce gas costs.

By intelligently splitting a large order and dynamically allocating it across multiple optimal paths, the system can complete complex cross-market arbitrage and liquidity extraction within milliseconds. According to a 2025 study by Zhang and Tessone on an enhanced DEX routing method, the improved routing method can, compared with the baseline, increase DEX execution efficiency and the benefits to all parties (the specific figures depend on their evaluation criteria and dataset) [47].

The intelligentization of liquidity routing profoundly changes the competitive landscape of decentralized exchanges. In the traditional model, a protocol's competitiveness depends mainly on its liquidity depth and fee level. But in the era of intelligent routing, an isolated liquidity pool no longer constitutes a competitive barrier, and capital flows frictionlessly toward the protocol that prices most efficiently. This means that even if a protocol's liquidity depth is not the deepest, an intelligent routing system will still preferentially use it if it prices most efficiently. This competitive model encourages all protocols to pursue the highest pricing efficiency, thereby improving the market quality of the entire ecosystem.

At the same time, the intelligentization of liquidity routing also intensifies the technological competition among routing algorithms. Aggregators with the strongest computing power and the best models will dominate the allocation of order flow, thereby exerting a decisive influence on the survival of the underlying protocols. This gives rise to a new market structure: a small number of aggregators with the most advanced AI capabilities control the majority of order flow, while the underlying protocols are reduced to liquidity providers for these aggregators. This trend toward centralization may have a negative impact on the long-term health of DeFi, and so excessive centralization must be prevented through open standards and transparent algorithms.

31.9 Semantic oracles

31.9.1 The capability boundary of traditional oracles

The oracle mechanism is the core infrastructure connecting on-chain smart contracts with the off-chain real world. In Chapter 28's analysis of the fourth layer of the trust stack, the oracle is defined as a one-way data-transmission channel whose main function is to obtain the latest price of a particular asset from an external data source and convert it into an on-chain-verifiable structured format [48]. Although this design has supported the prosperity of DeFi over the past few years, it has also established the fundamental limitations of the traditional oracle.

The first limitation lies in the strict constraint on data type. Traditional oracles are designed to transmit highly structured numerical data, including exchange asset prices, interest-rate indices, weather data, or other already-quantified indicators. The common feature of these data is that they have a clear numerical representation and can be directly processed by a smart contract's conditional logic. When facing unstructured information, however, traditional oracles are powerless. A public company's financial report, for example, is usually a PDF document of hundreds of pages, containing complex textual descriptions, tables, footnotes, and explanations of accounting adjustments. A traditional oracle cannot automatically extract the key metric "third-quarter revenue" from such a document, still less understand the complex accounting treatment of "one-time gains" or "nonrecurring losses" in the report. Likewise, the key information in a legal contract or a government statement is often implicit in natural-language expression, and a traditional oracle cannot parse its semantic meaning. This dependence on structured data strictly confines the scope of oracle applications to asset classes that already have market quotes or have already been quantified.

The second limitation stems from the inability to handle the ambiguity of natural language. Natural language is inherently full of ambiguity, metaphor, context dependence, and polysemy. When a contract condition is written as "the company's performance declines significantly," for example, the word "significantly" may have entirely different meanings in different contexts. In the medical field, "significant" may mean a statistical p-value below 0.05; in finance, it may mean a year-over-year decline of more than 20%; and in everyday language, it may be merely a vague qualitative description. A traditional oracle cannot understand this semantic flexibility; it can only mechanically look up a predefined numerical value or category. If a contract condition involves any expression requiring subjective judgment or semantic understanding, a traditional oracle is not up to the task.

The third limitation lies in the inability to handle "subjective judgment" questions. In the real world, many important events or conditions are inherently subjective. The question "whether a company has violated the financial covenants in its loan agreement," for example, involves not only numerical computation but also the legal interpretation of contract terms. The question "whether a politician's remarks constitute hate speech" involves an understanding of cultural background, historical context, and social norms. The question "whether an artwork has collection value" involves aesthetic judgment and market expectation. Traditional oracles were not designed with such questions in

mind, because these questions cannot be resolved by a simple data query. They require a system able to understand context, weigh multiple factors, and reason and judge.

The vision of "perpetual futures on everything" described in Chapter 27 has at its core the aim of creating continuous derivatives markets for any asset or any definable event. This means the system must be able to handle an extremely diverse range of underlying assets, from traditional financial assets to emerging non-standard assets, and from explicit numerical conditions to complex event determinations. This poses new technical requirements for the oracle. First, the oracle must be able to accept any contract condition described in natural language, not just a few predefined data types. Second, the oracle must be able to understand the semantic meaning of these natural-language conditions, including handling ambiguity, identifying implicit assumptions, and understanding context. Third, the oracle must be able to find or infer the relevant verification information in the real world, even if that information does not exist in a standardized market-quote form. Finally, the oracle must be able to perform some degree of reasoning and judgment to determine whether an event that occurred in the real world satisfies the contract condition.

The design of traditional oracles cannot meet these requirements at all. Even if we try to extend the existing oracle architecture, we can only handle more kinds of structured data; we still cannot cross into the domain of unstructured information and semantic understanding. This is precisely why "perpetual futures on everything" has been strictly limited technically. So long as the oracle remains in the role of "data transmission," the feasibility boundary of perpetual futures on everything will long be confined to the range of assets that have already been fully marketized and standardized. For all non-standard assets and events requiring semantic understanding, traditional oracles cannot meet these requirements.

31.9.2 Semantic oracle architecture

The semantic oracle represents a paradigm shift in oracle architecture; by embedding large language models (LLMs) into a network of nodes, it achieves the leap from mere "data transmission" to "semantic understanding." The core advantage of the semantic oracle is its ability to read and understand complex conditions described in natural language and to convert unstructured textual information into deterministic instructions that a smart contract can execute. The architecture of this new type of oracle usually comprises four mutually coordinated layers, each bearing a different functional responsibility [49].

The input layer is responsible for receiving and aggregating unstructured information from multiple heterogeneous sources. Unlike a traditional oracle, which can receive only predefined data requests, the input layer of a semantic oracle can handle natural-language queries in any form and perform preliminary semantic normalization, ensuring that different expressions with the same meaning are recognized as equivalent conditions. The input layer simultaneously aggregates information from multiple sources—news reports, social media, official announcements, financial databases, and descriptions of on-chain activity—to provide high-quality material for subsequent in-depth analysis.

The LLM understanding layer is the cognitive core of the semantic oracle. Advanced

NLP models at this layer perform text comprehension and semantic parsing, key-condition extraction, ambiguity identification and resolution, and polysemy handling. When the Federal Reserve releases a monetary-policy statement, for example, the model must not only identify the keyword "rate cut" but also understand its magnitude, timetable, and economic impact; when a financial report mentions "30% revenue growth," it must distinguish year-over-year from quarter-over-quarter and total revenue from segment-level revenue.

The verification layer is designed to address the hallucination, bias, or manipulation that a single model may produce. This layer, through a multi-node consensus mechanism, requires multiple independently running LLMs to assess the same event independently [50], uses a sentence-level encoder to map text into a vector space and compute semantic similarity, and combines a truth-discovery algorithm to weight and aggregate the results based on nodes' historical credibility [49]. Experimental data show that even in the extreme case where 40% of the nodes are maliciously manipulated, this consensus mechanism can still maintain a relatively high accuracy.

The output layer is responsible for converting the consensus-verified semantic conclusion into an on-chain-verifiable structured data format, including a Boolean value, a numerical value, or other predefined data types. The output layer also attaches the model's confidence score for the conclusion or an uncertainty interval, enabling a smart contract to distinguish high-confidence from low-confidence determinations and to take conservative action—such as delaying settlement or triggering human arbitration—when facing a highly uncertain situation.

Figure 31-9 compares the semantic oracle with the traditional oracle along six capability dimensions: the traditional oracle is well-developed in the transmission of structured numerical values but has zero capability along dimensions involving natural-language understanding; the semantic oracle, by embedding LLMs, achieves a qualitative change from infeasible to feasible, although its precision is still constrained by model hallucination and the consensus mechanism.



Figure 31-9. A comparison of the capability dimensions of the semantic oracle and the traditional oracle (conceptual illustration; not empirical data; based on the author's theoretical derivation of architectural characteristics [48][49])

31.9.3 Unlocking perpetual futures on everything

The introduction of the semantic oracle fundamentally changes the conclusions of Chapter 27 regarding the feasibility of underlying assets; its unlocking effect manifests in bringing non-standard assets and complex events, which traditional oracles cannot handle, into the scope of programmable finance. This unlocking effect can be clearly demonstrated through three typical scenarios, each representing a new class of financial application that could not previously be realized.

Political-event perpetual futures are the first domain to be significantly unlocked. Traditional prediction markets rely on human arbitration when handling political events, a process that is time-consuming and prone to dispute. A semantic oracle can monitor and parse official legislative documents and authoritative news sources in real time; when a contract condition is set as "whether a particular bill is signed in a particular year," it can automatically confirm whether the condition is satisfied by analyzing the bill's text, signing records, and official statements, eliminating the delay and subjectivity of human intervention. The technical feasibility of political-event perpetual futures, however, is not the same as legal feasibility. In most major jurisdictions, such contracts face serious legal obstacles. The CFTC takes a cautious stance in regulating event contracts and has questioned the legal status of some political-event contracts. Multiple countries classify political-event betting as gambling rather than a financial activity,

applying gambling regulations rather than financial regulations. The semantic oracle therefore unlocks more of a technical-feasibility boundary, and the actual deployment of these products remains subject to the evolution of each jurisdiction's regulatory stance.

Corporate-performance perpetual futures likewise benefit from the semantic oracle's comprehension capability. Traditional oracles cannot directly extract specific financial metrics from a financial-report PDF, and the ingestion of structured data usually takes weeks. A semantic oracle can parse financial-statement text in real time, accurately identify specific metrics, understand the terminological differences between GAAP and IFRS, and distinguish one-time gains from regular operating performance, enabling microlevel events based on corporate fundamentals to become the underlying of perpetual futures.

Natural-language-condition derivatives represent the ultimate unlocking effect brought by the semantic oracle. A trader can create a contract using any natural-language condition, for example "the global average temperature in 2025 exceeds the pre-industrial level by 1.5 degrees Celsius" or "a particular country's unemployment rate exceeds 10% in a particular quarter." So long as the condition can be verified by public information, the semantic oracle can undertake the responsibility of parsing and determining it, achieving perpetual futures on everything in the true sense.

The semantic oracle's enhanced capability is equally significant for the CEX perpetual futures market. In centralized exchanges' business expansion, listing a new perpetual futures underlying has long been constrained by the availability of a reliable price-feed source. Traditional oracles can support only assets that already have continuous market quotes, which confines the range of CEX perpetual futures underlyings to mainstream crypto assets and a few traditional financial assets. The introduction of the semantic oracle enables the CEX to break through this limitation, bringing non-standard indices (such as a particular DeFi-sector index or a Layer 2 ecosystem-activity index) and macro-event contracts (such as central-bank rate decisions or major regulatory-policy changes) into the range of perpetual futures underlyings. This expansion of underlyings not only enriches the CEX's product line but, more importantly, provides traders with fine-grained risk-hedging tools. For example, a development team focused on a DeFi lending protocol can hedge the systemic risk of its protocol token through a DeFi-sector-index perpetual futures.

Table 31-6 summarizes the unlocking effect of the semantic oracle on different types of underlying assets. The table compares the feasibility assessments of the traditional oracle and the semantic oracle one by one across five categories: crypto assets, foreign exchange, stock indices, carbon credits / house-price indices, and natural-language-condition events. The most striking contrast in the table appears in the last row: natural-language-condition events are "wholly infeasible" under the traditional oracle, whereas the semantic oracle raises them to the level of "possible but with limited precision"—and this breakthrough is precisely the technical basis on which the vision of "perpetual futures on everything" can move from theory to practice.

Underlying-asset type	Traditional-oracle feasibility	Semantic-oracle feasibility	New possibilities unlocked
Crypto assets	Naturally feasible	Can be enhanced	Multi-source verification improves manipulation resistance
Major foreign-exchange pairs	Exchange data obtainable	Can be enhanced	Semantic understanding of central-bank statements and expectation pricing
Stock indices	Difficult outside trading hours	Feasible	Fills the pricing vacuum during non-trading hours such as the Asian session
Carbon credits / house-price indices	No continuous market quotes	Difficult but possible	Infers prices from statistical data and transaction records
Natural-language-condition events	Wholly infeasible	Limited precision	Turns any non-standard event into a settleable contract underlying

Table 31-6. The matrix of the semantic oracle's unlocking effect on the feasibility of "perpetual futures on everything" (Data source: theoretical derivation based on model capability [49])

31.9.4 The trust problem of semantic verification

Although the semantic oracle greatly expands the application boundary of smart contracts, it also introduces an entirely new trust problem: how can we verify that an LLM's "understanding" of complex semantics is accurate and error-free? Unlike a traditional oracle that transmits deterministic numerical values, a semantic oracle outputs a probability-based estimate, which makes the fairness of liquidation depend directly on the quality of the AI's estimate. If the model hallucinates and generates a seemingly reasonable but actually erroneous interpretation of information, it may cause an enormous position to be wrongly liquidated based on an erroneous price or condition [51]. This risk poses a severe challenge to the stability of the DeFi system.

To meet this challenge, the system must establish a multi-level mechanism of verification and checks and balances. Multi-node consensus is the first line of defense and the most basic protective measure. As stated earlier, by requiring multiple independently trained LLMs to assess independently and aggregating the results, one can effectively reduce the risk of a single model's hallucination. The design of this multi-node consensus requires careful consideration of several key parameters. The first is the choice of the number of nodes. In theory, the more nodes, the better the system can resist attacks by malicious nodes. More nodes, however, also mean higher cost and a longer consensus time. In practice, a balance between security and efficiency is usually needed. A common choice is to use an odd number of nodes, such as five, seven, or 11, so as to avoid a tie. The second is the criterion for selecting nodes. Not all LLMs are suitable to join the consensus network. The system needs to assess each model's historical accuracy,

whether it is easily deceived by particular types of adversarial examples, and whether it has known biases. The third is the setting of the consensus threshold. The system needs to decide how many nodes' concurring opinions constitute a valid consensus. For example, is it necessary for all nodes to agree, or only a majority, or a weighted majority? Different threshold settings affect the system's security and availability.

Even with a sound multi-node consensus mechanism, however, the system still faces a deeper risk: because mainstream LLMs have a relatively high degree of homogeneity in their training data and algorithmic architecture, they may jointly make the same mistake when facing certain particular edge cases [52]. This phenomenon, called "strategy homogenization," means that multiple seemingly independent models may in fact fail in the same way. For example, if all LLMs are trained on similar internet data, they may all inherit the same biases or misinformation present in that data. If a particular piece of misinformation is widely disseminated on the internet, all models may learn this misinformation and thus give the wrong answer when facing the related question. This risk means that a simple majority-vote mechanism may fail in certain scenarios—the concurring opinion of multiple nodes does not necessarily mean the conclusion is correct. This common-mode failure risk also extends to the level of adversarial attack. A prompt-injection attack can manipulate an LLM's output by embedding special instructions in the input text; if an attacker constructs a carefully designed piece of misinformation containing prompt-injection content able to bypass the models' safety filters, then even with multi-node consensus, all nodes using the same base model may give the same wrong result. Mitigation strategies should include mandating that nodes in the consensus network use base models from different vendors, implementing strict input-sanitization mechanisms, and introducing non-LLM verification paths as a heterogeneous security layer.

An adversarial-verification mechanism therefore becomes a necessary complement. In this mechanism, the system deploys not only a "primary model" responsible for generating the conclusion but also a "challenger model" specifically designed to look for logical flaws or factual errors. The challenger model's goal is to generate adversarial examples or raise counter-evidence, forcing the primary model to provide a more rigorous logical basis for its conclusion. This adversarial process is analogous to cross-examination in a courtroom, where the prosecution and the defense each try to find the flaws in the other's argument. In the context of the semantic oracle, the challenger model tries to interpret the same information from different angles, look for possible alternative explanations, or raise relevant information the primary model may have overlooked. For example, if the primary model's conclusion is "the company claims 50% revenue growth in its financial report," the challenger model might raise "but this growth comes mainly from a one-time gain and does not represent the growth of the regular business" or "this figure is calculated in nominal dollars, and the inflation-adjusted growth rate is in fact much lower." Only when the primary model's conclusion can withstand the challenger model's repeated attacks will the conclusion be adopted by the system. This courtroom-debate-like adversarial process can significantly improve the robustness of semantic parsing, because it forces the system to consider multiple perspectives and possible alternative explanations.

To incentivize the challenger model to perform its duty conscientiously, the system can

design a corresponding economic-incentive mechanism. For example, if the challenger model successfully finds an error in the primary model, the challenger can obtain a reward, while the primary model is penalized. This incentive mechanism ensures that the challenger has an incentive to look for flaws conscientiously rather than perfunctorily. At the same time, to prevent the challenger from maliciously raising groundless objections, the system can require the challenger to provide evidentiary support for its objections, and if an objection is proven to be wrong, the challenger is likewise penalized.

Even with a sound multi-model consensus and adversarial verification, the semantic oracle still needs to set an uncertainty threshold. When the system assesses that the uncertainty interval of a particular event exceeds a preset safety threshold, it must trigger an automatic fallback mechanism. This fallback mechanism may include suspending trading in the relevant contract, delaying settlement until more information is obtained, or triggering a human-arbitration process. This conservative approach ensures that the system does not make a hasty decision when facing a highly uncertain situation, thereby protecting the interests of market participants.

As the ultimate backstop, for a semantic determination involving extremely high value or serious dispute, the system must retain the power of human arbitration. This does not mean a return to the human-arbitration model of traditional prediction markets, but rather a more fine-grained hybrid model. In most cases, the system relies on the automated semantic oracle to make determinations and settle. But when the system detects a high degree of uncertainty or dispute, it automatically submits the case to a decentralized team of human arbitrators. This team may be composed of experts from different backgrounds and professional fields, who vote based on the relevant evidence. This hybrid model retains both the high efficiency of the automated system and the irreplaceable value of human judgment in handling extreme, unknown situations. This reflects the fact that, in the evolution toward "model is law," human common sense and judgment still play an important role.

31.10 From code is law to model is law

As AI systems gradually replace the hardcoded rules in DeFi infrastructure, the operating logic of market microstructure undergoes a profound paradigm shift. This transformation changes the system's core decision mechanism from deterministic code to a probabilistic model, bringing both an enormous opportunity for adaptive optimization and new governance challenges and systemic risks. The early-blockchain ideal of "code is law" was once seen as the ultimate governance solution, but as market complexity has increased and extreme events have occurred frequently, the limitations of this paradigm have become ever more apparent. At the same time, the powerful capability that AI models display in handling complex decisions makes "model is law" a possible alternative. Both paradigms, however, have fundamental flaws, and the true breakthrough lies in how to deeply integrate the two, forming a hybrid architecture that both retains the safety floor of code's determinism and fully harnesses the model's adaptive-optimization capability.

31.10.1 The strengths and limitations of code is law

The ideal of "code is law" forms the cornerstone of the first generation of DeFi infrastructure. Under this paradigm, all business logic, risk parameters, and execution conditions are turned into immutable smart-contract code. The achievement of this architecture is manifested mainly in the absolute determinism and verifiability it provides. Once a smart contract is deployed, its execution path depends entirely on the input state and the predefined logical branches, with no ambiguity or room for human discretion. This determinism ensures that the system runs as expected in all situations, thereby establishing a high degree of execution reliability in a trustless environment. Formal verification technology further strengthens this advantage, using mathematical proof to ensure that a smart contract's logic conforms to its expected properties and thereby eliminating conventional security vulnerabilities at the code level [35]. In addition, the complete transparency of the code enables any participant to audit the system's operating mechanism, ensuring that the rules apply fairly to all. This censorship resistance is a characteristic that traditional financial systems cannot achieve; any attempt to change the rules must go through an explicit on-chain governance process and cannot be carried out through covert back-end operations.

The determinism of code also brings another key advantage: the system's behavior is fully predictable. Participants can precisely calculate how the system will react in any given market state, which makes risk assessment relatively intuitive. In a fixed-rate lending protocol, for example, participants know that the interest-calculation formula is entirely deterministic and will not change because of some centralized entity's subjective judgment. This predictability is a foundational condition for building market confidence, and is especially prominent in transactions involving large sums.

The rigidity of hardcoded rules, however, exposes significant limitations when facing complex and variable market microstructure. Deterministic code cannot adapt to unforeseen extreme market situations, and when facing complex scenarios such as an instantaneous liquidity drought or an abrupt shift in correlation structure, preset linear rules often prove too simple and too slow. A dynamic margin rate computed by a fixed formula, for example, cannot capture the nonlinear interactions among multidimensional market variables, causing the system to be possibly overly conservative under normal conditions yet inadequately prepared for extreme tail risk. The 2016 Ethereum DAO incident offers an important case. The smart-contract code of the time was logically "correct," but when facing the unforeseen attack vector of a reentrancy attack, it could not adaptively adjust its defensive strategy at all. The determinism of code proved a limitation in such a scenario, because once the vulnerability was discovered, the system could not respond to the threat through dynamic adjustment and could only passively wait for human intervention.

In addition, the static nature of parameter settings leaves the system without real-time adaptive capability. In a traditional governance architecture, adjusting risk parameters depends on a human proposal-and-voting process, which is not only time-consuming but also hard-pressed to keep up with the millisecond-level tempo of market changes. When market volatility doubles within seconds, adjusting risk parameters through a governance vote is entirely unrealistic. A deeper problem is that code logic is inher-

ently constrained by Boolean conditional judgments and cannot handle complex semantic conditions or continuous probabilistic assessments. When it is necessary to judge the toxicity of an order or predict the liquidation probability of a position, a simple threshold-trigger mechanism proves inadequate. Code can only say "if the margin rate falls below 50%, then liquidate," but it cannot say "based on the current market liquidity, the depth of the liquidation queue, and cross-market correlation, this position's liquidation probability is 73.2%, and therefore the trader should be proactively alerted." This simplified handling not only reduces the system's capital efficiency but also limits the resilience of the market infrastructure when responding to complex attacks and anomalous volatility.

Another implicit limitation of code lies in its dependence on parameter tuning. When designing a smart contract, developers must preset all parameter values, such as liquidation thresholds, risk weights, and fee ratios. The choice of these parameters is often based on historical data and the designer's subjective judgment, but the market environment is continually evolving. Once parameters are hardcoded, changing them requires a complex upgrade process, which in a decentralized system often means obtaining community consensus. This leaves the system either with parameter settings too conservative to cope with the worst case, or too aggressive and prone to collapse under extreme conditions.

31.10.2 The potential and risks of model is law

When AI models are introduced as the core decision engine of market infrastructure, the nature of the system's rules evolves from deterministic code into a probabilistic model. Under the "model is law" paradigm, the system's behavior is no longer determined by fixed conditional statements but is driven by the model's real-time assessment of multidimensional market states. This shift endows the infrastructure with the ability to handle complex semantics and make continuous judgments. AI models can simultaneously process thousands of input variables, discover nonlinear patterns that human designers cannot foresee, and dynamically adjust risk parameters according to real-time liquidity conditions, liquidation-queue density, and cross-market spillover effects.

This adaptive capability significantly improves the system's capital efficiency and resilience. Through continuous learning and optimization, the model can achieve more precise risk pricing than a hardcoded formula, reducing the capital sequestration caused by over-conservatism and taking predictive preventive measures before risk actually materializes. In a predictive liquidation system, for example, the model no longer waits for the margin to fall below a threshold to trigger liquidation, but, based on a continuous assessment of the liquidation probability, guides the gentle reduction of a position at an early stage, thereby effectively preventing the systemic risk of a liquidation cascade. The model can learn that a particular type of market condition often precedes a price collapse and can therefore act in advance. In addition, the model can handle complex semantic conditions, such as the multidimensional combined condition "when the market is in a panic state and liquidity providers are withdrawing," rather than merely a simple single threshold such as "price below X." This enables the system to make fine-grained responses to subtle changes in market microstructure.

The model's capacity for continuous optimization is also a major advantage over hard-coded rules. As new market data continually accumulate, the model can automatically adjust its parameters and strategy without human intervention. This means the system can gradually adapt to the evolution of the market environment rather than being locked into past parameter settings. In the field of high-frequency trading, this real-time learning capability has been shown to significantly improve trading performance and risk-management outcomes.

A probabilistic decision mechanism, however, inevitably introduces new dimensions of risk. The most notable flaw of AI models is their unpredictability and lack of interpretability. When a deep neural network decides to raise the margin rate for a certain class of assets, it often cannot provide a humanly understandable causal explanation. This "black box" nature directly challenges a decentralized system's basic requirements for transparency and auditability. If participants cannot understand the logic behind a system's decisions, the foundation of trust will be severely weakened [53]. In traditional finance, regulators can require a bank to explain its risk decisions, but for a deep-learning model such a requirement often cannot be met.

In addition, when facing a "black swan" event that never appeared in the training data, the model's behavior often exhibits extreme fragility. The model may make wholly unreasonable decisions in a brand-new market environment, leading to systematic mispricing or improper liquidation. The 2012 Knight Capital algorithmic-runaway incident illustrates this risk. Because of an incomplete deployment of new trading code, the firm triggered an obsolete legacy function that had already been retired, and the trading system placed orders out of control for approximately 45 minutes, producing approximately \$440 million in realized losses (with the final related pre-tax loss, including subsequent costs, at approximately \$461 million). This incident shows that even a tested algorithmic system can run out of control because of a deployment defect or an unanticipated operating state.

Adversarial fragility is another key source of risk. Machine-learning models are prone to carefully crafted adversarial-example attacks, and a malicious participant can deceive the model by creating particular market signals, inducing it to make parameter adjustments favorable to the attacker [54]. An attacker might, for example, "train" the model to produce a particular bias through a series of carefully designed order flows, so as to gain an advantage at a critical moment. This new form of "model attack" is more covert and harder to defend against than a traditional code-vulnerability exploit, because it does not require finding a logical error in the code but instead exploits the model's own learning mechanism.

Finally, because the model is continually updated during ongoing learning, its behavioral characteristics undergo "model drift" over time, which means the system's rules today may differ from its rules tomorrow, thereby undermining participants' expectations of the system's stability [27]. This is a serious problem for traders who need to plan over the long run. In addition, the model's training process may itself introduce bias. If the training data are insufficiently representative, or if there are flaws in feature engineering, the model may learn spurious correlations, leading to systematic decision errors. These risks make "model is law" infeasible as an independent

governance paradigm.

31.10.3 The hybrid architecture

To resolve the inherent contradiction between determinism and adaptivity, the next generation of market infrastructure must adopt a "code plus model" hybrid architecture. The design principle of this architecture is not to make an either-or trade-off between "code is law" and "model is law," but to integrate the two through deep nesting. The core insight of the hybrid architecture is this: hardcoded code defines the system's inviolable safety floor, and within that floor, the AI model is granted ample freedom to optimize for efficiency.

Figure 31-10 shows the nested design logic of the hybrid architecture: the outer layer, composed of hardcoded smart contracts, constitutes an absolute safety boundary; the inner layer, composed of the AI model, performs adaptive optimization within the operating space bounded by the safety floor; and the two layers are connected through a "model proposes, code reviews" verification mechanism.

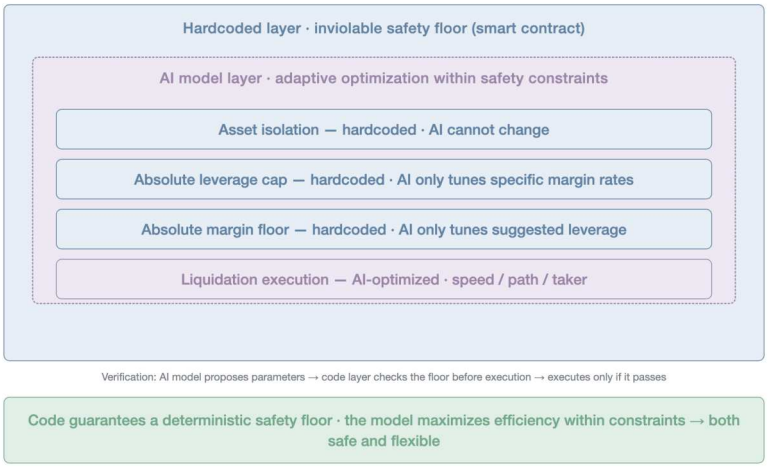


Figure 31-10. The design of the "code plus model" hybrid architecture (conceptual illustration; based on the author's theoretical derivation of a governance framework for decentralized infrastructure)

In the outer layer of the hybrid architecture, the system sets absolute safety boundaries through hardcoded smart contracts. These floor rules include an asset-isolation mechanism, the maximum permitted leverage multiple, the minimum maintenance-margin rate, and explicit liquidation-trigger conditions. The execution of these rules has absolute priority and is not affected by the output of any AI model. No matter how optimistic the market state assessed by the internal model, or how aggressive its recommended parameters, the hardcoded layer will resolutely reject any operation that breaches the safety floor. This design preserves the core value of "code is law" in guarding against extreme tail risk, ensuring the system's survivability in the worst case.

The design principles of the floor warrant special attention. An effective floor must

satisfy several key conditions: first, it must be absolute and non-negotiable, and cannot be breached for any reason. Second, it must be verifiable, so that any participant can independently verify whether the system has complied with the floor. Third, the floor must be loose enough to allow the model to perform meaningful optimization within it, rather than so strict that the model cannot function. Fourth, the floor must be resistant to circumvention. For example, if the floor stipulates a maximum leverage of 10×, but the model can circumvent this limit through a combination of multiple accounts, then the floor is ineffective. The hybrid architecture must therefore include additional checking mechanisms to ensure that the model cannot circumvent the floor through creative means.

Within the operating space bounded by the safety floor, the inner layer of the hybrid architecture is led by the AI model. The model is authorized to perform high-frequency adaptive optimization, including dynamically adjusting the specific margin rate, setting personalized risk parameters, executing predictive interventions, and optimizing the speed and the counterparty of the liquidation process. For example, when market volatility rises but has not yet triggered the hardcoded liquidation threshold, the model can proactively raise the margin requirement for particular high-risk positions, or advise the trader to close out early. This inner-layer optimization significantly improves the system's speed of response to changes in market microstructure, enabling the infrastructure to maximize capital efficiency under normal conditions while smoothly defusing risk in the early stage of its accumulation.

The mechanism by which code verifies the model's output in the hybrid architecture is the key link ensuring the system's safety. The model does not directly control the execution engine but instead generates parameter-adjustment recommendations or operation instructions. Before these recommendations take effect, they must undergo strict review by the hardcoded logic. The code layer acts as a strict verification mechanism; it does not assess the economic reasonableness of the model's recommendation but only verifies whether the recommendation violates any preset safety floor. The specific workflow is as follows: first, the model analyzes the current market state and system state and generates a parameter-adjustment recommendation, such as "raise account A's margin rate from 50% to 55%." Second, this recommendation is sent to the code-verification layer. The verification layer performs a series of checks: whether the new margin rate is still above the hardcoded floor of the minimum maintenance-margin rate, whether account A's total leverage is still below the hardcoded floor of the maximum leverage, and whether any other operation prohibited by hardcoded rules is present. If all checks pass, the instruction is executed; if any breach of the floor is found, the code directly rejects the recommendation and may trigger a downgrade of the system to a more conservative fallback state. This "model proposes, code reviews" workflow successfully harnesses to the greatest extent AI's advantages in handling complex semantics and multidimensional optimization, while ensuring the system's absolute safety.

The design advantage of this architecture is that it achieves a mode combining "limited trust with independent verification." The system trusts the model's optimization capability but strictly verifies its output. This avoids both the rigidity that comes from relying entirely on code and the uncontrollability that comes from relying entirely on the model.

31.10.4 Model auditing

Shifting from auditing deterministic code to auditing a probabilistic model constitutes the most severe governance challenge under the hybrid architecture. Traditional code auditing focuses on the correctness of the logic and the completeness of the execution paths, with the goal of finding potential vulnerabilities and ensuring that the code's behavior conforms to expected specifications. Although this auditing process is complex, because of the deterministic nature of code, its results are verifiable and final [55]. Code auditing can use mathematical tools such as formal verification to ensure the code's correctness through symbolic execution or theorem proving. An auditor can trace every execution path of the code and verify that, under all possible inputs, the code produces the expected result.

The complexity of auditing an AI model, however, increases exponentially. Model auditing requires not only assessing the reasonableness of the algorithmic architecture but also examining in depth the quality of the training data, the bias in feature engineering, and the model's performance in historical decisions. A deep neural network may contain millions of parameters, and a minute change in any one of them may cause a significant change in the output. This makes the traditional code-auditing method entirely inapplicable. An auditor cannot understand the model's decision logic by reading the code line by line, because the model's decisions are produced by complex mathematical operations and nonlinear transformations.

Because of the inherent "black box" nature of deep-learning models, requiring the model to provide a precise causal explanation for each of its microlevel decisions is often technically unrealistic. To meet this challenge, the governance framework must introduce new practical schemes to ensure that the model's behavior is traceable and accountable. One feasible mechanism is to require the AI system to periodically generate and publish standardized "decision reports." These reports should record in detail the key decisions the model made within a particular period, clarify the main input features on which the decisions relied and their weights, and contrast the model's decisions with the expected output of purely hardcoded rules under the same conditions [56].

The specific design of the decision report should include the following. First, the report should record all important decisions above a certain threshold, such as all decisions that led to a margin-rate adjustment of more than 5%. Second, for each decision, the report should list the top ten most influential input features and their contribution to the decision. Third, the report should include a comparative analysis of the model's decision with a pure hardcoded rule. For example, if a hardcoded rule would liquidate a position under the given market conditions but the model decided not to liquidate, the report should explain the model's reasoning. Fourth, the report should include the model's confidence indicator—that is, how certain the model is of its decision. If the confidence is low, this may indicate that the model is handling an unfamiliar market environment.

These decision reports should be stored on the blockchain, forming an immutable audit trail. In this way, anyone can look back through history to see the decisions the model made at a particular point in time and their reasoning. Through this approach, even if it is impossible to fully reveal the model's internal nonlinear mapping process, governance

participants can still assess the reasonableness of the model's behavior and identify potential systematic biases by analyzing the statistical relationship between inputs and outputs. If the model is found to systematically make unreasonable decisions under a particular type of market condition, this can trigger a retraining of the model or an adjustment of the floor.

The probabilistic nature of model decisions also gives rise to complex questions about the accountability mechanism. When the model's adaptive adjustments cause some participants to suffer unexpected losses, the traditional liability-attribution framework based on deterministic rules no longer applies. In the case of hardcoded rules, if the system causes a loss, the liability is clear: either the rule design was flawed, or the rule execution was faulty. But for the model, the situation is far more complex. The model's decisions are based on the patterns it has learned, and these patterns may themselves be flawed, but the flaw may be due to insufficient training data, improper feature engineering, or an improper choice of model architecture.

To ease this contradiction, the concept of "model insurance" provides a market-based mechanism for risk transfer. A protocol operator can establish a dedicated insurance fund for the decision behavior of the AI infrastructure; when the model's output causes an anomalous loss exceeding what the hardcoded floor anticipates, the injured party can obtain compensation from the fund. The design of this mechanism requires special attention to several aspects. First, the scope of insurance must be clearly defined. What types of loss should the insurance cover? All losses caused by model decisions, or only those that breach particular conditions? Second, the pricing method must reflect the model's actual risk. If the model's historical performance shows that it is prone to error under particular market conditions, the insurance premium should rise accordingly. Third, the claim-trigger condition must be objective and verifiable. A claim cannot be based on subjective judgment but must be based on explicit conditions verifiable on-chain. For example, a claim might be triggered when the loss caused by the model's decision exceeds a certain threshold and this loss is due to the model's decision rather than to market volatility.

This mechanism turns the long-tail risk brought by the model's unpredictability into a quantifiable and priceable insurance cost, thereby strengthening participants' trust in the hybrid architecture. Participants know that even if the model makes an unreasonable decision, they have insurance as a last line of protection. At the same time, the level of the insurance premium also provides a market signal of the model's quality. If a model's insurance premium is high, this indicates that the market considers the model to be very risky.

Finally, the system must have a sound "revertibility" design as the ultimate safety guarantee. When facing anomalous model performance, a severe adversarial attack, or a fundamental structural rupture in the market environment, the system should be able to cut off the AI model's control automatically or through an emergency governance process and seamlessly revert to a base state driven purely by hardcoded rules. This design establishes code's status as an always-available safety backup and sends a clear signal to the market: the model is an enhancing component that improves efficiency, but it is not the sole dependency for keeping the system running. The description of

"seamless reversion" above, however, must confront the practical challenges of engineering implementation. When the AI model has dynamically lowered the margin rate of a certain class of assets to a relatively low level during operation to optimize capital efficiency, while the hardcoded default is set relatively high, the parameter jump at the instant of reversion may itself trigger large-scale margin calls or liquidation events. The actual reversion mechanism should therefore be designed as a gradual downgrade rather than an instantaneous switch: after detecting an anomaly, the system first limits the model's adjustment magnitude, then gradually returns the parameters to the hardcoded defaults. The reversion process must also handle the race condition in which the AI system's uncompleted operations and the hardcoded system run simultaneously, ensuring that no double execution or missed execution occurs.

The engineering implementation of revertibility must consider several aspects. First, the system must be able to detect the model's anomalous behavior. This can be achieved by monitoring the degree of deviation between the model's decisions and its historical patterns. If the model makes a large number of decisions inconsistent with its historical behavior within a short time, this may indicate that the model has been attacked or has undergone severe drift. Second, the system must be able to trigger the reversion quickly. This cannot depend on human intervention but must be automatic. Once an anomaly is detected, the system should immediately stop using the model's output and switch to hardcoded rules. Third, the system must be able to recover after reversion. This means that during the reversion period, the system must still be able to operate normally, though its efficiency may decline. Fourth, the system must be able to diagnose the root cause of the problem. After reversion, an in-depth analysis is needed to determine why the model malfunctioned—whether because of an adversarial attack, training-data contamination, model drift, or another reason.

Through these multi-level auditing, accountability, and reversion mechanisms, the hybrid architecture maintains the safety floor of DeFi infrastructure while embracing the efficiency gains that AI brings. This architecture is not a perfect solution, but it represents the best practice, under current technological conditions, for balancing security and efficiency. As AI technology develops further and interpretability methods improve, this hybrid architecture is expected to keep evolving, gradually raising the model's trustworthiness and autonomy.

31.11 Chapter summary

This chapter has advanced three core theoretical contributions. First, the four-stage model of AI penetration into microstructure (assistance → execution → autonomy → adversarial competition) provides a systematic framework for understanding the evolution of AI's role in the market. The current market is in the critical transition from Stage 2 to Stage 3, in which market making and arbitrage, because of their heavy dependence on speed and multidimensional optimization capability, have become the fastest-penetrating domain, while governance participation remains at the proof-of-concept stage. Second, the dual-role framework reveals the economic logic of AI's evolution from a value-extracting market participant into value-creating market infrastructure: participant AI's profit margin is continually compressed under competitive

pressure, whereas infrastructure AI sustains a persistent capacity for value capture on the strength of network effects, economies of scale, switching costs, and the data-feedback effect. This framework explains why capital and technological resources are accelerating their migration toward the infrastructure domain. Third, the "code plus model" hybrid architecture provides a design paradigm for the next generation of market infrastructure: hardcoded rules set an inviolable safety floor, within which the AI model is granted ample space for adaptive optimization, and determinism and flexibility are unified through a "model proposes, code reviews" workflow.

This chapter has identified three key findings and risks. Strategy homogenization is the most prominent new form of systemic fragility in the AI era: algorithms trained on similar architectures and data may, when facing an anomalous signal, make wholly identical defensive reactions within milliseconds, triggering an "AI flash crash" of extremely short duration but extremely great destructive power. This mechanism turns the volatility distribution from mildly right-skewed into bimodal—that is, normal-regime volatility declines but tail risk rises—and market resilience exhibits a bimodal character of "extremely high in the normal regime, extremely low in the tail." At the same time, AI's effect on market quality is highly nonlinear: transaction cost and the efficiency of price discovery improve significantly under normal conditions, but fairness deteriorates substantively because of the widening speed gap. At the infrastructure level, the six core AI-driven systems (the dynamic risk engine, the intelligent matching engine, the predictive liquidation system, the dynamic pricing mechanism, the compliance-monitoring system, and liquidity-routing optimization) are upgrading the static protocol architecture into an adaptive system, among which the predictive liquidation system, by moving risk management forward from passive triggering to probabilistic prediction, can significantly reduce the scale of liquidation cascades under extreme conditions. The breakthrough of the semantic oracle, in turn, fundamentally unlocks the technical bottleneck of "perpetual futures on everything," enabling unstructured information such as political events, corporate performance, and arbitrary natural-language conditions to be brought into the on-chain financial system.

The analysis in this chapter also reveals several open questions. How should the boundary of AI agents' participation in decentralized governance be delineated, and should the permissions for parameter optimization be distinguished from those for constitutional decisions in a tiered manner? How can an audit and accountability framework for probabilistic models be established, and can a "model insurance" mechanism effectively transfer the long-tail risk that unpredictability brings? Can the risk of strategy homogenization be effectively mitigated through a "heterogeneity incentive" mechanism, or will it keep intensifying as AI technology converges further? Can the hallucination and bias problems of the semantic oracle be adequately controlled through multi-node consensus and adversarial-verification mechanisms? The answers to these questions will determine whether an AI-driven autonomous financial network can achieve a sustainable balance among efficiency, security, and fairness.

References

[1] Trestman, M. (2013). The Cambrian explosion and the origins of embodied cognition.

Biological Theory, *8*(1), 80-92.

[2] Plotnick, R. E., Dornbos, S. Q., & Chen, J. (2010). Information landscapes and sensory ecology of the Cambrian Radiation. *Paleobiology*, *36*(2), 303-317.

[3] Hirshleifer, D. (2015). Behavioral finance. *Annual Review of Financial Economics*, *7*(1), 133-159.

[4] O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, *116*(2), 257-270.

[5] Aquilina, M., Budish, E., & O'Neill, P. (2022). Quantifying the high-frequency trading "arms race". *The Quarterly Journal of Economics*, *137*(1), 493-564.

[6] Dou, W. W., Goldstein, I., & Ji, Y. (2025). AI-powered trading, algorithmic collusion, and price efficiency. *NBER Working Paper 34054*. https://doi.org/10.3386/w34054

[7] Liu, Q., Miao, D., Wang, C., & Wang, C. D. (2025). Strategic homogeneity in trading: Amplifying market volatility and the emergence of mini-crashes. *SSRN Electronic Journal*, 5239024. https://doi.org/10.2139/ssrn.5239024

[8] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71-100. https://doi.org/10.1016/0304-405x(85)90044-3

[9] Cont, R., & Xiong, W. (2024). Dynamics of market making algorithms in dealer markets: Learning and tacit collusion. *Mathematical Finance*. https://doi.org/10.1111/mafi.12401

[10] Zhao, M., & Linetsky, V. (2021). High frequency automated market making algorithms with adverse selection risk control via reinforcement learning. *Proceedings of the Second ACM International Conference on AI in Finance*, 1-8.

[11] Herrmann, S., Muhle-Karbe, J., Shang, D., & Yang, C. (2020). Inventory management for high-frequency trading with imperfect competition. *SIAM Journal on Financial Mathematics*, *11*(1), 1-45. https://doi.org/10.1137/18m1207776

[12] Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217-224. https://doi.org/10.1080/14697680701381228

[13] Ken Research. (2024). *Global high frequency trading market report*. https://www.kenresearch.com/global-high-frequency-trading-market

[14] Hofweber, T., Bergl, J., Reyes, I., & Sadovnik, A. (2025). The Black Tuesday Attack: how to crash the stock market with adversarial examples to financial forecasting models. *arXiv preprint arXiv:2510.18990*. https://doi.org/10.48550/arXiv.2510.18990

[15] Raboun, A., Briere, M., & Lehalle, C.-A. (2021). Liquidity provision and market-making in different uncertainty regimes: Evidence from the COVID-19 market crash. *Available at SSRN 3815169*. https://doi.org/10.2139/ssrn.3815169

[16] Gao, K., Vytelingum, P., Weston, S., Luk, W., & Guo, C. (2022). High-frequency financial market simulation and flash crash scenarios analysis: An agent-based modelling approach. *Journal of Artificial Societies and Social Simulation*, *27*(2), 8.

[17] Bookstaber, R., Foley, M. D., & Tivnan, B. F. (2016). Toward an understanding

of market resilience: Market liquidity and heterogeneity in the investor decision cycle. *Journal of Economic Interaction and Coordination*, *11*(1), 1-27.

[18] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201-2238. https://doi.org/10.1093/rfs/hhn098

[19] Goldblum, M., Schwarzschild, A., Patel, A., et al. (2021). Adversarial attacks on machine learning systems for high-frequency trading. In *Proceedings of the Second ACM International Conference on AI in Finance* (pp. 1-9). https://doi.org/10.1145/3490354.3494367

[20] Fernández Vicente, Ó. (2025). Market making strategies with reinforcement learning. *arXiv preprint arXiv:2507.18680*. https://doi.org/10.48550/arXiv.2507.18680

[21] Dahlhaus, R., & Neddermeyer, J. C. (2014). Online spot volatility-estimation and decomposition with nonlinear market microstructure noise models. *Journal of Financial Econometrics*, *12*(1), 174-212.

[22] Commodity Futures Trading Commission. (2014). The flash crash: The impact of high frequency trading on an electronic market. *Office of the Chief Economist*.

[23] Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, *72*(3), 967-998.

[24] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307-327. https://doi.org/10.1016/0304-4076(86)90063-1

[25] International Monetary Fund. (2024). *Global Financial Stability Report: Advances in artificial intelligence: Implications for capital markets*. October 2024.

[26] Abbas, N., Cohen, C., et al. (2024). Artificial intelligence can make markets more efficient—and more volatile. *IMF Blogs*, October 15, 2024.

[27] Zenisek, J., Holzinger, F., & Affenzeller, M. (2019). Machine learning based concept drift detection for predictive maintenance. *Computers & Industrial Engineering*, *137*, 106031.

[28] Kearns, M., & Nevmyvaka, Y. (2013). Machine learning for market microstructure and high frequency trading. In *Algorithmic trading: New realities for traders, markets and regulators*.

[29] Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *Review of Financial Studies*, *27*(8), 2267-2306. https://doi.org/10.1093/rfs/hhu032

[30] Chi, T., He, N., Hu, X., & Wang, H. (2024). Remeasuring the arbitrage and sandwich attacks of maximal extractable value in Ethereum. *arXiv preprint*, arXiv:2405.17944.

[31] Sun, S., Qin, M., Zhang, W., Xia, H., Zong, C., & Duan, J. (2023). TradeMaster: A holistic quantitative trading platform empowered by reinforcement learning. *Advances in Neural Information Processing Systems*, *36*, 1-15.

[32] U.S. Securities and Exchange Commission. (2014). SEC Charges New York-Based

High Frequency Trading Firm With Fraudulent Trading to Manipulate Closing Prices. *Press Release 2014-229*.

[33] European Securities and Markets Authority. (2026). *Supervisory briefing on algorithmic trading in the EU* (ESMA74-1505669079-10311).

[34] Kurshan, E., Balch, T., & Byrd, D. (2025). The Agentic Regulator: Risks for AI in Finance and a Proposed Agent-based Framework for Governance. *arXiv preprint arXiv:2512.11933*. https://doi.org/10.48550/arXiv.2512.11933

[35] Bhargavan, K., Delignat-Lavaud, A., Fournet, C., Gollamudi, A., Gonthier, G., Kobeissi, N., … & Zanella-Béguelin, S. (2016). Formal verification of smart contracts: Short paper. In *Proceedings of the 2016 ACM Workshop on Programming Languages and Analysis for Security* (pp. 91-96).

[36] Han, J., Lee, J., & Li, T. (2025). A review of DAO governance: Recent literature and emerging trends. *Journal of Corporate Finance*, *90*, 102685. https://doi.org/10.2139/ssrn.5074046

[37] Tamai, S., & Kasahara, S. (2024). DAO voting mechanism resistant to whale and collusion problems. *Frontiers in Blockchain*, *7*, 1405516.

[38] Stigler, G. J. (1957). Perfect competition, historically contemplated. *Journal of Political Economy*, *65*(1), 1-17. https://doi.org/10.1086/257878

[39] Yadav, Y. (2015). How algorithmic trading undermines efficiency in capital markets. *Vanderbilt Law Review*, *68*(1), 160-228.

[40] International Monetary Fund. (2024). *Artificial intelligence and its impact on financial markets and financial stability*. https://www.imf.org/en/news/articles/2024/09/06/sp090624-artificial-intelligence-and-its-impact-on-financial-markets-and-financial-stability

[41] Huang, Y. (2025). Deep learning-enhanced dynamic margin period of risk prediction for counterparty credit risk management: A multi-modal approach integrating market sentiment. *Proceedings of the 2nd International Conference on Computer Science and Management*, 1-8.

[42] Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, *130*(4), 1547-1621. https://doi.org/10.1093/qje/qjv027

[43] Cartea, Á., Duran-Martin, G., & Sánchez-Betancourt, L. (2023). Detecting toxic flow. *arXiv preprint arXiv:2312.05827* (revised 2026). https://doi.org/10.48550/arXiv.2312.05827

[44] Palaiokrassas, G., Scherrers, S., & Makri, E. (2024). Machine learning in DeFi: Credit risk assessment and liquidation prediction. *IEEE International Conference on Blockchain and Cryptocurrency*, 1-8.

[45] Popoola, N. T. (2023). Big data-driven financial fraud detection and anomaly detection systems for regulatory compliance and market stability. *Journal of Computer Applications Technology and Research*, *1*(1), 1-15.

[46] Tiwari, S., Ramampiaro, H., & Langseth, H. (2021). Machine learning in financial

market surveillance: A survey. *IEEE Access*, *9*, 156321-156338. https://doi.org/10.1109/access.2021.3130843

[47] Zhang, Y., & Tessone, C. J. (2025). Measuring DEX efficiency and the effect of an enhanced routing method on both DEX efficiency and stakeholders' benefits. *arXiv preprint arXiv:2508.03217*. https://doi.org/10.48550/arXiv.2508.03217

[48] Chainlink. (2024). *The Intersection Between AI Models and Oracles*. https://blog.chain.link/oracle-networks-ai/

[49] Xian, Y., Zeng, X., Xuan, D., Yang, D., Pei, C., Fan, P., & Liu, P. (2024). *Connecting Large Language Models with Blockchain: Advancing the Evolution of Smart Contracts from Automation to Intelligence*. arXiv preprint arXiv:2412.02263. https://doi.org/10.48550/arXiv.2412.02263

[50] Luo, H., Sun, G., Liu, Y., Zhao, D., Niyato, D., & … (2025). *A weighted byzantine fault tolerance consensus driven trusted multiple large language models network*. IEEE Transactions. https://doi.org/10.1109/tccn.2025.3620286

[51] Yao, J. Y., Ning, K. P., Liu, Z. H., Ning, M. N., Liu, Y. Y., & … (2023). *LLM lies: Hallucinations are not bugs, but features as adversarial examples*. arXiv preprint arXiv:2310.01469. https://doi.org/10.48550/arXiv.2310.01469

[52] Amiri-Margavi, A., Jebellat, I., Jebellat, E., & … (2025). *Enhancing answer reliability through inter-model consensus of large language models*. Conference on Artificial Intelligence. https://doi.org/10.1007/978-3-031-96235-6_22

[53] CFA Institute. (2025). *Explainable AI in Finance*. Research & Policy Center. https://rpc.cfainstitute.org/research/reports/2025/explainable-ai-in-finance

[54] Silva, S. H., & Najafirad, P. (2020). Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*. https://doi.org/10.48550/arXiv.2007.00753

[55] Almakhour, M., Sliman, L., Samhat, A. E., & Mellouk, N. (2020). Verification of smart contracts: A survey. *Pervasive and Mobile Computing*, *67*, 101227.

[56] Ganapathy, V. (2025). A comparative study of explainable artificial intelligence (XAI) techniques in financial auditing applications. *Edumania: An International Multidisciplinary Journal*, *3*(1), 1-15. https://doi.org/10.59231/edumania/9147

The Global Vision of Autonomous World Finance

According to a report published by CoinGecko in February 2026, the world's top 10 centralized exchanges (CEXs) and decentralized exchanges (DEXs) processed $92.9 trillion in notional (leveraged) perpetual futures trading volume across full-year 2025, a 64.6% increase over the prior year; decentralized perpetual futures exchanges reached an all-time high of $6.7 trillion in trading volume, up 346% year over year (this figure is notional volume, and because the source does not disclose its wash-trading filtering method, it should be read as an order-of-magnitude scale rather than as precise net demand) [1]. In stark contrast, the 2025 edition of the World Bank's Global Findex Database reports that roughly 1.3 billion adults worldwide remain unbanked [2][3], and the average cost of cross-border remittances has long held at around 6% (recent World Bank data put it at approximately 6.2% to 6.65%, still more than twice the Sustainable Development Goals' 3% target) [4]. The contrast between these two sets of figures reveals the central contradiction confronting autonomous finance: on-chain financial infrastructure has demonstrated the engineering capacity to substitute for traditional intermediaries, yet the scale of global financial exclusion shows that this capacity has not translated into universally accessible services. The central question of this chapter is whether the three-layer stack composed of perpetual futures, on-chain clearing and settlement, and artificial intelligence (AI) agents can address the genuine needs of financial inclusion while improving the efficiency of global capital allocation, and how its inherent tensions delimit the boundaries of this process.

The architectural logic of the three-layer stack is the starting point of the analysis. The asset layer converts underlying assets into on-chain perpetual futures; the infrastructure layer provides decentralized clearing and settlement together with embedded compliance; and the intelligence layer, through AI agents, sharply compresses the marginal cost of professional financial services. The synergy among the three layers within a permissionless network far exceeds the simple sum of each layer's functions (elaborated in Section 32.1), and its value can be understood only through systematic analysis.

On this basis, the discussion of financial inclusion shifts from an abstract vision to a concrete analysis of engineering feasibility. The core pain points faced by the roughly 1.7 billion to 4 billion people in emerging markets who lack basic financial services (a broader measure than the unbanked) [5][6]—namely foreign-exchange hedging, inflation-resistant savings, and low-cost cross-border payments—correspond directly to the en-

gineering properties of the three-layer stack in lowering friction costs and barriers to entry. Yet the real constraints on inclusion cannot be ignored. A lack of financial literacy means that low barriers may expose users without risk-management capacity to highly leveraged markets, and a large share of the world's population still has no internet access (approximately 2.2 billion people are entirely offline, and more than 3 billion others fall within mobile-network coverage yet do not use mobile internet—the so-called *usage gap* [7][8]; the full accounting is detailed in Section 32.2.5), so the digital divide constitutes a more fundamental barrier to access than financial exclusion itself. As the three-layer stack sharply compresses friction costs, the efficiency of global capital allocation also changes structurally: around-the-clock continuous pricing eliminates the opening gaps caused by information backlogs; synthetic exposure decouples the economic allocation of capital from the cross-border transfer of physical ownership; and cross-margining raises capital-utilization efficiency to a level unattainable under the traditional isolated-account system. But a highly interconnected network simultaneously breeds new forms of systemic risk, including liquidation cascades under global cross-margining, global flash crashes triggered by the homogenization of AI-agent strategies, and the system-wide propagation of an oracle single point of failure. These risks propagate on the order of milliseconds, compressing the window for human intervention to the extreme.

This analysis ultimately converges on three irreducible structural tensions. The trade-off between efficiency and stability is amplified to a global scale by the introduction of leverage and continuous trading; a fundamental conflict exists between permissionless access and the demand for consumer protection; and an enduring contest arises between a borderless liquidity network and the control that sovereign states exercise over monetary policy and capital flows. These tensions are not engineering defects awaiting repair but fundamental trade-offs that any financial system must make among competing value objectives. After reading this chapter, readers will be able to systematically understand the architectural logic and synergies of the three-layer stack, assess its real potential and constraints in financial inclusion and global capital allocation, and recognize that the mark of maturity for autonomous finance lies not in eliminating its inherent contradictions but in establishing governance mechanisms that seek a dynamic equilibrium within these long-term tensions.

32.1 The three-layer stack of autonomous finance

As defined in this chapter, *autonomous finance* refers to a paradigm of financial services whose defining feature is the autonomous decision-making capacity of AI agents and which runs on permissionless on-chain infrastructure. This concept has clear boundaries with adjacent terms: DeFi emphasizes the decentralization of governance structures, open finance emphasizes the openness of interfaces and data, and programmable finance emphasizes the ability of smart contracts to encode financial logic, whereas autonomous finance is distinctively positioned by the capacity of AI agents to complete the full closed loop of risk assessment, liquidity provision, and strategy execution without continuous human intervention.

The architecture of an autonomous finance network can be decomposed into three func-

tional layers: the asset layer, the infrastructure layer, and the intelligence layer. This three-layer analytical framework corresponds to the BIS Committee on Payments and Market Infrastructures (CPMI) layering of financial market infrastructures and to the DeFi composability stack proposed by Schär (2021) [9], but the distinctive contribution of this chapter's framework is to incorporate the AI intelligence layer as an independent functional layer, reflecting the fundamental impact of autonomous agents on the structure of the financial system. Each layer has made varying degrees of technical progress, but their value lies not in their individual functions but in the emergent effects produced when the three layers operate in concert within a single permissionless network. This section analyzes in turn the current maturity and core challenges of each layer, and on that basis argues why the synergy of the three-layer stack far exceeds the simple sum of each layer's functions.

The three functional layers are not simply stacked one atop another; rather, they form tight interdependencies along three dimensions—data flow, the transmission of trust, and the supply of liquidity (Figure 32-1).

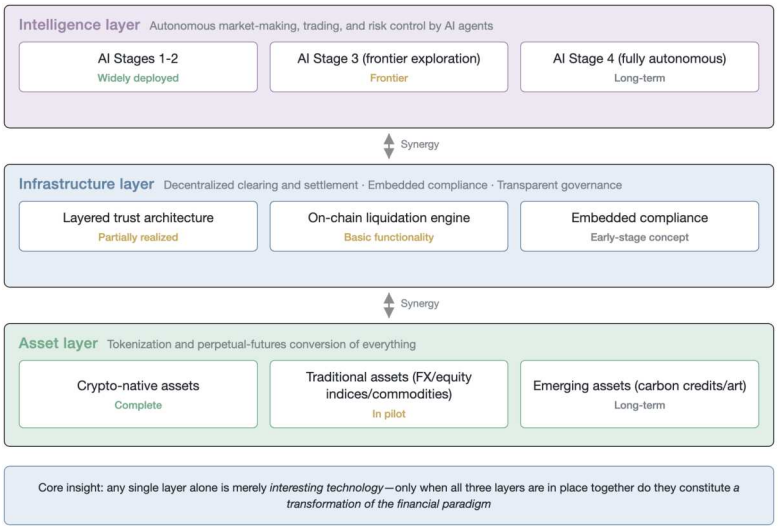


Figure 32-1. The three-layer stack architecture of autonomous finance (Data source: theoretical derivation in this chapter)

32.1.1 The asset layer

The asset layer is the cornerstone of the autonomous finance stack. In Chapter 27, we introduced the thought experiment of *perpetual futures on everything*, noting that the funding-rate anchoring mechanism is highly general: as long as a reliable external price source exists, any asset can be wrapped as a perpetual futures contract and traded. This mechanism substantially weakens the physical-delivery constraint of traditional finance (though it does not eliminate the constraint of geographic jurisdiction—that is, of regulation; see Section 32.5.3), allowing global investors to gain exposure to a given underlying through synthetic assets. In essence, the asset layer is a *value-*

mapping layer that converts diverse forms of real-world value—from crypto assets to commodities, from equity indices to exchange rates—into digital forms tradable on-chain. This conversion appears simple, but it involves complex problems of pricing, risk management, and liquidity aggregation.

At present, the development of the asset layer shows a marked phased character. The conversion of crypto-native assets into perpetual futures has largely matured at the product level, but the design of the underlying mechanisms is still evolving: the sharp divergence of funding rates under extreme conditions, the inconsistency of mark prices across exchanges, and the sustainability of insurance funds over a complete market cycle have yet to be fully tested. According to the report published by CoinGecko in February 2026, the world's top 10 centralized and decentralized exchanges processed as much as $92.9 trillion in notional (leveraged) perpetual futures trading volume across full-year 2025, an increase of 64.6% over 2024 [1]. For some crypto assets, perpetual futures volume has already surpassed spot volume, making them an important venue for price discovery and liquidity aggregation. Decentralized perpetual futures exchanges achieved explosive growth of 346%, reaching an all-time high of $6.7 trillion in volume [1]. The DEX perpetual futures volume should be interpreted with caution: token incentives and airdrop-points programs may have driven substantial wash trading, so the figure represents notional volume, and actual organic volume may be substantially lower. These data show that on-chain derivatives markets have evolved from a niche field into an important component of global financial trading. This growth is driven by several factors. First, perpetual futures offer greater capital efficiency than spot trading, allowing traders to obtain larger exposure with less capital. Second, 24-hour uninterrupted trading meets the needs of global investors, particularly those in different time zones. Third, the permissionless nature of decentralized platforms attracts users who are excluded from traditional exchanges.

Extending this model to traditional assets, however, remains at the pilot stage. Some frontier decentralized platforms have begun to experiment with perpetual futures on foreign exchange, equity indices, and commodities. For example, in discussing Hyperliquid's HIP-3 permissionless deployment mechanism, the same CoinGecko report notes that the platform can create a perpetual market for any asset with a reliable price feed, spanning commodities and even underlyings such as SpaceX and OpenAI [1]. These experiments demonstrate technical feasibility, but their trading scale and market depth remain far below those of crypto-native assets. This asymmetry reflects a deeper reality: although the technology is ready, the confidence of market participants, the clarity of regulatory frameworks, and the accumulation of liquidity still require time. The core challenges facing the conversion of traditional assets into perpetual futures include oracle reliability (equity prices require real-time, accurate feeds), regulatory uncertainty (attitudes toward derivatives trading differ widely across countries), and the cold-start problem for liquidity (a new trading pair needs enough participants before an effective market can form). Despite these challenges, we have already seen preliminary successes, which lay a foundation for the further expansion of the asset layer.

As for the *everything* vision of fully tokenizing or converting into perpetual futures such non-standard assets as carbon credits, real estate, or art, this remains a long-term goal. Such assets face not only more complex oracle-feed difficulties but also the con-

straints of legal title and physical enforcement that arise when real-world assets are brought on-chain. A real-estate perpetual futures contract, for example, requires a reliable house-price oracle, yet global house-price data are far less standardized than equity or commodity prices. Moreover, if settlement of the perpetual futures involves physical delivery, then cross-border legal and logistical costs rise sharply. The challenge for carbon-credit perpetual futures lies in the fact that carbon markets themselves are still in an early stage of development, with pricing mechanisms differing widely across countries and regions, which makes it difficult to build a unified on-chain carbon-credit market. Perpetual futures on art and collectibles, in turn, face the problems of authentication and the subjectivity of valuation. The evolutionary path of the asset layer is therefore a gradual one—from highly digitized native assets, gradually penetrating standardized traditional financial assets, and ultimately perhaps reaching non-standard physical assets. This process may take a decade or longer, but its direction is clear.

32.1.2 The infrastructure layer

If the asset layer defines *what* is traded, the infrastructure layer defines *how* to trade safely. In Chapters 28 through 30, we examined the decentralized clearing engine, the layered trust architecture, and embedded compliance mechanisms. Under the traditional system, clearinghouses and custodian banks rely on sovereign credit and vast legal-compliance apparatus to guarantee the finality of transactions. In the autonomous finance stack, these functions are hard-coded into immutable smart contracts and consensus protocols.

The on-chain clearing engine is the core component of this layer. Modern decentralized exchanges have already implemented cross-margining and cross-clearing functions similar to those of a traditional central counterparty (CCP), but in a permissionless environment these engines must cope with higher volatility and more extreme market conditions. In 2025, the open interest on decentralized exchanges surged by 229.6% (over the same period, open interest on centralized exchanges fell by 20.8%, reflecting a rotation of capital from CEXs to DEXs; note that 229.6% is a growth rate rather than a level), indicating that market participants' trust in on-chain infrastructure is rising substantially—they are not merely engaging in short-term speculation but also building long-term, substantive capital positions [1]. From a risk perspective, however, the rapid growth of open interest is at the same time a signal of accumulating systemic leverage, and readers should cross-reference the corresponding risk analysis in Section 32.4. This trust is built gradually: each time the system successfully weathers extreme conditions without failing, participants' confidence is further reinforced. Even so, under stress testing for capital efficiency and extreme tail risk, current on-chain clearing mechanisms remain functionally basic and not yet fully optimized.

The layered trust architecture seeks a balance between decentralization and performance. By building application chains (app-chains) optimized specifically for derivatives trading, platforms can achieve sub-second trade confirmation and high throughput while anchoring the final settlement layer to a more decentralized underlying network. This architecture has already been validated on some leading platforms, but cross-chain liquidity fragmentation and cross-domain interoperability remain pressing problems.

Imagine an investor who holds margin on chain A but wishes to trade an asset on chain B: this requires support from cross-chain bridges and liquidity protocols, which still carry technical and economic friction. Moreover, it is necessary to distinguish between *architectural decentralization* and *operational decentralization*. Most current derivatives app-chains anchor settlement architecturally to a decentralized base layer, yet remain highly centralized operationally: the sequencer is typically controlled by a single entity that orders transactions, the validator set is highly concentrated, and the platform retains emergency-intervention capabilities. Hyperliquid's 2025 JELLY incident—in which validators voted to freeze specific positions—shows that under extreme conditions the actual operating model of these platforms more closely resembles a centralized exchange with on-chain settlement than a genuinely decentralized protocol.

The least mature link in the infrastructure layer is embedded compliance. As Chapter 30 noted, for decentralized finance to enter the mainstream it must encode its responses to regulatory requirements. Technologies such as zero-knowledge proofs (ZKPs) and digital identity credentials provide a theoretical foundation for a system architecture in which lawbreaking is rendered *impossible*, but in actual deployment, how to establish a unified on-chain compliance standard across globally fragmented regulatory regimes remains at a very early, conceptual stage. The deeper obstacle is not only technical immaturity but also the fundamental incompatibility between zero-knowledge-proof compliance and existing legal frameworks: the U.S. Bank Secrecy Act requires financial institutions to be able to *produce* actual customer records, not merely to supply a cryptographic proof of compliance, while the EU's Anti-Money Laundering Directive requires transparent disclosure of beneficial ownership. A zero-knowledge proof can demonstrate that an address is not on a sanctions list, but it cannot satisfy regulators' *record production obligations*. U.S. anti-money-laundering rules, for example, conflict fundamentally with EU data-privacy rules, and there is as yet no clear answer to how a globalized on-chain system could satisfy both simultaneously.

32.1.3 The intelligence layer

The intelligence layer sits at the top of the autonomous finance stack. In Chapter 31, we analyzed the four stages of AI's penetration into on-chain finance. Although traditional quantitative trading has long been widespread, an autonomous finance network requires autonomous agents that can continuously assess risk, provide quotes, and execute complex strategies without human intervention.

At present, the use of AI in market making and algorithmic trading (that is, the first and second stages) is already widespread. In both centralized and decentralized markets, the overwhelming majority of liquidity is algorithm-driven. According to McKinsey's 2025 State of AI global survey, 62% of firms report at least experimenting with AI agents (about 23% have deployed them at scale in certain functions and 39% are at the experimental stage), indicating that the underlying technical foundation for the intelligence layer is accumulating at an accelerating pace [10]. On decentralized exchanges, because interfaces are fully open and smart contracts are composable, AI agents can conduct cross-market arbitrage and liquidity management more freely than in traditional financial systems. An AI market-making bot can provide liquidity si-

multaneously on multiple DEXs and automatically execute arbitrage when it detects a price inconsistency—all on a millisecond timescale.

Entering the third stage, however—highly autonomous AI conducting complex capital allocation and dynamic risk control within decentralized finance protocols—remains a frontier exploration. Such agents must handle risks specific to on-chain environments, including maximal extractable value (MEV) attacks, smart-contract vulnerabilities, and oracle latency. An AI risk-control system, for instance, must predict not only market risk but also the behavior of other AI agents and the possible failure modes of smart contracts. This is a far more complex problem than traditional risk control. As for the fourth stage—a fully autonomous market of *AI versus AI*, in which market microstructure is entirely dominated by machine logic—it remains a theoretical conception.

The core value of the intelligence layer lies in the way it substantially lowers the threshold and cost of liquidity provision. Most of what is currently called *AI market making* on-chain remains rule-based automated market making—adjusting automated market maker (AMM) parameters, running grid strategies, and the like—together with machine-learning-assisted parameter optimization, rather than end-to-end autonomous AI decision-making. Genuine autonomous market making is constrained by on-chain latency (approximately 200 milliseconds, compared with the microsecond-level response of centralized exchanges) and by pre-trade transparency (mempool visibility and the risk of sequencer front-running); these structural constraints make on-chain AI market making, at the current stage, closer to *intelligent automation* than to *AI autonomy*. In the incumbent financial architecture, market making is a capital- and technology-intensive privileged business in which only a few large institutions can participate. In the autonomous finance stack, open-source AI market-making bots combined with on-chain liquidity pools allow ordinary investors to participate in liquidity provision as well, thereby democratizing financial infrastructure. An individual holding a small amount of stablecoins can deploy a simple market-making bot to provide liquidity on a given trading pair and earn income from trading fees.

32.1.4 Synergies

Taken separately, each of the asset layer, the infrastructure layer, and the intelligence layer represents progress in a specific technology. Perpetual futures are an innovative financial derivative; on-chain clearing is an efficient piece of distributed-systems engineering; and AI trading agents are a successful application of machine learning. Developed in isolation, they would at most be *interesting technologies*. But when the three layers interweave and fall into place at once within a single permissionless network, they constitute a paradigm shift in finance.

This synergy is manifested first in the mutual reinforcement of liquidity and trust. The asset layer's *perpetual futures on everything* substantially expands the boundary of tradable underlyings, but without the decentralized clearing and immutable ledger provided by the infrastructure layer, traders would be unable to establish trust in a global network that lacks traditional legal protection. Conversely, without the around-the-clock, deep liquidity provided by the intelligence layer's AI agents, perpetual futures on long-tail or emerging assets would become unattractive because of excessively wide

bid-ask spreads. A perpetual futures contract on an emerging-market currency pair, for example, would have prohibitively high trading costs without sufficient liquidity, losing its value as a hedging instrument. This positive reinforcing cycle also operates in reverse: when liquidity declines, trust erodes with it, triggering capital withdrawal that further depletes liquidity, forming the liquidity-spiral mechanism described by Brunnermeier and Pedersen (2009) [11]. The key question is what conditions trigger the system to switch from the positive cycle to the negative one—an issue developed further in the systemic-risk analysis of Section 32.4.

Second, the synergy is manifested in the reconciliation of compliance and innovation. The embedded compliance being explored in the infrastructure layer allows the asset layer to satisfy the regulatory requirements of a given jurisdiction while preserving permissionless access. The intelligence layer's risk-control agents can read these compliance parameters in real time and automatically adjust trading strategies, ensuring that capital is allocated globally without crossing regulatory red lines. This hard-wiring of compliance and trading logic at the code level is unimaginable in traditional systems built from human compliance officers and cumbersome paper documents. A zero-knowledge proof can demonstrate that a trader satisfies a specific compliance condition (for example, that the trader is not a U.S. person) without revealing the trader's true identity.

Finally, the combination of the three layers unlocks capital efficiency. In traditional finance, an investor who wants to hedge foreign-exchange risk, invest in U.S. technology stocks, and short a commodity all at once must cross multiple time zones, open different brokerage accounts, and endure lengthy settlement cycles. In the autonomous finance stack, through the composability of smart contracts, an AI agent can, within milliseconds and using the same stablecoin margin, establish all three positions simultaneously with the support of a decentralized clearing engine. All friction is abstracted away, and capital flows freely, without borders and without time-zone lag.

This synergy also appears along the risk-management dimension: when the three-layer stack is fully in place, risk is no longer isolated but understood and managed by the entire system—the diversification of the asset layer lowers single-asset risk, the transparency of the infrastructure layer supports real-time monitoring, and the automation of the intelligence layer enables rapid response.

Although the components of the autonomous finance stack differ in maturity, its core skeleton is already in place. The success of crypto-native assets has proven the feasibility of the underlying logic, while the evolution of infrastructure and the spread of AI agents are accelerating the network's expansion into the broader field of traditional finance. This is not merely an upgrade of trading tools; it is a fresh answer to the fundamental question of who should build financial systems, who should run them, and whom they should serve.

From a systems-theory perspective, the core value of the three-layer stack lies in its emergence. When the asset layer provides sufficiently diverse underlyings, the infrastructure layer provides sufficiently reliable clearing mechanisms, and the intelligence layer provides sufficiently efficient liquidity, the system as a whole spontaneously evolves new functions and possibilities. These new functions are not preplanned by any single layer's

designers but emerge naturally from the interaction of the three layers. Cross-asset automatic hedging strategies, dynamic risk management, and even new forms of financial products, for example, may all arise spontaneously within this system. The magnitude of the synergy can be illustrated with a conceptual example. In a three-asset portfolio of BTC, the S&P 500, and gold, because the correlation matrix among the three exhibits a substantial negative-correlation component in most market environments, the aggregate margin requirement under cross-margining can be 30% to 50% lower than under the isolated-account model. This estimate, however, requires three important qualifications: (1) the 30% to 50% margin saving assumes that the correlation structure remains stable under normal market conditions; (2) under stress scenarios, cross-asset correlations tend toward 1 (systematically demonstrated by Forbes and Rigobon (2002) [12]), at which point the margin saving may narrow to as little as 5% to 15%; and (3) no on-chain DEX currently offers perpetual futures on BTC, the S&P 500, and gold simultaneously, so this portfolio serves only as a conceptual example rather than a currently realizable trading strategy. This means that the same amount of principal can support greater risk exposure within the three-layer stack, but this gain in capital efficiency also implies greater system fragility to extreme jumps in correlation. When cross-asset correlations tend toward 1 under stress, the margin buffer contracts far faster than under the isolated-account system (see Section 32.4.1 for a dedicated analysis of cross-margining risk under stress scenarios).

As the foregoing analysis shows, the components of the three-layer stack differ significantly in maturity, and this difference means that the pace of the whole system's evolution will be constrained by how quickly its weakest link can be overcome. Table 32-1 summarizes the core functions, current maturity, and key bottlenecks of each layer of the three-layer stack, providing a baseline reference for the later sections' analysis of the real-world constraints on the three-layer synergy.

Layer	Core function	Current maturity	Key bottleneck
Asset layer	Converts underlyings into on-chain perpetual futures	Crypto-native assets mature; traditional assets at the pilot stage; non-standard assets a long-term prospect	Oracle reliability, regulatory clarity, liquidity cold start
Infrastructure layer	Decentralized clearing and settlement, and embedded compliance	Clearing engine functionally largely complete; cross-chain interoperability fragile; embedded compliance at the concept stage	Cross-chain liquidity fragmentation, unification of global compliance standards
Intelligence layer	Autonomous market making, risk control, and strategy execution by AI agents	Market making and arbitrage widely deployed (stages one and two); autonomous capital allocation at the frontier (stage three)	Perception of on-chain-specific risks, explainability, strategy homogenization

Table 32-1. Maturity comparison across the three-layer stack (Data source: based on the analysis in Sections 32.1.1–32.1.3)

The engineering properties of the three-layer stack—permissionless access, extremely low barriers to participation, and globally accessible infrastructure—correspond directly to the scale of global financial exclusion. Translating this technical potential into financial inclusion, however, requires distinguishing the essential difference between *lowering barriers to entry* and *effective empowerment*.

32.2 The structural constraints on financial inclusion

The existence of billions of unbanked people worldwide makes financial inclusion not merely a moral appeal but a systemic problem with clear engineering constraints. The technical properties of the three-layer stack in lowering friction costs and barriers to entry correspond directly to the core pain points these populations face—foreign-exchange hedging, inflation-resistant savings, and low-cost cross-border payments. Beginning from the genuineness of the demand, this section examines the inclusive potential of perpetual futures, on-chain infrastructure, and AI agents, then confronts the unavoidable structural constraints of financial-literacy gaps, the digital divide, and the *inclusion trap*.

32.2.1 Genuine demand

When we speak of financial inclusion, we must first establish the genuineness and urgency of these needs. According to the 2025 edition of the World Bank's Global Findex Database, roughly 1.3 billion adults worldwide remain unbanked (account ownership has risen to about 79%; the earlier literature's range of 1.3 billion to 1.7 billion reflects differences across survey years such as 2017, 2021, and 2025), a group concentrated primarily in developing countries in Asia, Africa, and Latin America [2] [3]. Behind these numbers lie genuine economic hardship and the erosion of wealth. Small-business owners in emerging markets face the daily risk of local-currency depreciation, yet no effective foreign-exchange hedging instrument is available to them. Traditional commercial banks typically do not offer forward FX contracts or options to small customers, and even when they do, the high fees and margin requirements deter micro and small enterprises. This scarcity of financial instruments leaves such firms defenseless against global macroeconomic fluctuations, and their profits may shrink sharply because of exchange-rate movements.

Beyond exchange-rate risk, savers in developing countries also face persistent erosion from inflation. In many emerging-market economies, inflation has long held at elevated levels; average inflation in emerging markets and developing economies in 2024, for example, was about 8% (approximately 8.1% per the International Monetary Fund (IMF) World Economic Outlook), while the 2025 forecast is about 5.3% (on the basis of the IMF's October 2025 World Economic Outlook) [13]. In this environment, if savers can only hold their money in local currency at home or in local bank accounts with extremely low yields, the real purchasing power of their wealth will inevitably shrink. Yet because of capital controls, high account-opening thresholds, and complex

cross-border compliance requirements, these savers can almost never obtain exposure to international assets. Unlike investors in developed countries, they cannot hedge local inflation risk by buying an S&P 500 index fund or U.S. Treasuries. This absence of investment channels is, in essence, a structural form of wealth deprivation.

In addition, the world's tens of millions of cross-border remittance workers likewise bear the heavy costs of an inefficient financial system. These workers leave their hometowns to labor in foreign lands and remit their savings home to support their families. According to World Bank and IMF data, however, the average cost of personal cross-border remittances has long held at around 6% (recent World Bank data put it at approximately 6.2% to 6.65%, still more than twice the Sustainable Development Goals' 3% target), and in some regions it is even higher [14] [15]. This means that tens of billions of dollars each year are consumed by financial intermediaries in the friction of cross-border transfers—money that could otherwise have improved the lives, education, and health of poor families. These needs—foreign-exchange hedging, inflation-resistant savings, and low-cost cross-border payments—are not false demand created by financial engineering but genuine pain points that billions of people worldwide face every day.

32.2.2 The inclusive functions of perpetual futures

Confronted with these genuine financial needs, the traditional financial system—because of its high operating costs, complex compliance requirements, and path dependence on high-net-worth clients—has consistently failed to provide effective solutions. The perpetual futures examined in Chapter 27, once combined with decentralized infrastructure, show the potential to become a powerful instrument of inclusion. Foreign-exchange perpetual futures give small-business owners in emerging markets a hedge against local-currency depreciation at a cost significantly below that of traditional solutions. Unlike a traditional bank's forward contract, participating in on-chain FX perpetual futures requires no complex credit review or high account-opening threshold. A small-business owner needs only a smartphone wallet and a small amount of stablecoins to establish, on a decentralized exchange, a perpetual futures position opposite to the firm's actual exposure. Because perpetual futures are tightly anchored to the spot price through the funding-rate mechanism, this hedge can precisely offset the losses caused by local-currency depreciation; and on liquid major trading pairs, the empirical analysis of on-chain stablecoin foreign exchange by Adams et al. (2023) [16]—using EUROC/USDC as a proxy for EUR/USD—shows that the on-chain trading price deviates from the traditional-market exchange rate by an average of about 5 basis points during normal trading hours. For a $500 on-chain cross-border transfer, the cost in the best case is about $4.80, roughly 83% lower than the $28 charged through traditional banking channels. Liquidity-provider fees for on-chain AMM spot swaps are typically about 1 to 5 basis points (0.01% to 0.05%), far below the spreads and credit-review costs embedded in traditional bank forward contracts, substantially lowering the economic threshold for small conversions. The cost-advantage figures above, however, derive primarily from empirical analysis of major currency pairs such as USDC/EUROC; for the small-currency pairs of emerging markets, on-chain FX liquidity remains limited and spreads may reach 50 to 200 basis points. Moreover, the full-stack cost must also include the on-ramp and off-ramp

costs between fiat and stablecoins (typically 2% to 5% in emerging markets), which substantially narrows the cost advantage of on-chain solutions for end users who still need fiat conversion.

For savers in developing countries facing inflationary erosion, equity-index perpetual futures and commodity perpetual futures offer a channel across geographic and institutional barriers. In the traditional financial world, an ordinary saver in India or Nigeria who wishes to invest in the U.S. stock market must overcome layer upon layer of obstacles—foreign-exchange controls, opening an overseas securities account, and meeting minimum-capital requirements—which is entirely impossible for the great majority of people. Through synthetic-exposure perpetual futures (whose frictionless mechanism for decoupling an asset's economic attributes from physical ownership is detailed in Section 32.3.3), with only $100 worth of stablecoins this saver can gain precise on-chain exposure to the S&P 500 or gold, fundamentally breaking the monopoly under which *high-quality assets are reserved for the rich and for residents of developed countries.*

Going further, the combination of stablecoins and perpetual futures can, in a sense, substitute for—or even surpass—the function of a traditional savings account. In a country with high inflation and an unstable local currency, holding a dollar-pegged stablecoin is itself a defensive strategy. If the user further uses these stablecoins as margin to provide liquidity or adopt a particular hedging strategy in the perpetual futures market, the user can also earn a continuous stream of funding-rate income. In a market environment of strong long demand, this funding-rate-based yield is often far higher than the deposit rates of traditional banks. This combination not only protects the purchasing power of the principal but also creates positive cash flow, providing the unbanked with a fully functional *synthetic yield strategy.* Yet the risk profile of this strategy differs fundamentally from that of traditional savings: the funding rate can remain negative for extended periods, at which point the *yield* turns into a continuous cost outflow; even under low leverage, extreme conditions can still trigger margin calls or forced liquidation; and stablecoins themselves carry depegging risk, as the 2022 collapse of UST to zero and the temporary fall of USDC to $0.87 in March 2023 both illustrate. A user who understands such a strategy as *savings* without fully recognizing these risks is a textbook case of the inclusion trap discussed in Section 32.2.5. In addition, Rajan (2006) [17] argued that an inherent tension exists between the proliferation of financial instruments and the diffusion of systemic risk: as more participants gain access to complex instruments, the total quantity of risk-taking and the pathways of its transmission expand in tandem. The democratization of perpetual futures faces the same dilemma, and its potential negative effects are discussed in detail in Section 32.2.5.

32.2.3 The inclusive properties of the infrastructure

The reason perpetual futures can be transformed from a high-risk speculative instrument into a vehicle for financial inclusion is fundamentally that the on-chain infrastructure on which they run possesses three core properties absent from the traditional financial system: permissionlessness, extremely low barriers, and global accessibility. First, permissionlessness is the most powerful force for breaking down financial exclu-

sion. In traditional finance, know-your-customer (KYC) and anti-money-laundering (AML) requirements—though motivated by the reasonable goal of preventing financial crime—in effect exclude enormous numbers of people. Hundreds of millions worldwide are permanently locked out of the formal financial system because they lack official identification, have no fixed address, or have a blank credit history. The execution logic of smart contracts on decentralized exchanges contains no conditional test of a user's identity attributes; they verify only cryptographic signatures and margin adequacy. This architecture, which executes undifferentiated logic for every participant meeting the verification conditions, gives those abandoned by traditional systems an equal right to participate. Under the current legal framework, however, permissionlessly offering leveraged derivatives to the public constitutes a violation in multiple jurisdictions. The Commodity Futures Trading Commission (CFTC) has brought enforcement actions against offshore decentralized protocols such as Ooki DAO, alleging that they offered unregistered derivatives products to U.S. users. This tension between inclusive value and compliance obligation is precisely the concrete mapping of the trilemma analyzed in Section 32.5.2.

Second, on-chain infrastructure lowers the barrier to participation in financial instruments to a level the traditional system cannot reach. On traditional derivatives markets such as the Chicago Mercantile Exchange, participating in futures trading typically requires thousands or even tens of thousands of dollars of initial margin, which necessarily confines the market to institutional investors and high-net-worth individuals. By contrast, because the automated liquidation mechanism of smart contracts eliminates the fixed costs of human intervention and intermediaries, the minimum margin requirement for on-chain perpetual futures can be as low as $1 to $10 (Hyperliquid and other decentralized platforms, for example, allow positions to be opened with single-digit dollar amounts). This micro-level divisibility lets financial instruments precisely fit the actual capital scale of low-income populations in developing countries. A small vendor with only a few dollars of surplus a day can use the most advanced financial derivatives to manage micro-level risk. But the technical minimum threshold does not equal economic accessibility. Take a position with $10 of margin and 10x leverage ($100 notional): a single open-and-close round trip of taker fees plus spread may consume 0.5% to 1% of the margin, and once funding rates are added, the cost efficiency of a very small position is significantly lower than that of a larger one. The gap between technical reachability and economic viability must be assessed with care.

Finally, the global accessibility of on-chain systems eliminates the financial-island effect created by geographic borders. Traditional financial markets are highly fragmented, and the *regulatory pricing islands* discussed in Chapter 18 are commonplace in reality. Residents of different countries can access only the limited products that local financial institutions are willing or permitted to offer. Financial infrastructure deployed on a public blockchain, by contrast, is a single global state machine. As long as there is an internet connection, a farmer in a remote African village and a hedge fund manager on Wall Street access exactly the same liquidity pool and face exactly the same execution logic and transaction costs. This globally unified infrastructure substantially narrows the geographic disparity in the supply of financial services.

32.2.4 The democratization of AI financial advisers

If perpetual futures and on-chain infrastructure solve the problems of the *supply* and *reach* of financial instruments, the introduction of AI agents addresses the core *cognition* and *execution* bottlenecks of financial inclusion. In the traditional wealth-management industry, professional financial-advisory service is an expensive luxury. A qualified private banker able to provide asset allocation, dynamic hedging, tax optimization, and risk management typically serves only high-net-worth individuals with investable assets exceeding $1 million and charges thousands to tens of thousands of dollars a year in advisory and management fees [18]. For the vast majority of ordinary people worldwide, professional financial advice at this level is entirely inaccessible. They can only rely on their own intuition, limited common sense, or unreliable recommendations from social networks to make financial decisions, which often leads to serious consequences.

The intelligence layer of the autonomous finance network is democratizing this expensive professional service through large language models (LLMs) and specialized reinforcement-learning algorithms. An AI financial agent can be regarded as a tireless, emotionless *around-the-clock private banker* with extremely low marginal cost. Consider a small-business owner in Jakarta. She need not understand complex funding-rate mechanisms or implied volatility; she need only state her need to her AI agent in natural language: "Please protect the rupiah I've set aside for next month's inventory purchases from losses caused by dollar appreciation." The AI agent automatically converts these funds into stablecoins, seeks the deepest-liquidity rupiah-to-dollar perpetual futures on the optimal decentralized exchange, computes the precise hedge ratio, executes the trade automatically, and, when market conditions change or the Federal Reserve (Fed) issues an unexpected statement, automatically adjusts the position within the limits allowed by block-confirmation time.

This *AI-as-a-service* model not only lowers the operational threshold; more importantly, it bridges the vast gap in ordinary users' financial literacy. An AI agent can monitor a portfolio's risk exposure in real time and set strict stop-loss logic, preventing users from taking on devastating losses through cognitive bias or emotional loss of control. In traditional finance, risk control is imposed by institutions to protect their own interests; in an autonomous finance network, the AI agent is the user's personal fiduciary, whose sole objective is to maximize the user's utility and control downside risk. By encapsulating complex financial-engineering logic beneath a friendly natural-language interface, AI agents allow the most advanced derivatives instruments to serve safely those with the least financial knowledge—and this is the essence of what technology contributes to financial inclusion. However, behavioral-finance research shows that lowering transaction friction may actually intensify overtrading and erode retail returns (Barber and Odean, 2001 [19]). While a natural-language interface lowers the operational threshold, it may also mask the complexity of the underlying derivatives position: when a user issues a simple hedging instruction, the user may not understand that they are opening a leveraged perpetual futures position that carries liquidation risk. Moreover, AI agents at present bear no fiduciary duty and are subject to no regulatory accountability, and when a strategy recommended by an AI leads to a margin call or blow-up, how responsibility is allocated among the user, the developer, and the protocol remains an unresolved legal question.

The foregoing analysis shows that the path to realizing financial inclusion involves three progressive layers—identifying genuine demand, matching technical tools, and responding to structural constraints (Figure 32-2).

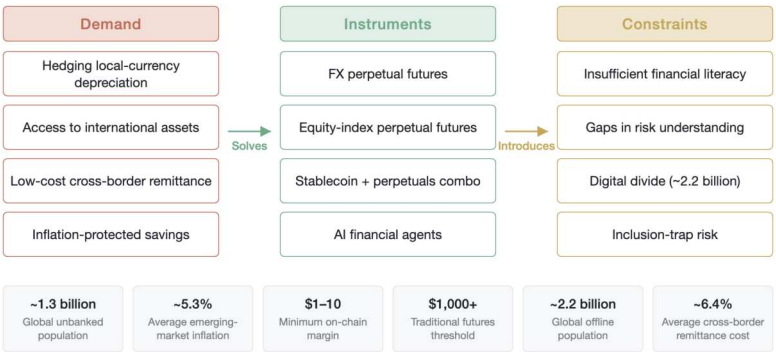


Figure 32-2. An analytical framework for financial inclusion: demand, tools, and constraints (The figure is organized in three columns—demand, tools, and constraints; the left column lists three categories of core pain points, namely foreign-exchange risk, inflation erosion, and cross-border remittance costs, each broken down into specific items that map respectively to the tool solutions in the middle column and the structural constraints in the right column; the framework structure is a theoretical derivation of this chapter; the empirical anchors at the bottom of the figure have been updated to the latest verifiable values: the global unbanked population is approximately 1.3 billion, per the 2025 edition of the World Bank's Global Findex Database, survey year 2024 [2][3]; average inflation in emerging markets and developing economies is approximately 5.3%, per the IMF's October 2025 World Economic Outlook projection, whereas the approximately 8% cited in the main text is the 2024 realized value, the two being different measures; the global offline population is approximately 2.2 billion, per the International Telecommunication Union's Facts and Figures 2025 [7]; the average cost of cross-border remittances is approximately 6.4%, per the World Bank's Remittance Prices Worldwide for the third quarter of 2025, with a global mean of 6.36%; the on-chain minimum margin of approximately $1 to $10 and the traditional-futures threshold of more than $1,000 are product facts, detailed in Section 32.2.3; empirical data as of June 2026)

32.2.5 Real constraints

Yet the analysis of the preceding 31 chapters has repeatedly shown that any leap in technological paradigm comes with irreducible costs. As we celebrate the inclusive vision that the three-layer stack brings, we must remain soberly realistic and confront the real constraints that impede this vision. These constraints are not temporary technical problems but structural features of inclusive finance itself.

The first is the fundamental deficiency of financial literacy and risk understanding. Although AI agents can assist decision-making to a large degree, the ultimate bearer of risk is still the user. Perpetual futures are in essence a leveraged derivative; while they can effectively hedge foreign-exchange or inflation risk, they also introduce entirely new liquidation risk, smart-contract-vulnerability risk, and oracle-failure risk. For the user, the nature of the risk has shifted, but the total quantity of risk has not necessarily decreased. This is the concrete manifestation, within inclusive finance, of the distinction between *risk transfer* and *risk elimination* discussed in Chapter 25. If a user who wholly lacks financial common sense blindly participates in highly leveraged trading

merely because the barrier to entry has fallen, then this *inclusion* is in essence the exposure of users who lack risk-management capacity to higher leverage risk. In fact, the constraint lies not only in the absence of financial literacy. Behavioral-finance research shows that even sophisticated investors with professional knowledge systematically exhibit overconfidence and loss aversion, which makes derivatives products fundamentally at odds with the human cognitive architecture rather than merely a knowledge gap that education can bridge. Not only does it fail to improve users' economic circumstances; it can accelerate the destruction of their wealth. This phenomenon has many precedents in financial history. In the 2008 subprime mortgage crisis, many low-income American households were encouraged to enter complex mortgage products they did not understand, ultimately leading to mass loss of housing. Similar tragedies have played out repeatedly in microfinance, where many products billed as *inclusive*, but usurious in nature, ultimately drove poor populations into deeper debt traps.

Second, we must confront the harsh digital divide. On-chain finance's *permissionlessness* rests on the implicit premise of *possessing the capacity for digital access*. In the real world, however, technology itself is a powerful barrier to entry. According to data from the International Telecommunication Union (ITU) and the GSMA, a large share of the world's population still lacks access to mobile internet (the ITU estimates that about 2.6 billion people are entirely offline—a 2023 figure since revised down to about 2.2 billion for 2025 [7]; the GSMA estimates that a further 3 billion-plus people fall within mobile-network coverage yet do not use mobile internet, the so-called *usage gap* [8]); they either lack a smartphone or cannot obtain a stable, reliable internet connection. This number even exceeds the size of the world's unbanked population. For those in conditions of extreme poverty and infrastructural scarcity, the most advanced on-chain perpetual futures and AI agents are meaningless. They cannot participate in the autonomous finance network—not for lack of permission, but for lack of the most basic means of technical access. The existence of this digital divide makes the inclusive promise of autonomous finance inherently limited. If the expansion of the autonomous finance network merely gives better investment tools to a middle class that already owns digital devices and internet access, while further marginalizing the bottom 3 billion, then this technological evolution will aggravate rather than alleviate global inequality. This is a disquieting reality but one that must be faced.

Finally, we must be wary of the so-called *inclusion trap*. In financial history, cases of predation in the name of *inclusion* are common, from subprime mortgages to certain usurious forms of microfinance, all of which drove vulnerable groups deeper into debt. If the actual operating result of the autonomous finance network is to draw more retail investors who lack risk tolerance into a complex market they do not understand at all, making them the liquidity counterparties of institutional market makers and high-frequency trading algorithms, then it degenerates into a mechanism for *democratizing risk*—pushing risk downward. This democratization of risk is fundamentally different from financial inclusion. The former spreads the burden of risk evenly across everyone, including those least able to bear it; the latter helps people manage risk better through education, protection, and empowerment. Financial inclusion is by no means simply *lowering barriers to entry*; it must combine *effective tools*, *systematic education*, and *embedded risk control*. The three-layer stack provides powerful tools, but how to build

a supporting system of social education, how to design ethical norms that prevent the technology from being abused, and how to ensure that the most vulnerable groups do not become the system's sacrifice will be the key constraints determining where this vision ultimately leads. These questions cannot be answered by technology itself; they require the deep involvement of society, policymakers, and ethicists.

The real constraints on inclusion show that the accessibility of financial instruments is only half the equation; the other half is whether users can wield these instruments effectively while fully understanding the risks. Even as the three-layer stack lowers barriers to entry, it is also reshaping the way global capital is allocated.

32.3 The globalization of capital-allocation efficiency

Global capital markets are nominally highly interconnected, but at the microstructural level they remain deeply divided by geographic barriers, session limits, and unequal access. Through mechanisms such as synthetic exposure, around-the-clock continuous pricing, and cross-margining, the three-layer stack fundamentally changes how capital crosses these barriers. This section analyzes the concrete manifestations of the current fragmentation of global capital allocation, assesses the efficiency gains of the three-layer stack in eliminating temporal and spatial friction, and asks about the distributional effects of this efficiency dividend: who are the main beneficiaries, and who will bear the costs of creative destruction.

32.3.1 The current fragmentation of capital allocation

The fragmentation of global capital markets is manifested first in entrenched geographic barriers. The financial markets of different countries are strictly separated by legal jurisdictions, clearing systems, and monetary boundaries. An ordinary investor in Jakarta who wishes to buy U.S. technology stocks must clear numerous obstacles, including opening an offshore securities account, converting currency, and paying steep cross-border remittance fees. Conversely, capital in New York seeking to invest in high-yield Indian bonds likewise faces complex qualified-foreign-investor quotas and capital controls. This geographic segregation produces pronounced market segmentation, so that assets with identical risk characteristics may be priced very differently across markets [20].

Beyond geographic barriers, session limits are another important factor behind market fragmentation. Modern economic activity proceeds around the clock, and information is generated continuously in every corner of the globe, yet the trading hours of core capital markets still adhere to an industrial-age timetable. The core trading session of the New York Stock Exchange, for example, is limited to 9:30 a.m. to 4:00 p.m. Eastern Time. Major geopolitical events, corporate earnings releases, or macroeconomic data published outside these 6.5 hours cannot be immediately reflected in prices. This session limit artificially severs the immediate transmission between information and price, producing sharp price gaps at the open [21]. Information not priced in time not only increases market uncertainty but also leaves investors across time zones in a passive position when managing risk.

In terms of access mechanisms, the traditional financial system embeds a deep inequality between institutions and retail investors. Large multinational institutional investors can obtain exposure to almost any market and any asset through global custodian-bank networks, prime brokerage services, and complex derivatives instruments. They enjoy privileged channels that cross geographic and temporal barriers. For the vast majority of retail investors worldwide, however, this capacity for global allocation is out of reach. They are confined to local markets and can only passively accept the inflation risk of their own currency and the systemic risk of a single market. This inequality of access, in essence, reserves the world's best investment opportunities and hedging instruments for groups that already possess a capital advantage [22].

The cost of this fragmentation is high: it directly produces enormous losses in the efficiency of global capital allocation. When capital cannot flow freely and frictionlessly, funds cannot reach the places with the highest expected returns in time. In some emerging markets there may be highly promising innovative firms and high-yield projects whose development is constrained by the absence of channels for global capital injection; meanwhile, in developed economies, large amounts of surplus capital may be able to circulate only among low- or even negative-yielding assets. The fragmentation of global capital allocation not only lowers overall economic productivity but also aggravates the imbalance in wealth distribution.

These three forms of fragmentation—geographic barriers, session limits, and unequal access—together constitute the microstructural basis of the current losses in global capital-allocation efficiency; Figure 32-3 contrasts the current state of fragmentation with the target state after integration by the three-layer stack.

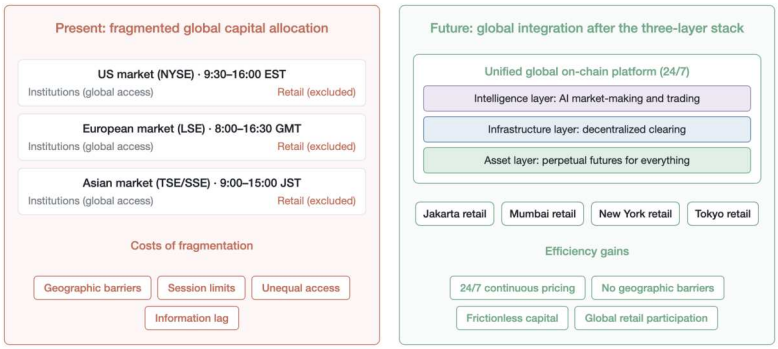


Figure 32-3. Fragmented global capital allocation versus integration by the three-layer stack (Data source: theoretical derivation in this chapter)

32.3.2 Around-the-clock continuous pricing

In an autonomous finance network, session limits are eliminated and replaced by an around-the-clock global continuous-pricing mechanism. As the Tesla-stock simulation in Chapter 27 showed, around-the-clock pricing substantially reduces the magnitude of opening gaps caused by weekend information backlogs. In that thought experiment, a sudden event occurring over the weekend no longer had to wait for the Monday-morning

opening bell; market participants could trade and hedge promptly once information was generated. Existing crypto-market data show, however, that around-the-clock operation does not mean an even distribution of liquidity. Public data show that order-book depth for major trading pairs exhibits a strong diurnal cycle, falling by about 40% during the low-activity hours after the U.S. market close relative to the intraday peak (estimated from public data on single trading-pair samples such as those on Binance); what declines over the weekend, however, is chiefly trading-volume share and the spread, while order-book depth itself does not fall in step. Low-liquidity windows can transform gap risk into thin-market abnormal-volatility risk, and the slippage a stop-loss order experiences in a liquidity vacuum can far exceed expectations. This effect especially threatens retail users in emerging markets, who may trade during the windows of lowest liquidity in the global cycle.

When this mechanism extends from a single asset to all major global assets, the resulting efficiency gain is structural. From U.S. equities to European sovereign bonds to Asian agricultural commodities, all assets trade continuously on a single global platform with no concept of a close. Because trading hours cover all geographic time zones with no market-closed window, the price-discovery process is no longer disturbed by artificially set trading periods. New information generated at any moment—whether macro data from Tokyo, a central-bank decision from London, or a technological breakthrough from Silicon Valley—can be translated into movements in global asset prices with extremely low latency.

This continuous-pricing mechanism reduces the excess volatility that arises in traditional markets from information backlogs. In fragmented markets, unpriced information accumulated during market closures is often released all at once at the instant of the open, causing liquidity to dry up and prices to swing violently. In an around-the-clock market, by contrast, information is digested smoothly and continuously. Investors in different time zones can participate in the market according to their own schedules: Asian investors digest Asian information during their day, and European and American investors take up the relay during their respective active hours. This relay not only provides continuous liquidity but also allows prices to reflect an asset's true fundamentals more accurately [21].

More importantly, continuous pricing eliminates the timing-mismatch risk of cross-market arbitrage. In the traditional system, because different markets' trading hours do not overlap, cross-market hedging often faces enormous execution risk. Attempting to hedge a spot exposure in the U.S. market with a derivative in the Asian market, for example, requires bearing the price-movement risk created by the gap between the two markets' opening times. On a unified around-the-clock platform, by contrast, the prices of all related assets update synchronously, so complex cross-market, cross-asset-class arbitrage and hedging strategies can be executed precisely, further improving the pricing efficiency of global markets.

32.3.3 Frictionless capital flows

Beyond temporal barriers, innovations in the asset and infrastructure layers let the autonomous finance network achieve frictionless global capital flow. The core of this

shift lies in the mechanism of synthetic exposure, and in particular the widespread use of perpetual futures. In the traditional model, by contrast, obtaining the economic return of an asset generally requires physical settlement—that is, actually buying and holding the asset. This transfer of physical ownership comes with cumbersome legal procedures, cross-border settlement delays, custody fees, and foreign-exchange conversion costs [23].

Perpetual futures largely decouple an asset's economic attributes from its physical ownership. An investor in Mumbai who is bullish on the long-term prospects of the U.S. technology sector no longer needs to go through a complex account-opening and currency-purchase process to buy a real Nasdaq index fund. Through perpetual futures built on on-chain infrastructure, the investor need only use stablecoins as margin to obtain price-movement returns identical to those of holding the real stock. In this process, no real stock crosses a border and no cross-border capital flow occurs in the traditional sense, yet the purpose of capital allocation is perfectly achieved. All friction costs—from clearinghouse fees to custodian-bank cuts to the spreads of foreign-exchange market makers—are sharply compressed. This decoupling conceals an extreme-scenario risk not yet fully examined: when the underlying asset's liquidity dries up or the protocol suspends trading in a specific asset, holders of synthetic exposure may find their exit channel closed, trapped in a state of *phantom liquidity* in which a position is nominally liquid but in fact cannot be closed. The 2008 synthetic CDO crisis already demonstrated that the liquidity illusion of synthetic instruments can shatter in an instant under stress.

The cross-margining mechanism of on-chain infrastructure pushes this frictionless flow further still. In a traditional fragmented account system, capital is used very inefficiently. An investor's funds may be scattered across a stock account, a foreign-exchange account, and a commodity-futures account, each of which must independently maintain high margin. Even if the investor holds a large unrealized gain in one account, it cannot be used directly to meet a margin call in another. This siloing of funds forces investors to hold large idle cash buffers. In the unified clearing network built by the three-layer stack, by contrast, the same stablecoin margin can support cross-asset positions at the same account level and, by exploiting negative correlations among assets, lower the overall margin requirement (the mechanism and its quantitative range are detailed in Section 32.1.4), giving the same principal far greater risk exposure and allocation flexibility than an isolated-account system. A structural vulnerability of this unified-margin architecture is single-collateral concentration risk. When all cross-asset positions are denominated in a single stablecoin, collateral value and margin adequacy become perfectly positively correlated. During the March 2023 episode in which USDC depegged to \$0.87, a cross-margining system denominated purely in USDC would face an approximately 13% decline in the margin adequacy ratio across all positions, triggering cascading liquidations unrelated to the performance of any underlying asset. It is precisely to avoid this concentration risk that traditional prime brokers accept a diversified basket of collateral.

32.3.4 Welfare analysis and distributional effects

When the friction of global capital allocation is minimized and efficiency rises sharply, we must ask: who exactly enjoys this efficiency dividend? Traditional financial innovation has often benefited Wall Street's large institutions first, but the welfare distribution brought by the autonomous finance network presents a strikingly different picture.

The greatest beneficiaries are not the multinational institutions that already occupy the top of the global financial system. For these institutions, complex legal structures, vast compliance teams, and expensive financial infrastructure have long since established dedicated channels for global capital allocation. For them, the three-layer stack merely lowers marginal transaction costs rather than providing a capability from scratch. Those who gain an enormous welfare improvement are the groups long excluded from global capital markets, especially retail investors and small and medium-sized enterprises in emerging markets [24]—as noted in Section 32.2.1, the roughly 1.3 billion unbanked people worldwide are concentrated primarily in developing countries.

For these groups, the autonomous finance network provides global access previously unavailable. In the past, they were trapped in their countries' fragile financial systems, forced to bear the twin blows of hyperinflation and currency depreciation yet unable to hedge effectively. Now, with only an internet-connected smartphone, they can obtain exposure to the world's best assets at an extremely low threshold, or use foreign-exchange hedging instruments once affordable only to multinational corporations. This equalization of access not only protects their wealth from the erosion of a local economic crisis but also, for the first time, gives them a chance to share in the dividends of global economic growth. This shift—from a privilege of the rich to a tool for everyone—is the core embodiment of financial inclusion at the level of capital allocation (the optimistic assessment in this section remains subject to the threefold constraints of financial literacy, the digital divide, and regulatory adaptation discussed in Section 32.7; the greatest beneficiaries are more likely to be middle-income groups that already possess digital access and financial awareness, rather than the most financially excluded at the very bottom).

Any major technological and institutional change, however, comes with creative destruction, and the rise of the autonomous finance network is no exception. The rise in efficiency and the elimination of friction will inevitably deal a heavy blow to the intermediaries of the traditional financial system. In a fragmented global market, vast profits arise precisely from these frictions. Cross-border remittance companies survive on steep fees, global custodian banks extract rents through complex settlement networks, and foreign-exchange market makers earn rich spreads in opaque quotes. Once the three-layer stack automates and makes transparent these functions through code and algorithms, the business models of these traditional intermediaries will face an existential challenge. Moreover, synthetic exposure that functionally bypasses capital controls may provoke a defensive reaction at the sovereign level. Countries that maintain capital-account controls may respond by banning stablecoin transactions or blocking on-chain financial infrastructure, which would nullify the inclusive gains their citizens should in theory receive. Welfare analysis must therefore incorporate a key variable: the very efficiency gains described in this chapter are highly likely to provoke

offsetting regulatory countermeasures.

This distributional effect reveals an inherent tension in the development of the autonomous finance network. On one hand, by raising the efficiency of global capital allocation, it creates enormous consumer surplus for the broad population of emerging-market users; on the other, it ruthlessly eliminates the rents earned by traditional financial intermediaries. The transition to this new paradigm is therefore bound to be far from smooth and will face fierce resistance from vested interests. Over a longer horizon, however, the tendency of technological progress to drive down transaction costs and raise capital-allocation efficiency has powerful momentum. When capital is no longer constrained by geographic and institutional barriers and can move to where expected returns are highest, the way the global economy operates will change profoundly.

Frictionless capital flow, however, means that panic, error, and liquidation can likewise propagate across borders with extremely low latency. Systemic risk has not disappeared; it has shifted from the balance sheets of centralized institutions into the topology and algorithmic logic of a decentralized network.

32.4 New systemic risks

Even as the three-layer stack eliminates traditional financial friction, it also removes the function of that friction as a risk buffer, thereby giving rise to entirely new forms of systemic risk. When all assets are interlinked through cross-margining, millions of AI agents share homogeneous strategies, and the oracle becomes the single data interface for the entire system, the global propagation of a local shock is measured in milliseconds. Starting from a globalized version of the reflexivity triangle, this section examines the global flash-crash risk induced by AI-strategy homogenization, the fragility of the oracle as a critical single point of failure, and the challenge that the new paradigm of *too interconnected to fail* poses to the theory of financial stability.

32.4.1 The globalization of the reflexivity triangle

In Chapter 20 we examined the reflexivity triangle operating within a single exchange or a single asset: a price decline raises volatility, higher volatility forces market makers and traders to reduce leverage, and the forced liquidations of the deleveraging process further depress the price. This positive-feedback loop is the microstructural basis of the extreme conditions that occur frequently in crypto markets. When all assets are interlinked through on-chain infrastructure, however, the transmission range of this reflexivity triangle undergoes a qualitative jump, expanding from *a single exchange* to *all markets worldwide*.

In the traditional financial system, natural firewalls exist between asset classes and geographic markets. Although sharp volatility in the U.S. stock market affects global sentiment, the isolation of settlement systems, differences in time zones, and the constraints of capital controls mean this effect typically takes several days to transmit fully to emerging-market foreign-exchange or commodity markets. In the three-layer stack of autonomous world finance, these physical and institutional frictions are wholly eliminated. More critically, if the on-chain system implements the global cross-margining

mechanism described in Chapter 29, the liquidation of one asset propagates instantly to all others through margin linkages.

Imagine a scenario: the local currency of an emerging-market country suffers a sudden depreciation, leaving many small-business owners who use foreign-exchange perpetual futures to hedge facing margin calls. Because these users hold global cross-margin accounts, the system automatically liquidates the other assets they hold to cover the losses on the FX position. This cross-asset passive selling drives down the prices of U.S. equity indices and gold, in turn triggering liquidations for other investors who hold only those assets. In this process, the surge in volatility prompts AI market makers worldwide to widen spreads or withdraw liquidity simultaneously, so that an originally tiny sell order inflicts an enormous price impact. The funding-rate mechanism unique to perpetual futures further amplifies this cascade: when crowded longs push the funding rate to an extreme level, the holding cost for longs rises sharply, forcing more longs to close out and exit, thereby forming a positive-feedback loop of *funding rate* $\rightarrow$ *holding cost* $\rightarrow$ *forced liquidation*—an amplification channel unique to perpetual futures that does not exist in traditional margin cascades. This is the ultimate, amplified version of the cross-dimensional contagion of Chapter 20: a global liquidation cascade. In such a state, no asset is a safe harbor, because the liquidity pools of all assets are interconnected on the same underlying infrastructure. This interconnection amplifies a minor tremor in a local market into a global systemic crisis whose destructive power far exceeds the contagion effects of traditional financial crises [25]. The equilibrium model for clearing obligations in networked financial systems established by Eisenberg and Noe (2001) [26] provides a starting framework for understanding this contagion mechanism, but its core assumptions (a known, fixed network topology and synchronous clearing) differ fundamentally from the on-chain perpetual futures ecosystem. On-chain systems have a dynamic, permissionless topology (new interconnections can appear at any time without central visibility), block-by-block asynchronous clearing (the clearing order is determined dynamically by MEV auctions), and a first-come-first-served rather than pro-rata loss-allocation mechanism. The E&N model therefore requires substantial theoretical extension along three dimensions—topological dynamism, asynchronous clearing, and loss-allocation rules—before it can apply to contagion analysis in a DeFi environment.

The pioneers of reflexivity theory long ago warned that financial markets do not always tend toward equilibrium but often fall into self-reinforcing states of disequilibrium. In traditional isolated markets, such a state of disequilibrium eventually ends when funds are exhausted or external intervention arrives. But in a globally interconnected autonomous finance network, although the depth and breadth of the capital pool increase, the paths of contagion also become infinitely many. The collapse of one asset is no longer an isolated event but becomes the starting node that triggers a chain of liquidations. This networked reflexivity makes the system appear highly resilient and efficient in normal times, yet under stress it may exhibit systemic fragility.

The foregoing scenario analysis shows that the contagion path of the reflexivity triangle in a globally interconnected environment is far more complex than in a single market: the three links of price decline, volatility surge, and liquidity withdrawal are no longer confined to a single asset or a single exchange but unfold synchronously across all asset

classes and geographic markets through the multilateral linkages of cross-margining (Figure 32-4).

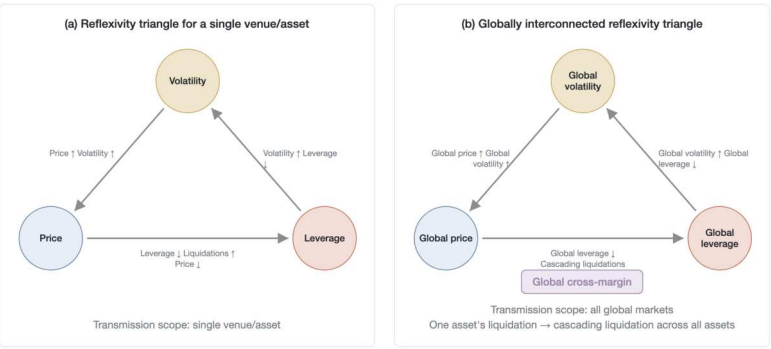


Figure 32-4. The globalized transmission mechanism of the reflexivity triangle (Data source: theoretical derivation in this chapter)

32.4.2 A global AI flash crash

Chapter 31 discussed the *strategy-homogenization* risk that arises when AI acts as a market participant. In a globally interconnected financial network, this risk will be amplified exponentially. As large language models and reinforcement-learning algorithms become widespread in finance, we can foresee that the overwhelming majority of autonomous trading agents worldwide will be driven by base models supplied by a few top technology companies. This means that although these agents serve different clients and have different risk preferences, the underlying logic by which they process information, assess risk, and execute trades will be highly similar.

If AI market-making agents worldwide use similar strategies, anomalous data or an extreme event in one region could trigger the synchronized reaction of AI agents worldwide within milliseconds. Consider the U.S. stock market *flash crash* of May 6, 2010, when high-frequency trading algorithms collectively withdrew liquidity in the face of uncertainty: the Dow Jones Industrial Average plunged 998.5 points (approximately 9%) intraday, of which, within about five minutes (2:42 to 2:47 p.m.), it fell nearly 600 additional points on top of the earlier decline, bringing the cumulative peak-to-trough drop for the day close to 1,000 points, before recovering most of that decline within about 20 minutes (by 3:07 p.m.) [27]. That event was confined to the U.S. stock market alone and was caused mainly by simple rule engines rather than advanced AI. The fundamental driver of the 2010 flash crash, however, was microstructure fragility—a single large sell order triggering a liquidity vacuum—rather than strategy homogenization. For the risk of an AI-driven synchronized crash, a more apt historical precedent is the 2007 quant factor crisis, in which strategy crowding among quantitative hedge funds produced synchronized position reversals and a liquidity spiral. Even imposing a mandatory *algorithmic diversity* requirement cannot fundamentally eliminate the herding effect, because all AI agents are trained on the same historical data, observe the

same on-chain signals (which are transparent by design), and optimize the same objective function. Herding is an emergent property of a shared information environment, not merely a technical flaw stemming from a shared code base. In autonomous world finance, the destructive power of a *global AI flash crash* would far exceed that of 2010.

In an extreme scenario, when a piece of breaking news is fed into the global agent network, a large number of AI agents may reach the conclusion *reduce risk exposure* within the same millisecond. They would simultaneously cancel buy orders and issue sell orders, driving the liquidity of all major global assets to zero in an instant. Because the matching speed of on-chain transactions is limited only by the blockchain's block-generation time and network latency, such a flash crash might last only seconds. Yet within those few seconds, the liquidity vacuum could cause prices to deviate sharply, triggering the liquidation logic hard-coded into smart contracts. Human traders would not even have time to understand what was happening, let alone intervene. By the time the system finally returns to equilibrium, the transfer and destruction of wealth are often already irreversible.

This machine-driven panic lays bare the fragility of market microstructure. The high-frequency trading arms race analyzed by Budish et al. (2015) [28] takes on a new dimension in the era of AI agents: when competition shifts from a millisecond-level speed advantage to the homogenization of model architecture and training data, the source of systemic risk shifts from the technical failure of an individual institution to the collective behavior pattern of the entire agent population. More worrying still, modern AI models are typically *black boxes* whose decision processes are hard for humans to understand fully. After a *global AI flash crash*, post-hoc analysis might take months, and identifying the exact trigger might become impossible. This lack of explainability further undermines the confidence of market participants, making the system's post-crisis recovery even more difficult. To address this risk, future autonomous finance systems may need to introduce *algorithmic diversity* requirements or distributed circuit-breaker mechanisms—but these would inevitably sacrifice a portion of efficiency.

32.4.3 Oracle single-point-of-failure risk

When we examined *perpetual futures on everything* in Chapter 27, we noted that the oracle is the sole bridge connecting off-chain real-world data to on-chain smart contracts. In a globally interconnected financial network, this *oracle problem* will evolve from a local technical challenge into a critical single point of failure for the entire system. The traditional financial system relies on distributed price-discovery and settlement networks: if the New York Stock Exchange's data feed fails, the Chicago Mercantile Exchange or the London Metal Exchange can continue to operate. In an on-chain perpetual futures system, however, if a widely used oracle network is attacked or fails, the consequences will be severe.

Suppose an oracle network that provides price feeds for trillions of dollars of assets worldwide broadcasts an erroneous price to the chain because of a software bug or a failure of its economic-incentive mechanism. In a traditional market, such an obviously erroneous trade would be caught and canceled by the exchange's human monitoring team. But in the on-chain world where code is law, smart contracts trust the oracle's

data unconditionally. This erroneous price would immediately trigger erroneous liquidations of all perpetual futures that depend on the oracle. Because of the global cross-margining mechanism described earlier, these erroneous liquidations would set off real chains of blow-ups, turning a data error into a genuine systemic collapse.

Although the industry is working to reduce this risk through multi-source oracles, decentralized oracle networks, and cryptographic-verification techniques based on zero-knowledge proofs, these measures can only mitigate the problem, not eliminate it [29]. Multi-source oracles may provide false redundancy: if most price sources share upstream dependencies (the majority of feeds ultimately derive from the same three or four centralized exchanges), then multi-source aggregation does not genuinely diversify risk. Moreover, oracle extractable value (OEV) creates a structural economic incentive for validators or MEV searchers to profit from liquidation events by manipulating the timing of oracle updates. As long as an on-chain system must rely on external data to execute financial contracts, the oracle will always be the weak link that can trigger a system-wide catastrophe. The potential payoff from hacking an oracle grows geometrically as the on-chain financial network expands in scale, which will attract the world's most capable malicious actors to search for every tiny vulnerability in the system. This asymmetric attack-defense game makes oracle security a persistent structural risk facing autonomous finance systems.

At a deeper level, oracles are not merely conduits that relay data; they in fact perform a function of *adjudicating truth*. When a dispute arises in the real world (for example, when the definition of an asset diverges or an exchange suspends trading under extreme conditions), the oracle must decide what data to broadcast to the chain. This adjudicative power often lacks sufficient transparency and accountability in decentralized systems. If the governance of an oracle network is controlled by a few interest groups, they could manipulate price feeds to reap enormous profits in on-chain derivatives markets. Oracle risk is therefore not merely technical but also political and economic.

32.4.4 Too interconnected to fail

Network effects are deeply double-edged. A globally interconnected autonomous finance network creates extremely high efficiency by eliminating intermediaries and geographic barriers, but it also creates the systemic fragility of *too interconnected to fail*. In the 2008 global financial crisis, the core dilemma regulators faced was *too big to fail*: the balance sheets of a few giant financial institutions were so large that their collapse would destroy the entire credit system. In autonomous world finance, the form of risk has shifted: an individual node may not be large, but because of the system's high interconnection, the failure of any single node may propagate instantly to all others.

This interconnection is manifested at multiple levels: cross-margining at the asset level, shared liquidity pools at the infrastructure level, and strategy homogenization at the intelligence level. When these levels interweave, the system's complexity exceeds the comprehension of any single participant or regulator. The *immediacy* and *around-the-clock continuity* of on-chain systems further worsen this problem. Traditional financial markets provide *cooling-off periods* through market closures and weekends, giving participants time to digest information, raise capital, and coordinate rescues. A global

network that never closes strips away this precious buffer.

When a crisis occurs, transmission speed is measured in milliseconds and seconds, and the window for human intervention is extremely compressed or eliminated altogether. We have built a financial machine more efficient and more inclusive than ever before, but it is also a machine that by design lacks the traditional physical damping. How to establish a new *shock-absorption mechanism* in a global network with no centralized intervener, no market closures, and driven entirely by code and algorithms will be a core challenge of the autonomous finance era. This shift in systemic risk requires us to rethink the definition of financial stability, moving from a focus on the health of a single institution to a focus on the resilience of the entire network topology.

Under the *too interconnected to fail* paradigm, traditional regulatory tools (such as capital-adequacy requirements or stress tests) may no longer apply, because these tools are typically based on static balance-sheet snapshots and cannot capture the complex nonlinear interactions of a high-frequency dynamic network. In the on-chain perpetual futures ecosystem, the risk nodes with high centrality are already clearly identifiable: core oracle networks (such as a few Chainlink/Pyth data sources), cross-chain-bridge liquidity hubs, and shared market-maker infrastructure. The failure of any one of these high-centrality nodes satisfies the contagion condition of *too interconnected to fail*, and its impact will spread along the network topology to all dependent nodes. Future financial-stability frameworks may need to draw on epidemiology or complex-systems theory, focusing on monitoring the density of connections among network nodes, identifying potential super-spreader nodes, and designing smart-contract circuit breakers that can automatically sever contagion chains when a local crisis erupts. Yet on-chain circuit breakers face a fundamental paradox: when one protocol halts trading while interconnected protocols keep running, the result is not orderly market stabilization but arbitrage dislocations and trapped positions. The governance of circuit-breaker parameters is itself an unresolved problem: DAO governance votes are too slow for millisecond-scale cascades, while protocol admin keys reintroduce the very centralization risk that permissionless design is meant to eliminate. This is an underexplored field, but it is vital to the long-term survival of autonomous finance.

One systemic-risk channel the foregoing analysis has not adequately addressed is the *funding monoculture* of dollar stablecoins. The three-layer stack is denominated in dollar stablecoins throughout, and so it natively inherits the transmission effects of Federal Reserve monetary policy. A rate-hiking cycle raises the opportunity cost of stablecoin capital while compressing the total value locked (TVL) and liquidity across all on-chain venues. Empirically, the most likely trigger of a systemic shock is not a cascade of emerging-market currency crises but a dollar-liquidity squeeze that hits all on-chain markets at once, because they share a single base quote currency. This stablecoin monoculture stands in deep tension with the *sovereign plurality* narrative of Section 32.5.3: a financial system that claims to serve a diverse set of global sovereigns nonetheless depends, at the funding layer, entirely on the monetary policy of a single central bank.

The three risk dimensions analyzed in Sections 32.4.1 through 32.4.3—cross-margining cascades, AI flash crashes, and oracle failures—are not mutually independent events.

If an oracle price deviation triggers a cross-margining cascade while AI market makers synchronously withdraw liquidity upon detecting the price anomaly, the joint activation of the three risk vectors will produce a shock-amplification effect far exceeding the sum of each analyzed independently. This *risk resonance* scenario represents the system's true worst case and is also the extreme scenario a stress-testing framework must cover.

The foregoing analysis of systemic risk reveals a deeper structural problem: the three-layer stack contains three irreducible tensions that must be continuously managed rather than eliminated once and for all by technical means.

Figure 32-5 visualizes these three tensions as three pairs of mutually constraining forces—efficiency versus stability, permissionlessness versus consumer protection, and global integration versus sovereign plurality—and indicates the direction in which each can be addressed: gains in efficiency inevitably bring new fragilities; ZKP credentials can partly reconcile the conflict between permissionlessness and consumer protection but cannot fully eliminate the tension; and in the contest between global integration and sovereign plurality, states choose among embracing, restricting, and vacillating. Section 32.5 develops each of these three trade-offs in turn.

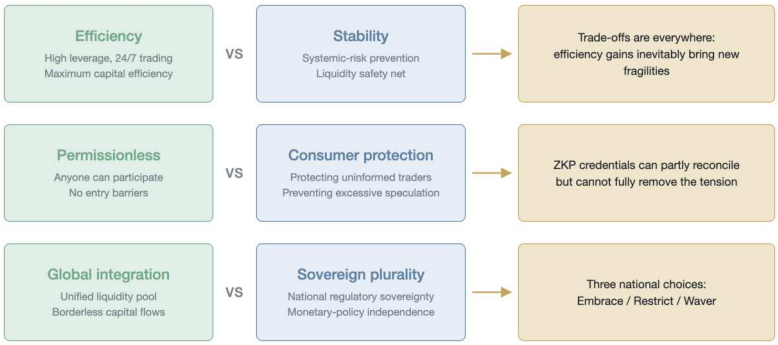


Figure 32-5. The three irreducible tensions in an autonomous finance network (Data source: theoretical derivation in this chapter)

32.5 Three irreducible tensions

The analysis of the two preceding sections shows that the efficiency gains and the systemic risks of the three-layer stack are not independent variables but the two faces of the same set of design choices. This inherent tension can be summarized as three irreducible structural trade-offs: efficiency versus stability, permissionlessness versus consumer protection, and global integration versus sovereign plurality. This section works through the concrete manifestations and amplification mechanisms of each tension in an autonomous finance network, and argues that treating these tensions as fundamental trade-offs requiring continuous management—rather than as engineering defects awaiting repair—is the epistemological premise for understanding autonomous finance.

32.5.1 Efficiency versus stability

The core function of a financial system is capital allocation, and the efficiency of capital allocation is often negatively correlated with the system's stability. In an autonomous finance network, this trade-off can be measured conceptually through two sets of indicators: efficiency indicators (capital-utilization efficiency and price-discovery latency) and stability indicators (tail-risk probability and liquidation-cascade depth), between which a structural negative correlation exists. This chapter offers a conceptual framework for identifying the structure of the trade-off rather than a formal mathematical model; a full formal treatment—such as a constrained optimization linking capital-utilization efficiency to the probability of systemic tail risk—awaits future research. An increase in leverage raises capital-utilization efficiency but necessarily entails a higher probability of liquidation cascades [30]; around-the-clock continuous pricing shortens price-discovery latency but removes the physical circuit-breaking window that traditional markets use to relieve panic. As the global liquidation-cascade mechanism analyzed in Section 32.4.1 shows, cross-asset cross-margining transmits the volatility of a single asset to the entire network. The crypto-market liquidation storm of October 2025—in which more than $19 billion in leveraged positions were forcibly liquidated within roughly 24 hours, the largest single-day liquidation on record (this is measured as liquidation volume, distinct from the contraction in open interest over the same period)—validated this transmission path [31]. The AI-strategy-homogenization risk analyzed in Section 32.4.2 further intensifies this tension: when large numbers of agents adopt homogeneous strategies, their synchronized reaction patterns amplify market volatility to the scale of a millisecond-level flash crash [32][33]. Every gain in efficiency accumulates deeper fragility in the system's foundations; this trade-off is a structural constraint of financial systems that cannot be circumvented by a more advanced technical architecture.

32.5.2 Permissionlessness versus consumer protection

Permissionless participation is the cornerstone of decentralized finance, but an irreconcilable tension exists between full permissionlessness and adequate consumer protection. This tension can be formalized as a trilemma: fully permissionless access, adequate consumer protection, and a deep liquidity network—at most two of which can be satisfied simultaneously. Abandoning access restrictions exposes participants who lack financial literacy to risks they cannot assess (the structural constraint analyzed in Section 32.2.5) [34][35]; imposing access screening reintroduces centralized gatekeepers and excludes the very marginalized populations who need these tools; and abandoning the first two in pursuit of liquidity alone produces an unregulated, highly leveraged casino. This structure is isomorphic to the Mundell-Fleming trilemma in international finance [36]: just as a country cannot simultaneously achieve free capital flows, an independent monetary policy, and a fixed exchange rate, an autonomous finance network cannot simultaneously maximize openness of access, user protection, and market depth. Zero-knowledge-proof compliance credentials mitigate this conflict to some degree by lowering the cost of compliance enforcement [37], but proving that one meets the conditions is itself a permissioning mechanism. As long as the system needs to distinguish participants' eligibility, this structural tension will not disappear.

32.5.3 Global integration versus sovereign plurality

A fundamental conflict exists between the borderlessness of the autonomous finance network and the territoriality of sovereign states. At the network level, protocol standards on public blockchains and decentralized clearing and settlement connect global capital pools into a single liquidity network [38]; at the sovereign level, states rely on exclusive control over domestic currency issuance, capital flows, and financial institutions to implement monetary policy [39]. When citizens can freely convert their local currency into global stablecoins through the autonomous finance network and participate in on-chain trading, the effectiveness of traditional capital-control tools is substantially weakened. This conflict is driving a divergence of national strategies: some countries choose to embrace, obtaining the dividends of innovation by offering a friendly regulatory environment and exploring central bank digital currencies (CBDCs); some choose to isolate, seeking to exclude on-chain finance from the national economic system; and more countries seek a conditional balance between access to global liquidity and the preservation of policy independence. This divergence is not a temporary policy disagreement but the inevitable projection of the structural conflict between the network's global integration and the territorial exclusivity of sovereignty, and it will continue to shape the evolution of the international financial system. The dominance of USD stablecoins, moreover, grants the United States de facto extraterritorial regulatory authority over the entire on-chain financial system. Binance's $4.3 billion settlement and the sanctions against Tornado Cash show that even citizens of countries pursuing an *embrace* strategy remain subject to major-power enforcement jurisdiction when their transactions involve USD stablecoins, which makes the independence of all three strategies far more limited than it appears.

32.5.4 The epistemology of tension management

A mature theory of financial microstructure requires the recognition that the three tensions above represent fundamental trade-offs, not system bugs awaiting repair. Efficiency versus stability, permissionlessness versus consumer protection, and global integration versus sovereign plurality—these three pairs of opposites constitute the boundary conditions of the autonomous finance network. Different jurisdictions and groups of market participants will position themselves differently within these trade-offs according to their own value orientations: developed economies tend to sacrifice some efficiency in exchange for systemic stability, while emerging markets are more willing to bear volatility risk in exchange for access to global capital and the inclusiveness of financial instruments. This diversity is the inevitable projection of pluralistic values within a complex system. Recognizing that these trade-offs are irreducible marks a cognitive shift from techno-utopianism to realist financial engineering.

The maturity of the three-layer stack lies not in eliminating these contradictions but in establishing mechanisms with sufficient resilience to maintain a dynamic equilibrium within these long-term structural tensions. This equilibrium is continually recalibrated as technology evolves, market cycles turn, and the geopolitical environment shifts.

Under the constraint of the three tensions, the future of autonomous finance is not a fixed endgame but a complex adaptive system that requires continuous design and

adjustment.

32.6 From vision to roadmap

Granting the premise of three irreducible tensions, the evolutionary path of autonomous finance is not unpredictable. Depending on different combinations of regulatory attitude, technological maturity, and the speed of market adoption, one can sketch three sharply different scenarios: gradual convergence, dual-track coexistence, and disruptive substitution. This section analyzes the driving conditions and possible outcomes of each path, identifies the key uncertainty variables that will determine the ultimate direction, and proposes *responsible optimism* as a basic stance for coping with these uncertainties.

32.6.1 Evolutionary paths

In examining the future of autonomous finance, we cannot give a deterministic timetable or a single endgame form. These three paths are not mutually exclusive; they may play out simultaneously in different jurisdictions or across different asset classes.

Path one is gradual convergence. Along this path, the boundary between traditional finance and on-chain finance gradually blurs, ultimately forming a hybrid global financial ecosystem. Large traditional financial institutions begin to recognize the enormous advantages of on-chain infrastructure in clearing-and-settlement efficiency, capital allocation, and around-the-clock trading, and progressively adopt these technologies. They may first pilot on-chain clearing engines in internal operations or specific B2B scenarios, then extend them to customer-facing businesses. At the same time, the regulatory framework gradually adapts, introducing privacy-preserving technologies such as zero-knowledge proofs to preserve a degree of permissionlessness while satisfying anti-money-laundering and know-your-customer requirements. In this scenario, the three-layer stack of autonomous finance (the asset layer, the infrastructure layer, and the intelligence layer) is absorbed and reshaped by the traditional financial system, becoming a new engine for raising global financial efficiency. The winners along this path include both large institutions that actively embrace innovation and emerging-market retail investors who receive better service through financial inclusion.

Path two is dual-track coexistence. In this scenario, on-chain finance and traditional finance coexist over the long term but remain relatively independent, serving different user groups and needs. Because of divergent regulatory policies and the difficulty of unifying technical standards, a clear barrier forms between the two financial systems. The traditional financial system continues to dominate mortgages, corporate financing, and large-scale asset management, while on-chain finance dominates crypto-native assets, high-frequency trading, and parts of emerging markets. This dual-track state produces a degree of market segmentation, and capital-allocation efficiency fails to reach its maximum. It also, however, provides a parallel financial network for the users excluded by the traditional system (such as the billions of unbanked people worldwide). In this parallel world, AI agents autonomously make markets and trade, and permissionless smart contracts manage perpetual futures on a variety of synthetic assets. This state

may persist for decades, until one side achieves a decisive breakthrough in technology or regulation.

Path three is disruptive substitution. In certain specific niche markets, on-chain finance, by virtue of its overwhelming advantages in efficiency and cost, achieves disruptive substitution of traditional finance. In derivatives trading, for example, on-chain perpetual futures—by virtue of around-the-clock continuous pricing, frictionless global capital flow, and cross-margining—rapidly absorb global liquidity, forcing traditional derivatives exchanges to the margins. In cross-border payments and remittances, solutions based on stablecoins and on-chain clearing sharply compress transaction costs, disrupting traditional correspondent-banking networks and remittance companies. In fields such as mortgages and large corporate financing, however—which depend heavily on offline credit assessment and complex legal enforcement—traditional finance retains an unshakable position. The evolution of this path is often accompanied by violent market turbulence and regulatory conflict; traditional financial institutions and vested interests may resist fiercely, but ultimately they must accept the reality that part of the market has been disrupted.

Each of these three paths corresponds to a different regulation-technology-market combination, and its ultimate direction is not predetermined but depends on how key variables actually evolve. Table 32-2 offers a structured comparison of the driving conditions, primary beneficiaries, and core risk of the three paths, providing a reference for assessing investment strategies and policy choices under different scenarios.

Path	Driving conditions	Primary beneficiaries	Core risk
Gradual convergence	Strong regulatory adaptability, active institutional adoption	Large institutions that embrace innovation, and emerging-market retail investors	Pace of innovation constrained by the pace of regulation
Dual-track coexistence	Regulatory divergence, difficulty unifying technical standards	The unbanked excluded by the traditional system	Market fragmentation leaves capital-allocation efficiency below its maximum
Disruptive substitution	Overwhelming on-chain efficiency advantage, breakthroughs in specific niche markets	Derivatives and cross-border-payment users	Severe market turbulence and resistance from incumbents

Table 32-2. Comparison of the three evolutionary paths (Data source: based on the analysis in Section 32.6.1)

Figure 32-6 shows the relative changes in the market share of traditional finance and on-chain finance under different scenarios from 2025 to 2040, as well as the evolution of a regulatory-adaptation-or-conflict index.

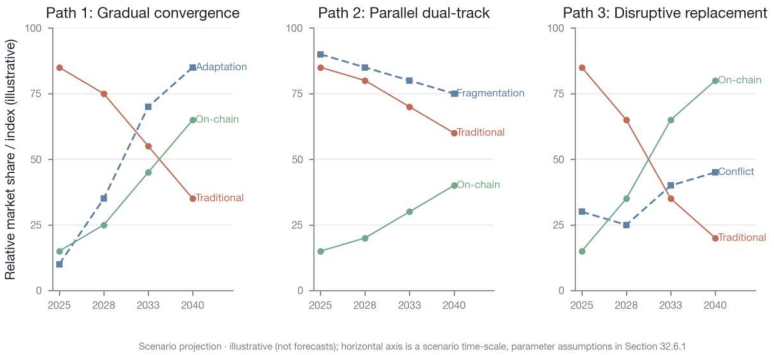


Figure 32-6. Scenario analysis of the three evolutionary paths of autonomous finance, 2025–2040 (Data source: scenario projection based on three variables—regulatory attitude, technological maturity, and market adoption speed; parameter assumptions detailed in the main text of Section 32.6.1)

32.6.2 Key uncertainties

Whichever path autonomous finance ultimately follows, its progress is deeply affected by several key uncertainties. These uncertainties constitute the blind spots in our predictions of the future and are also the variables we must focus on when building autonomous finance systems.

Regulation is among the most decisive of these uncertainties. A fundamental conflict exists between the global, borderless nature of autonomous finance and the sovereignty-based regulatory frameworks of traditional finance. The attitudes of different countries and regions toward on-chain finance will determine its room for development. Some countries may adopt an innovation-embracing attitude, seeking to attract capital and technology by establishing adaptive regulatory sandboxes; others may, out of concern for safeguarding financial sovereignty and preventing systemic risk, adopt severe restrictions or even outright bans. The vacillation or sudden reversal of regulatory policy can change the liquidity distribution and market structure of on-chain finance in an instant. How to find a balance point within the tensions of *permissionlessness* versus *consumer protection* and *global integration* versus *sovereign plurality* is an enormous challenge facing regulators.

Technological maturity is a second decisive variable. Although the three-layer stack of autonomous finance is conceptually relatively complete, in actual engineering it still faces many technical bottlenecks. The application of zero-knowledge-proof technology in embedded compliance remains at an early stage, and its computational efficiency and cost have not yet reached the requirements for large-scale commercial use. Cross-chain interoperability is still a fragile link; security vulnerabilities in cross-chain bridges have repeatedly caused enormous losses of funds, severely damaging market confidence. In addition, the widespread application of AI in finance brings the challenge of explainability. When millions of AI agents autonomously conduct high-frequency trading in a global network, the logic of their behavior is often a black box. Once a *global AI flash crash* occurs, it is difficult to trace its root cause quickly and take effective intervention measures. The pace at which these technical problems are solved will directly determine the speed at which autonomous finance matures.

The speed of market adoption and the tipping point of network effects are another key factor. The core value of a financial network lies in its liquidity, and liquidity in turn depends heavily on network effects. If on-chain finance cannot cross the adoption tipping point within a reasonable time and attract enough institutional capital and retail participation, it may remain forever at the margin of a niche market. Conversely, once this tipping point is crossed, network effects amplify exponentially, accelerating its substitution for or convergence with traditional finance. This adoption speed is not linear, however; it may be constrained by user habits, education costs, and switching costs. As noted earlier, pushing users who lack risk-management capacity into highly leveraged markets is not inclusion but a transfer of risk. If users lack the financial literacy needed to understand the risks of on-chain finance, then an overly rapid pace of adoption may instead cause large-scale wealth losses, provoking a social backlash and forceful regulatory intervention.

Finally, black swan events may become catalysts that either accelerate or impede the development of autonomous finance. The next major global financial crisis could expose the fragility of the traditional financial system, prompting a turn toward a more transparent, decentralized on-chain financial system; but by the same token, if the crisis is triggered by the systemic risks of on-chain finance (such as a global liquidation cascade caused by an oracle failure), it could destroy public trust in the new paradigm and set its development back by decades. The unpredictability of black swan events requires that, when designing autonomous finance systems, we treat antifragility as one of the core principles.

32.6.3 Responsible optimism

In the face of the evolutionary paths and key uncertainties above, this chapter takes the stance of *responsible optimism*. The three-layer stack provides new engineering possibilities, but technology itself does not determine outcomes; the same tools can become either an inclusive network or an instrument of regulatory arbitrage. The ultimate direction depends on choices of institutional design: whether education, risk control, and compliance can be embedded in the system's underlying logic determines whether autonomous finance empowers or preys.

Acknowledging systemic risk and irreducible tensions does not mean denying the potential value of autonomous finance. The builders of autonomous finance must seek a conditionally optimal path under the constraint of the three tensions, which requires combining the interdisciplinary wisdom of cryptography, financial engineering, economics, and law. More importantly, the evolutionary details of this complex adaptive system cannot be precisely predicted by anyone, so mechanisms for continuous iteration and agile adjustment are more valuable than a one-time perfect design.

32.7 Chapter summary

Starting from the architectural logic of the three-layer stack, this chapter analyzed the impact of autonomous finance along three dimensions—financial inclusion, global capital allocation, and systemic risk—and identified three irreducible structural tensions:

efficiency versus stability, permissionlessness versus consumer protection, and global integration versus sovereign plurality. Among the three possible evolutionary paths—gradual convergence, dual-track coexistence, and disruptive substitution—the ultimate form realized will depend on the interplay of regulatory attitude, technological maturity, and the speed of market adoption. The core conclusion of this chapter is that the three-layer stack possesses the engineering capacity to raise capital-allocation efficiency and expand the accessibility of financial instruments at the same time, but translating this capacity into genuine inclusion is limited by three constraints: financial literacy, the digital divide, and regulatory adaptation. Until these constraints are adequately addressed, the main beneficiaries of the three-layer stack will be middle-income groups that already possess digital access and basic financial awareness, rather than the most financially excluded at the very bottom. The value of autonomous finance lies not in building a perfect, contradiction-free system but in creating a governance mechanism capable of dynamically seeking the optimum within long-term tensions.

The next chapter turns from this chapter's forward-looking analysis to the theoretical level, systematically drawing together the analysis of the whole book—distilling its core propositions and a spectrum of confidence levels, clarifying unresolved questions and a research agenda, and offering the book's recommended courses of action for different groups of readers.

References

[1] CoinGecko. (2026). Is the Future of Crypto Perpetual? The Meteoric Rise of Perp DEXs. https://www.coingecko.com/learn/rise-of-perpetuals-and-perp-dexs

[2] World Bank. (2025). *The Global Findex Database 2025*. https://www.worldbank.org/en/publication/globalfindex

[3] World Bank. (2025). *Financial Inclusion Overview*. https://www.worldbank.org/en/topic/financialinclusion/overview

[4] World Bank. (2022). *Remittance Prices Worldwide* (Issue 44, December 2022). https://remittanceprices.worldbank.org/

[5] United Nations Development Programme. (2024). *UNDP annual report 2023*. https://www.undp.org/publications/undp-annual-report-2023

[6] World Bank. (2024, July 1). World Bank country classifications by income level for 2024-2025. *World Bank Blogs*. https://blogs.worldbank.org/en/opendata/world-bank-country-classifications-by-income-level-for-2024-2025

[7] International Telecommunication Union. (2025). *Facts and figures 2025: Measuring digital development*. https://www.itu.int/itu-d/reports/statistics/facts-figures-2025/

[8] GSMA. (2025). *GSMA calls for renewed focus on closing the Usage Gap as more than 3 billion people remain offline*. https://www.gsma.com/newsroom/press-release/gsma-calls-for-renewed-focus-on-closing-the-usage-gap/

[9] Schär, F. (2021). Decentralized finance: On blockchain- and smart contract-based financial markets. *Federal Reserve Bank of St. Louis Review*, 103(2), 153-174.

[10] McKinsey & Company. (2025). The State of AI: Global Survey 2025. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[11] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, 22(6), 2201-2238.

[12] Forbes, K. J., & Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance*, 57(5), 2223-2261. https://doi.org/10.1111/0022-1082.00494

[13] Statista. (2024). *Annual inflation rate of the Consumer Price Index in emerging markets and developing economies from 1980 to 2030.*

[14] World Bank. (2024). *Remittance Prices Worldwide, Issue 49.* https://remittanceprices.worldbank.org/

[15] International Monetary Fund. (2021). *A Smorgasbord of Factors Drive Remittance Costs.* https://www.elibrary.imf.org/view/journals/001/2021/199/article-A001-en.xml

[16] Adams, A., Lader, M-C., Liao, G., Puth, D., & Wan, X. (2023). On-chain foreign exchange and cross-border payments. *SSRN Working Paper No. 4328948.* https://ssrn.com/abstract=4328948 https://doi.org/10.2139/ssrn.4328948

[17] Rajan, R. G. (2006). Has finance made the world riskier? *European Financial Management*, 12(4), 499-533.

[18] NerdWallet. (2026). *How Much Does a Financial Advisor Cost in 2026?* https://www.nerdwallet.com/financial-advisors/learn/how-much-does-a-financial-advisor-cost

[19] Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1), 261-292. https://doi.org/10.1162/003355301556400

[20] Aitken, M., Chen, H., & Foley, S. (2017). The impact of fragmentation, exchange fees and liquidity provision on market quality. *Journal of Empirical Finance*, 41, 140-160. https://doi.org/10.1016/j.jempfin.2016.10.002

[21] BlackRock. (2025). *Market Spotlight: 24 Hour Trading - Implications for markets and investors.* https://www.blackrock.com/corporate/literature/whitepaper/blackrock-market-spotlight-24-hour-trading.pdf

[22] Ferreira, M. A., Massa, M., & Matos, P. (2010). Shareholders at the gate? Institutional investors and cross-border mergers and acquisitions. *The Review of Financial Studies*, 23(2), 601-644. https://doi.org/10.1093/rfs/hhp070

[23] Brunnermeier, M. K., De Gregorio, J., & Shin, H. S. (2012). *Banks and cross-border capital flows: Policy challenges and regulatory responses.* Committee on International Economic Policy and Reform.

[24] Koepke, R. (2019). What drives capital flows to emerging markets? A survey of the empirical literature. *Journal of Economic Surveys*, 33(2), 516-540. https://doi.org/10.1111/joes.12273

[25] Mieg, H. A. (2022). Volatility as a transmitter of systemic risk: Is there a structural

risk in finance?. *Risk Analysis*, 42(10), 2244-2258. https://doi.org/10.1111/risa.13564

[26] Eisenberg, L., & Noe, T. H. (2001). Systemic risk in financial systems. *Management Science*, 47(2), 236-249. https://doi.org/10.1287/mnsc.47.2.236.9835

[27] CFTC & SEC. (2010). *Findings Regarding the Market Events of May 6, 2010*. U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission.

[28] Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, 130(4), 1547-1621. https://doi.org/10.1093/qje/qjv027

[29] Albizri, A., & Appelbaum, D. (2021). Trust but verify: The oracle paradox of blockchain smart contracts. *Journal of Information Systems*, 35(2), 1-16. https://doi.org/10.2308/isys-19-024

[30] Wagner, W. (2011). Systemic liquidation risk and the diversity–diversification trade-off. *The Journal of Finance*, 66(3), 1141-1175. https://doi.org/10.1111/j.1540-6261.2011.01666.x

[31] FTI Consulting. (2025, December 24). *The crypto crash of October 2025: When leverage met liquidity*. https://www.fticonsulting.com/insights/articles/crypto-crash-october-2025-leverage-met-liquidity

[32] Oyeniyi, L. D., Ugochukwu, C. E., & others. (2024). Analyzing the impact of algorithmic trading on stock market behavior: A comprehensive review. *World Journal of Advanced Research and Reviews*, 21(1), 123-135.

[33] Asim, M., Khan, M. Y., & Shafi, K. (2024). Investigation of herding behavior using machine learning models. *Review of Behavioral Finance*, 16(3), 424-445. https://doi.org/10.1108/RBF-05-2023-0128

[34] Zetzsche, D. A., Arner, D. W., & Buckley, R. P. (2020). Decentralized finance. *Journal of Financial Regulation*, 6(2), 172-203. https://doi.org/10.1093/jfr/fjaa010

[35] OECD. (2025). *Improving the digital financial literacy of crypto-asset users*. OECD Publishing. https://www.oecd.org/en/publications/improving-the-digital-financial-literacy-of-crypto-asset-users_19cfecad-en/full-report.html

[36] Obstfeld, M., & Taylor, A. M. (2004). *Global capital markets: Integration, crisis, and growth*. Cambridge University Press.

[37] Yuyama, T., Katayama, K., & Brigner, P. (2023). Proposal of principles of DeFi disclosure and regulation. In *International Conference on Financial Cryptography and Data Security* . Springer.

[38] BVNK. (2025). *Blockchain in cross-border payments: a complete 2025 guide*. https://bvnk.com/blog/blockchain-cross-border-payments

[39] Klagge, B., & Martin, R. (2005). Decentralized versus centralized financial systems: is there a case for local capital markets? *Journal of Economic Geography*, 5(4), 387-421. https://doi.org/10.1093/jeg/lbh050

Summary and Research Outlook

In 2016, Arthur Hayes launched perpetual futures on BitMEX—a derivative contract that strips away the expiration date and uses the funding rate in place of delivery to anchor to the spot price. This design echoed the academic vision of macro-market instruments that Robert Shiller (1993) [1] had proposed, though that original proposal was never implemented in traditional finance. By 2025, nine years later, average daily notional trading volume in perpetual futures had settled in the hundreds of billions of dollars (according to CoinGlass's *2025 Crypto Derivatives Market Annual Report*, average daily notional turnover across crypto derivatives that year was roughly $260 billion, of which perpetual futures accounted for the overwhelming majority) [2]; they contributed more than 60% of the information share in Bitcoin's price discovery [3]; and derivatives turnover ran at roughly 3 times or more that of the corresponding spot market (reaching 5 to 10 times for major coins on a single leading exchange) [2]. This stands in sharp contrast to the early market documented a decade earlier by Makarov and Schoar (2020), which was characterized by cross-exchange spot arbitrage and price dislocations [4]. What began as a design compromise to solve practical problems in crypto trading has become one of the highest-volume categories of financial product in the world, and it is now expanding into foreign-exchange, equity-index, and commodity markets. From the perspective of microstructure theory, what makes this evolution significant is that a synthetic-exposure instrument that removed the constraints of expiry and delivery reshaped, within just a few years, where price discovery occurs, how liquidity is supplied, and how volatility is generated. This fact raises the central question of this chapter: after 33 chapters of systematic analysis of this mechanism and the markets in which it operates, which reliable regularities has the analysis ultimately revealed, and which unresolved problems does it leave behind?

This book takes institutional design as its through-line and begins with the game-theoretic mechanism of the funding rate. The funding rate is not a simple price-correction term but an institutional device that recasts price convergence as a problem of incentive compatibility, letting market participants' self-interested behavior anchor the price spontaneously, without intervention by any central authority. Over nine years of market operation, this mechanism has been tested by repeated episodes of extreme conditions, and its effectiveness has become a high-confidence empirical fact.

In the analysis of liquidity, the market maker's profit equation contains a structural flaw: adverse-selection risk and inventory costs rise in tandem as market volatility intensifies [5], so the production cost of liquidity is positively correlated with demand for it, leaving liquidity scarcest precisely when it is most needed. This countercyclical

fragility is a necessary consequence of profit maximization, not a behavioral bias. Along the volatility dimension, high-leverage liquidation cascades and the periodic pulses of the funding rate—beyond the traditional categories of informational and noise volatility [6]—create an institution-driven source of volatility that is independent of both fundamental information and trader irrationality [7].

Building on this, the book reframes market quality as a trade-off problem across five dimensions—transaction costs, price discovery, resilience, fairness, and accessibility—and argues that this trade-off presents a markedly different priority structure under normal, stressed, and crisis conditions. At the engineering level, trust decomposes into five independent layers: asset custody, matching and execution, clearing and settlement, data feeds, and governance decisions. Each layer can independently seek the optimal balance between trust-minimization and performance (Chapters 28–29). The evolution of artificial intelligence (AI) from a market participant into infrastructure marks a paradigm shift from "code is law" to "model is law." The methodological stance running through the analysis is that institutions determine microstructure, trade-offs are everywhere, and normal and crisis conditions follow different dynamics.

As the culmination of the book, this chapter distills the entire argument into seven core propositions and five cross-cutting themes. It uses a confidence spectrum to distinguish which conclusions have been cross-validated by multiple methods, which are supported only by preliminary evidence, and which remain logical inference or open conjecture, and it explicitly lays out the knowledge gaps and unresolved problems in the current theoretical framework. Beyond this, from the standpoint of disciplinary contribution, the chapter identifies five epistemological legacies that perpetual futures, as an object of analysis, bequeath to the theory of financial microstructure: making the institution-driven source of volatility explicit, a theory of the endogenous fragility of liquidity supply, the engineering of trust, a theory of 24/7 continuous markets, and the economics of synthetic exposure. It then maps out paths forward for researchers, practitioners, regulators, and students. After reading this chapter, the reader will possess a complete cognitive map distinguishing the known, the credible, the speculative, and the unknown, and can turn to Section 33.6 to find the path of action suited to their own role.

33.1 Core propositions

This section sets out the book's seven core propositions. Each proposition states its central conclusion and logical basis, and points to the chapters that support it. Section 33.2 presents the cross-cutting methodological perspective, and Section 33.5 sets out how the book advances the existing literature on financial microstructure.

In essence, perpetual futures are a general-purpose mechanism for generating synthetic exposure, not one confined to crypto assets. Traditional derivatives are constrained by the need for physical delivery and by expiration dates, which slice the time dimension into discrete contract cycles. Perpetual futures strip away both of these constraints and use the funding rate to maintain long-run convergence between the contract price and the spot price, providing traders with pure, open-ended price exposure [7]. This mechanism internalizes physical delivery and storage costs as a dynamic holding fee,

giving market participants synthetic exposure to any continuously priceable asset under a transparent cost structure. Its generality means that, wherever a reliable price-signal source and sufficient liquidity depth exist, perpetual futures can extend to domains such as foreign exchange, commodities, and macroeconomic indicators (Chapters 2–3 and Chapter 27). Contracts for difference (CFDs) and total return swaps in traditional finance likewise provide synthetic exposure, but they are limited by bilateral counterparty risk and finite market depth [8]. Perpetual futures reduce counterparty concentration through on-chain clearing and a multilateral funding-rate mechanism. Nevertheless, the collapse of Terra/LUNA in May 2022 shows that when the anchoring mechanism degenerates from an external spot price into a self-referential, endogenous loop, synthetic exposure can suffer catastrophic depegging. In that episode, the combined market capitalization of LUNA and UST—roughly $40 billion, of which UST accounted for about $18 billion—evaporated within about a week; this is the loss in that ecosystem's market capitalization, which is distinct from the whole-market loss figure the media often cite. The current confinement of perpetual futures to a few dominant asset classes stems more from regulatory constraints, liquidity depth, and path dependence than from the mechanism itself; but, as the depegging case above shows, that mechanism also has its boundaries once the external anchor degenerates into an endogenous loop.

The funding rate is an incentive-compatible device for achieving price anchoring in a decentralized environment, and its design logic is of a piece with the mechanism-design theory of Myerson (1981) [9] and Hurwicz (1972) [10]: in the absence of a central clearinghouse, the funding rate dynamically adjusts the holding costs borne by longs and shorts, so that price convergence emerges from participants' individual profit-maximizing behavior. When the contract trades at a premium, longs pay shorts; when it trades at a discount, the reverse holds, incentivizing arbitrageurs to eliminate price deviations of their own accord. At the same time, this periodic payment mechanism introduces a pulse effect into the microstructure: the mechanical repositioning around settlement times constitutes a source of price volatility independent of information flow. Different participant types respond to the rate in different ways—arbitrageurs are highly sensitive, trend traders relatively insensitive—giving the system a self-regulating capacity that maintains price anchoring without any central decision-maker. Budish et al. (2015) [11], in their analysis of the high-frequency trading arms race, offer a complementary perspective: frequent batch auctions, as an alternative market-design mechanism, likewise aim to achieve a superior market equilibrium by altering the structure of the game (Chapter 10).

Liquidity is not an inherent property that markets possess by nature; it is a scarce resource produced by market makers who bear adverse-selection risk and inventory costs. The market maker's profit equation contains a structural flaw: production cost is positively correlated with market demand [5]. As market volatility intensifies, adverse-selection risk and inventory risk rise together, forcing market makers to widen spreads or withdraw depth, so that the supply of liquidity dries up rapidly at the very peak of demand. This countercyclical fragility is not a behavioral bias on the part of market makers but a structural result of profit maximization under the assumptions of standard market-making models. In crypto markets, which lack a lender of last resort and

circuit-breaker protections, the collective retreat of market makers can readily trigger liquidation cascades and liquidity spirals [12] [13], amplifying a micro shock into a systemic crisis. The "Black Thursday" event of March 12, 2020, is a textbook case of this mechanism: over the two trading days of March 12–13, Bitcoin fell by roughly 50% cumulatively; on BitMEX alone, liquidations during this crash exceeded $700 million (on a broader, roughly 24-hour rolling-window basis, Chapter 12 gives a figure of more than $1.4 billion; the two figures use different statistical conventions), while bid-side order-book depth shrank to about $20 million at the most extreme moment, and liquidity evaporated far faster than participants could replenish it (Chapters 19–21).

Volatility is not a single risk metric but a mixture of four distinct drivers: informational volatility arises from fundamental updates; liquidity volatility arises from insufficient order-book depth; mechanistic volatility arises from liquidation cascades and funding-rate pulses; and transmission volatility arises from the amplification of local shocks by cross-market arbitrage networks [14]. The first two have been studied extensively by traditional microstructure theory; the latter two are products of the institutional design of perpetual futures and are independent of external information shocks [7]. Distinguishing the four sources of volatility bears directly on assessments of market quality and on regulatory decisions: a price move of the same magnitude, if driven by information updates, indicates that the market is efficiently digesting new information; if dominated by liquidation cascades, it points instead to a defect in system design (Chapters 22–24).

Market quality is a five-dimensional composite of transaction costs, price discovery, resilience, fairness, and accessibility, and inherent trade-offs exist among these dimensions [15]. Extremely low transaction costs may sacrifice market resilience, and an absolute insistence on fairness may constrain market-maker participation; no design achieves the optimum on all dimensions at once. These trade-offs are also state-dependent: the priority structure of market quality differs markedly across the three operating states of normal, stressed, and crisis conditions. A design that pursues efficiency under normal conditions may become a source of fragility in a crisis. Robust market design must therefore take the worst-case scenario as its binding constraint, ensuring that the system can still maintain its core functions under extreme shocks (Chapter 25).

Trust can be decomposed into five independent layers: asset custody, matching and execution, clearing and settlement, data feeds, and governance decisions. A traditional centralized exchange concentrates all five layers of trust in a single architecture, which confers a performance advantage but introduces a systemic single-point-of-failure risk. The collapse of FTX in November 2022 confirmed this judgment: roughly $8 billion in client funds was misappropriated under an opaque monolithic architecture, and the concentration of custody, matching, and settlement caused internal controls to fail completely. A fully on-chain model achieves maximal trust-minimization but is constrained by the throughput bottleneck of the blockchain. The theory of layered trust offers a solution that transcends this binary choice: at each layer, one independently seeks the optimal trade-off between trust and performance. For example, the asset-custody layer adopts a maximally trust-minimized on-chain mechanism to safeguard security, while the matching-and-execution layer retains an off-chain engine in pursuit of speed. This approach turns trust from an abstract institutional problem into a configurable

engineering parameter (Chapters 28–29).

AI's role in financial markets is evolving from participant to infrastructure. Early AI applications concentrated on algorithmic trading and high-frequency market making and were, in essence, sophisticated participants extracting value from the market under a given set of rules. As model capabilities have advanced, AI has begun to take on system-level functions such as dynamically adjusting margin parameters and parsing unstructured data, marking a paradigm shift from "code is law" to "model is law." The contribution of algorithmic trading to market price discovery already rests on a solid empirical footing [16]; building on this, AI further gives the system an adaptive, dynamic responsiveness that lets it optimize operating parameters in real time across a multidimensional state space. However, homogeneous models may, under certain conditions, trigger decision resonance and create a novel type of liquidity crisis. How to guard against AI-driven endogenous fragility while improving efficiency is a central question for the next generation of microstructure design (Chapter 31).

33.2 Theoretical threads running through the book

The seven core propositions are the book's explicit conclusions. Beneath them, five methodological threads run through the analysis across the book's eleven parts. These threads are not answers to any particular phenomenon but recurring analytical perspectives and epistemological stances that constitute the book's underlying logical framework.

Traditional microstructure theory treats market institutions as exogenous variables and analyzes price discovery and market-making behavior under a given set of rules. The book's first methodological thread treats the institution itself as an endogenous variable: the core microstructure features of perpetual futures markets can all be traced back to the underlying institutional design. The high-leverage margin regime endogenously creates mechanistic volatility (Proposition 33.1.4); the periodic settlement of the funding rate produces price pulses independent of information flow (Proposition 33.1.2); the absence of an expiration date removes the friction of contract rollovers but also removes the deadline for natural convergence; and around-the-clock trading eliminates overnight risk accumulation but also strips away the shock-absorber effect of traditional markets. Institutions are not the backdrop of the analysis but the first cause that shapes the form of the market.

This perspective offers direct guidance for policy and design. The effective way to change a microstructure phenomenon is not investor education or another layer of external regulation but a change in the very institutional design that produces it. The institutional-shaping effect remains strong across all parts of the book and peaks in Part 3 (the funding-rate mechanism). In evaluating any financial innovation, the first question is not its functional features but how its institutional framework alters the incentive structure of participants.

There is no cost-free efficiency improvement in financial markets. The book's second thread reveals the universality of trade-offs: high leverage lowers the cost of capital while amplifying systemic fragility (Propositions 33.1.3 and 33.1.4); market-maker priv-

ileges enhance liquidity but sacrifice fairness (Proposition 33.1.5); the permissionless principle improves accessibility but increases the risk of abuse; physical and logical boundaries exist between trust-minimization and performance-maximization (Proposition 33.1.6); and the rigid enforcement of compliance stands in tension with the flexibility needed for crisis response. These trade-offs are not design flaws but intrinsic features of a multi-objective optimization problem.

Acknowledging the universality of trade-offs is a design philosophy, not a pessimistic judgment about the evolution of markets. Effective system design does not seek to eliminate trade-offs but to find, under given constraints, the current optimal position on the Pareto frontier, while recognizing that this position moves dynamically as technological conditions and the regulatory environment change. The strength of this thread rises in the later parts and peaks in Part 8 (market quality): as the analysis shifts from single-dimension optimization to holistic assessment, trade-offs become the central issue.

The normal market and the crisis market operate on different logics and cannot be explained within a single unified theoretical framework. This thread ties together several of the book's propositions: liquidity is abundant under normal conditions but evaporates in a crisis (Proposition 33.1.3); volatility is predominantly informational under normal conditions but is taken over by mechanistic factors in a crisis (Proposition 33.1.4); and the priorities of market quality flip fundamentally across states (Proposition 33.1.5). Analysis that does not distinguish between states is not only incomplete but can lead to dangerous design decisions, because the optimal solution under normal conditions may be precisely the worst choice in a crisis.

State dependence requires market designers to build systems for two utterly different worlds at once. AI market makers can effectively enhance liquidity under normal conditions, but their homogeneous strategies may cause their collective behavior to converge during extreme events (Proposition 33.1.7). The strength of this thread rises gradually from Part 1 and stays high beyond Part 7, reflecting that the deeper one's understanding of the market becomes, the more state dependence emerges as an unavoidable core methodology.

Among all governance tools, transparency has the lowest cost and the highest net benefit. This thread connects propositions at both the micro and macro levels: at the micro level, price transparency directly affects arbitrage efficiency and the anchoring effectiveness of the funding rate (Proposition 33.1.2); at the macro level, the architectural transparency provided by on-chain infrastructure—guaranteed by cryptography and distributed consensus rather than by manual audit—turns trust from a dependence on institutions into a technological guarantee (Proposition 33.1.6). The systemic harm done by the opaque architecture of centralized exchanges has been confirmed repeatedly, and the core logic of embedded compliance is precisely to use architectural transparency to reshape the governance paradigm.

Transparency is not without limits. In certain settings, full transparency may conflict with privacy protection or, in AI-driven games, expose strategies. But as a foundational governance principle, increasing transparency is always an effective means of reducing information asymmetry and guarding against systemic risk. This thread peaks in Part 4 (price discovery) and again in Parts 9–10 (governance and engineering), reflecting the

dual role of transparency at both the micro-efficiency and macro-governance levels.

Technology changes the rules of financial markets but not the human nature of participants. On-chain transparency eliminates information asymmetry, smart contracts sharply reduce the possibility of asset misappropriation, and AI may remove humans' emotional reactions in extreme conditions. Yet the pursuit of high leverage, the tendency toward power concentration in decentralized governance, and the novel resonance risk created by the homogenization of algorithmic strategies all show that the basic features of human nature have merely taken on new forms as technology evolves (Propositions 33.1.3, 33.1.6, and 33.1.7). The need for trust is displaced by technology across different layers but never disappears.

This recognition constitutes the book's core design philosophy: robust system design should assume that participants will do the worst out of self-interest, and then use architecture and incentive mechanisms to make such behavior physically or logically impossible. This thread runs at a gentle baseline strength across all parts of the book and peaks in Part 11 (the future outlook): when looking ahead to an AI-dominated market, a clear-eyed understanding of human nature becomes especially important.

33.3 The confidence spectrum of the conclusions

Academic honesty requires researchers not only to state what they have found but also to mark clearly how confident they are in those findings. The analysis spanning 33 chapters has advanced a great many theoretical frameworks, empirical observations, and forward-looking inferences, but the epistemological status of these conclusions is not equal: some rest on massive volumes of high-frequency data and rigorous mathematical proof, some depend on preliminary observations of early protocols, and still others are logical extrapolations about future technological evolution. This section divides the book's core findings into four epistemological tiers, from highest to lowest confidence (high confidence, medium-high confidence, medium confidence, and exploratory conjecture), and, at the far end of the spectrum, sets out the current knowledge gaps. The confidence grading rests on three dimensions: the rigor of the theoretical foundation (whether there is a formalized causal logic), the breadth of empirical support (whether it has been cross-validated across independent datasets), and the adequacy of stress testing (whether it has been tested under extreme market conditions). Overall, the higher-confidence findings cluster at the micro-mechanism level, while the lower-confidence inferences mostly concern the direction of macro-level evolution—a distribution that reflects the difference between the rigor of microstructure analysis of local mechanisms and the inherent uncertainty in predicting system-level evolution.

High-confidence findings are conclusions that possess a solid theoretical foundation, have been cross-validated by multiple independent empirical studies, and have been tested under repeated episodes of extreme market conditions. These findings constitute the core premises for understanding perpetual futures markets.

The effectiveness of the funding-rate mechanism is among the most firmly established of these findings. Since BitMEX introduced perpetual futures in 2016, nine years of market data show that the funding rate can effectively maintain convergence between

the contract price and the spot price without intervention by any central authority [7]. Its core logic is pure economic reasoning: when the contract trades at a premium, longs pay shorts, directly eroding their return and incentivizing rational traders to close positions or open arbitrage positions. This self-regulating feedback mechanism has remained effective across bull, bear, and range-bound markets, and rate anomalies during extreme events have returned to normal within short cycles. He et al. (2024) [7] find that, over their sample period, the magnitude of the deviation between perpetual and spot prices declined by an average of about 11% per year, which the authors attribute to increasing arbitrage capital and intensifying competition (that is, the market gradually becoming more efficient) rather than to the speed at which a deviation reverts to the no-arbitrage price at any given point in time.

The leading role of perpetual futures in price discovery for major crypto assets is likewise a high-confidence finding. Applying the information-share methodology established by Hasbrouck (1995) [17], Alexander et al. (2020) [3] show empirically that the crypto derivatives market contributes more than 60% of the price-discovery information share; Makarov and Schoar (2020) [4] further confirm the key role of cross-market arbitrage in narrowing spreads; and Cong et al. (2023) [18] caution that wash trading must be filtered out when evaluating trading-volume data. High leverage and low friction give informed traders higher returns in perpetual futures markets, thereby concentrating the price-discovery function on the derivatives side.

On market-maker behavior, adverse-selection costs and inventory risk remain the core variables determining the supply of liquidity [5] [19]. Although the microstructure of crypto markets differs markedly on the surface from that of traditional equity markets, the market maker's profit equation—in which the sum of adverse-selection costs, inventory risk, and institutional costs determines quoted depth and spread width—is clearly reproduced in the high-frequency order-book data of perpetual futures, confirming the cross-market applicability of classical microstructure theory.

Volatility clustering in perpetual futures markets, and its anomalous intensity, have also been amply verified [14]. GARCH-model estimates [20] and high-frequency realized-volatility analysis consistently show that volatility shocks in this market are extremely persistent. The clustered arrival of information, the amplifying effect of leverage on price fluctuations, and the removal of the traditional overnight buffer by around-the-clock trading together produce volatility clustering more pronounced than in traditional markets. The roughly 50% cumulative decline in Bitcoin over March 12–13, 2020, and the multi-day amplification of volatility it triggered, are a textbook case of this feature.

The monolithic architecture of centralized exchanges carries systemic trust risk. The collapse of FTX in November 2022—roughly $8 billion in client funds misappropriated under an opaque architecture, with an internal backdoor program allowing affiliated parties to bypass automatic liquidation—confirmed that when matching, clearing, and custody are concentrated in a single opaque entity, the principal-agent problem is amplified to a systemic scale. This lesson became a core force driving on-chain financial infrastructure toward a layered trust architecture.

Medium-high-confidence judgments are conclusions that are logically well argued and supported by early empirical data but have not yet weathered a complete economic

cycle or extreme stress testing. At the current level of understanding, these judgments are highly reliable, yet room must be left for future data to revise them. The theoretical framework that decomposes the volatility of perpetual futures into four sources—informational, liquidity, mechanistic, and transmission (see Proposition 33.1.4)—belongs to this category. This framework has considerable explanatory power at the conceptual level, clearly distinguishing healthy price discovery from the noise introduced by institutional design. Precisely quantifying and separating these four volatility sources, however, still poses methodological challenges at the empirical level: existing econometric tools often require fairly strong identifying assumptions when handling highly nonlinear and endogenous high-frequency data, so the specific decomposition proportions still carry a nontrivial confidence interval (this identification difficulty is developed in detail in Section 33.4.1).

With regard to infrastructure architecture, we have ample reason to believe that, over the long run, a layered architecture will outperform both extremes, whether fully on-chain or fully off-chain. Decomposing trust into distinct layers—asset custody, matching and execution, clearing and settlement, and so on—and independently seeking the optimal trust-performance trade-off at each layer has a clear advantage in engineering terms. The asset-custody layer requires the highest security and should therefore adopt architectural trust solutions such as multi-signature wallets or hardware wallets; the matching layer has the highest performance requirements and can adopt a centralized or semi-centralized design in exchange for speed; and the clearing layer must balance determinism against flexibility and can adopt a hybrid approach. This design preserves the asset security of the blockchain while breaking through the limits that the consensus mechanism imposes on matching speed. However, the stability of such a complex architecture when processing tens of thousands of orders per second, and the reliability of cross-layer state synchronization, still require confirmation from larger-scale real-world deployment. The problem of cross-layer state synchronization deserves particular attention: when orders generated at the matching layer must be finally settled at the clearing layer, how can the state of the two layers be kept consistent at all times? How can state splitting be prevented under extreme network latency or system failure? Solutions to these engineering challenges are still in the exploratory stage.

On liquidity governance, the protocol-led market-making vault model is regarded as a highly promising mechanism innovation. By pooling retail capital and adopting algorithmic strategies, it partly mitigates the problem of market-maker capital monopoly and provides the protocol with an endogenous liquidity buffer. Data from early decentralized exchanges show that this model can effectively compress bid-ask spreads and improve market depth under normal conditions. Whether, under a one-sided crash or extreme volatility, it will trigger a liquidity-spiral drain through strategy homogenization or excessive risk exposure, however, still awaits empirical testing (see Section 33.4.2). Likewise, gradual liquidation mechanisms can, in theory, substantially reduce the mechanistic volatility triggered by liquidation cascades relative to traditional cliff-edge liquidation. Early results from some exchanges bear this out. Yet how to dynamically calibrate the optimal liquidation-penalty parameters and step sizes for the liquidity characteristics of different assets remains a complex optimization problem.

Medium-confidence inferences are reasonable extrapolations based on current trends

and economic logic; their path to realization is constrained by multiple uncertain factors such as technological bottlenecks, regulatory policy, and strategic interaction among market participants, and they indicate possible directions of development. The foremost conjecture is that the perpetual futures mechanism will move beyond crypto assets into traditional financial domains such as foreign exchange, commodities, and macro equity indices. The generality of funding-rate anchoring means there is no fundamental obstacle to this cross-domain expansion at the mechanism level. The low-friction character of synthetic exposure holds enormous appeal for traders seeking pure price speculation. However, the deeply entrenched regulatory barriers of traditional financial markets, the inertia of existing clearing infrastructure, and the reliance of traditional assets on physical delivery constitute enormous institutional resistance. Just how broad and deep this expansion can ultimately become remains an open question.

Another important conjecture concerns the transformation of AI's role in financial markets. We expect AI to evolve gradually from a mere trading participant into market infrastructure itself. From high-frequency trading algorithms to dynamic risk engines and on to predictive liquidation systems, AI is taking over ever more foundational market functions. This evolution from value extraction toward value creation is, as a matter of economic logic, an inevitable result of the pursuit of efficiency. But the timetable for this process is hard to predict; it depends not only on breakthroughs in large language models and reinforcement learning but, more importantly, on the financial system's tolerance for uninterpretable algorithms.

As AI penetrates more deeply, "model is law" will become an important supplement to, and even a partial replacement for, the "code is law" paradigm. The rigid rules of smart contracts guarantee deterministic execution but also make them slow to respond to complex, unexpected events. Introducing well-trained and well-validated AI models to handle governance decisions and risk-parameter adjustments under ambiguous conditions is a reasonable direction for improving system resilience. However, this conjecture faces enormous governance challenges: How can the model's training data be kept from contamination? How can the model be prevented from hallucinating during extreme, long-tail events? How can probabilistic model outputs be incorporated into deterministic financial contracts? There are as yet no mature solutions to these questions, which makes this inference relatively uncertain.

Exploratory conjectures are long-range extrapolations that lack an empirical foundation; they serve more to expand the boundaries of understanding and to point out strategic directions worth long-term attention than to offer testable or immediately tradable short-term predictions. One bold conjecture is that decentralized perpetual futures markets will ultimately become the reference-pricing venue for the world's major assets. This conjecture rests on a premise: as on-chain liquidity continues to aggregate and transaction costs keep approaching their physical limits, the most informationally efficient market will naturally acquire pricing power. Should this vision come true, traditional exchanges would degenerate into followers of liquidity. But this process must overcome extremely high network-effect barriers, and it faces strong resistance from sovereign states unwilling to cede pricing power over core financial assets.

A further conjecture is the eventual formation of the "autonomous financial network"—

the three-layer stack, defined in Chapter 32, composed of an asset layer, an infrastructure layer, and an agent layer. In this conception, a financial stack made up of smart contracts, decentralized oracles, and AI agents would be able to issue assets, match trades, manage risk, and perform clearing entirely autonomously, without the intervention of any human intermediary. This is an extreme technological utopia; it requires the throughput of the underlying public chain to leap by orders of magnitude, cross-chain interoperability to achieve seamless connection, and the entire system to withstand nation-state-level computational attacks. For the foreseeable future, this conjecture will remain at the proof-of-concept stage.

A key technological conjecture underpinning the vision above is a breakthrough in semantic oracles. Traditional numerical oracles can only transmit discrete price data, whereas a semantic oracle parses natural language, legal text, or real-world events into verifiable on-chain state; its AI-infrastructure prerequisites are set out in Chapter 31. It is expected to understand and verify complex natural-language texts, legal contracts, and even real-world news events and to translate them into on-chain executable logic. If this technology can be realized, it will fully unlock the potential of "perpetual futures on everything," allowing any social event or economic indicator with consensus value to be financialized. However, the ambiguity of natural language and the limitations of AI comprehension mean that the reliability of semantic oracles will, for a long time to come, struggle to meet the exacting requirements of high-value financial transactions.

At the end of the confidence spectrum, we must honestly confront the questions that currently lie entirely beyond our capacity to know. Admitting ignorance is among the most valuable qualities in scientific research: it guards against theoretical hubris and points the way for future exploration. First, we do not know how large a liquidity shock the deep homogenization of AI strategies would trigger in an extreme black-swan event (the definition of its critical threshold is discussed in Section 33.4.1). The market has not yet experienced an "algorithmic flash crash" set off jointly by thousands of agents sharing similar underlying models and training data, and, for lack of a historical sample, this novel form of systemic risk cannot be accurately assessed by existing stress-testing models.

Second, we do not know where the optimal system-level leverage limit should be set in a decentralized perpetual futures market. The leverage ratio bears not only on capital efficiency but, more directly, on the destructive power of a liquidation cascade. It is a dynamically changing endogenous variable that depends on market depth, the volatility regime, and participants' risk preferences. The very search for a static, universal optimal leverage ratio may, in theory, be misconceived.

We also do not know whether fully decentralized governance mechanisms can remain effective in managing ultra-large-scale, extremely complex financial infrastructure. Historical experience shows that, as an organization grows in scale, decentralized governance often falls into inefficiency, capture by interest groups, or de facto oligarchic control. In an on-chain world unconstrained by a mandatory legal framework, how to design a governance mechanism that both prevents power concentration and preserves agile decision-making remains an unsolved sociological problem.

Moreover, we cannot foresee how sovereign states will ultimately mount a systematic re-

sponse to this permissionless, cross-border autonomous financial network. This already exceeds the scope of economics and microstructure theory and enters the domain of international political economy and national security strategy. In the contest between the coercive force of regulation and the censorship resistance of technology, whether the final equilibrium is outright suppression, conditional co-optation, or parallel coexistence, no one can give a definite answer at present.

Finally, we do not know how much of the theoretical framework built in this book will retain its explanatory power after a decade of technological iteration and market evolution. Market microstructure is a complex adaptive system in perpetual motion, especially in crypto finance, a field whose pace of evolution is measured in months. Perhaps in the near future, new institutional designs and technological breakthroughs will completely overturn the "stylized facts" we hold to be certain today. Yet it is precisely this endless evolution that gives the theory of financial microstructure its enduring research value. The current knowledge gaps are exactly the starting point for the next generation of researchers.

33.4 Unresolved problems and the research agenda

A good book not only answers questions but also poses new ones. In this systematic dissection of the microstructure of perpetual futures across 33 chapters, we have attempted to build a theoretical framework that explains the operating mechanisms of this emerging financial instrument. Yet as the analysis deepened, we became ever more aware of the limits of the current theory. Scientific progress often arises not from repeatedly confirming the known but from honestly confronting the unknown. What follows are problems that surfaced repeatedly over the course of the book's analysis but that, constrained by current data conditions, theoretical tools, or the stage of historical development, could not—and should not—be forced to a resolution within this book. These unresolved problems constitute the research agenda for the next stage of financial microstructure theory.

Figure 33-1 organizes the unresolved problems discussed in this section into a structured roadmap by research category (theoretical, empirical, and design) and time horizon (short-term, 1–3 years; medium-term, 3–7 years; and long-term, more than 7 years), offering researchers at different stages a set of prioritized entry points.

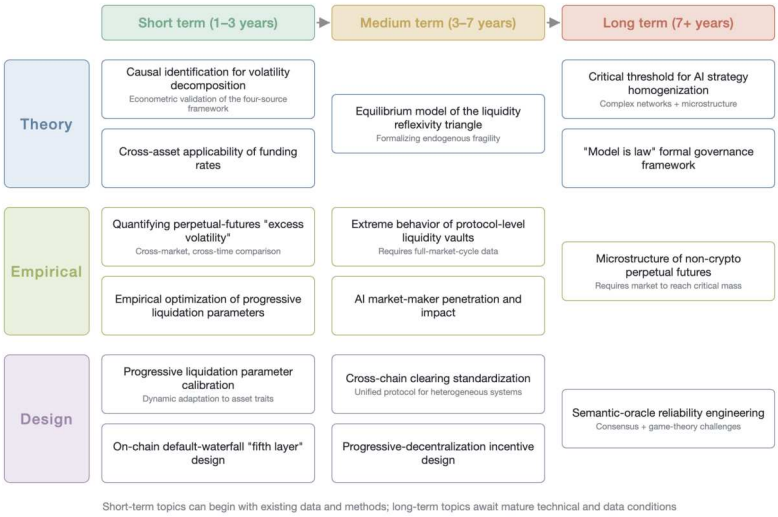


Figure 33-1. Research-agenda roadmap for unresolved problems: a research-priority matrix organized by category (theoretical/empirical/design) and time horizon (short-/medium-/long-term), where the entries are the unresolved problems raised in this book, compiled by the author and not empirical data

On the theoretical side, although the book has proposed a series of conceptual frameworks, translating them into rigorous formal models still faces major challenges. The first is the problem of rigorous causal identification for the four-source volatility decomposition. We deconstruct the volatility of perpetual futures into four sources—informational, liquidity, mechanistic, and transmission—and this conceptual distinction offers an important perspective for understanding market states. With the current econometric toolbox, however, separating these four components within high-frequency price sequences remains an unsolved problem. Existing decomposition methods often rely on strong assumptions—for example, that information shocks and liquidity frictions are mutually independent—but under extreme market conditions this independence assumption tends to fail. Future research will need to develop more refined identification strategies, possibly by introducing nonlinear dynamical models or drawing on higher-dimensional datasets, in order to quantify the different volatility sources precisely.

The second is the formal modeling of the liquidity reflexivity triangle. Through qualitative analysis, the book has revealed the reflexive loop among market-maker inventory, price volatility, and liquidity supply, pointing to the endogenous fragility whereby liquidity is scarcest when it is most needed. But turning this intuitive insight into a closed-form dynamic equilibrium model remains a formidable task. Traditional market-maker models typically assume an exogenous price process, whereas here the price process is itself endogenously determined by the collective behavior of market makers. Building such a macro-microstructure model that encompasses multiple equilibria and nonlinear feedback requires not only overcoming extremely high mathematical complexity but also striking a delicate balance between theoretical elegance and real-world explanatory power.

The third theoretical difficulty is defining the critical threshold for AI-strategy homogenization. As the penetration of algorithmic trading and AI market makers rises, strategy convergence becomes a systemic risk that cannot be ignored. When enough market participants use similar models and data sources, algorithms originally intended to dampen volatility may resonate under particular triggering conditions and set off an instantaneous drain of liquidity. But how much homogenization counts as "too homogeneous"? Is there a critical threshold that can be rigorously defined, beyond which market resilience undergoes a phase-transition-like collapse? Answering this question requires a deep fusion of complex-network theory with microstructure models—undoubtedly a major frontier for future theoretical research.

Finally, a formalized governance framework for "model is law" is likewise urgently needed. In exploring the evolution of AI from market participant to market infrastructure, we have foreseen a future in which algorithmic models may replace traditional code as the core of rule enforcement. Unlike deterministic smart-contract code, however, an AI model is inherently probabilistic, and its decision process often has the uninterpretable character of a black box. How can such a probabilistic, non-deterministic decision mechanism be incorporated into existing legal and economic frameworks of liability? When an autonomously operating risk-control model makes a misjudgment that causes enormous losses, how should responsibility be assigned? Building a formalized governance theory compatible with probabilistic logic is a prerequisite for realizing a truly autonomous financial network.

The validity of theory lies in empirical testing. On the empirical side, because the crypto-asset market has a short history and a complex data structure, many key hypotheses still await further verification. First and foremost is the precise quantification of the "excess volatility" of perpetual futures. Although in theory high leverage and the funding-rate mechanism introduce additional volatility, accurately measuring the specific share of this mechanistic volatility in total volatility requires systematic comparison across markets, across time, and across asset classes. Current empirical research is often limited by a lack of out-of-sample data and struggles to control for confounding variables such as the macroeconomic environment and changes in regulatory policy. Future work will need to build a larger and more standardized microstructure database and apply advanced panel-data techniques before it can deliver statistically significant, definitive conclusions.

In addition, the real-world performance of protocol-based liquidity vaults under extreme market conditions remains an open empirical question. As an institutional innovation for addressing the endogenous fragility of liquidity, such vaults have shown potential to improve liquidity under normal conditions. Yet the true test of any financial institution always comes in a crisis. Because such mechanisms are still young, they have not yet weathered a systemic shock on the scale of the global financial crisis. When faced with the concurrent pressures of extreme selling, oracle failure, and network congestion, will these vaults serve as an effective buffer that stabilizes the market, or will design flaws accelerate their collapse? Answering this empirically will require more time and more "battle testing."

The true penetration and impact of AI market makers in crypto markets is likewise

hard to quantify empirically. Although algorithmic trading is widely seen as increasingly prevalent, the anonymity of on-chain identities and the concealment of trading strategies make it hard for researchers to distinguish accurately between the order flow of human traders and that of AI algorithms. The lack of reliable classification data makes it extremely difficult to assess AI's specific impact on market liquidity, price-discovery efficiency, and volatility characteristics. Future empirical research may need to draw on more advanced machine-learning classification techniques, or rely on high-granularity anonymized data provided at the exchange level, before it can gradually shed light on this question.

Finally, the market-microstructure characteristics of non-crypto-asset perpetual futures are an important empirical blind spot. As a thought experiment, the book has explored the possibility of extending the perpetual futures mechanism to traditional assets such as foreign exchange, equities, and commodities. At present, however, the trading volume on platforms attempting to launch such products remains too small to provide enough observations for reliable empirical analysis. Traditional assets and crypto assets differ fundamentally in their underlying value anchor, the structure of market participants, and the regulatory environment; how will these differences reshape the microstructure behavior of perpetual futures? Answering this question must wait until the relevant markets reach a certain degree of maturity and scale of liquidity before substantive empirical exploration can begin.

Beyond the theoretical and empirical challenges, in financial engineering and system design the book also leaves a series of problems unsolved. One of the most pressing is the design of a final recovery and resolution mechanism beyond the on-chain default waterfall. In the traditional central-counterparty clearing system, once all the front-line risk buffers have been breached, the ultimate backstop responsibility is typically borne by the central bank or the sovereign state; this is an external public-credit backstop, not an ordinary tier inside the default waterfall. In an on-chain financial network that pursues decentralization and permissionlessness, however, such an external government backstop does not exist. When an extreme black-swan event causes all clearing mechanisms within the system to fail and the bankruptcy shortfall cannot be covered, how should the system design its last line of defense? Is it an optimal auto-deleveraging mechanism, the issuance of a system debt token, or some socialized loss-mutualization protocol not yet invented? Finding an ultimate risk-buffer solution in the absence of sovereign-credit backing is a necessary path for on-chain finance to reach maturity.

The reliability engineering of semantic oracles (see Section 33.3.4) is another highly challenging design problem. Introducing large-language-model-based AI into an oracle network inevitably brings the risks of "hallucination" and adversarial attacks. How can a robust consensus mechanism be designed to ensure that multiple independent AI nodes, when confronting ambiguous or even maliciously constructed real-world information, output consistent and accurate semantic-parsing results? This is not only a technical problem in natural language processing but also a complex engineering challenge involving game-theoretic incentive design.

On the evolution of organizational architecture, the incentive design for "progressive decentralization" remains a problem not yet fully solved. In their initial phase, many

on-chain protocols depend on a core team for centralized development and operation, and promise to hand control gradually to the community. In reality, however, the team often lingers in the centralized phase for a long time—whether out of vested interests or distrust of the community's governance capacity—so that "progressive" becomes "never." How can sophisticated token-economics design and hard-coded smart-contract time-locks create an irreversible incentive mechanism under which the core team is not only willing but mathematically compelled to relinquish power step by step along the roadmap? This requires us to seek more powerful constraint tools within mechanism-design theory.

Finally, the standardization of cross-chain clearing constitutes a major design bottleneck for the future of multi-chain interoperability. As liquidity grows increasingly fragmented across different blockchain networks, it is becoming the norm for users to hold related assets and positions on multiple chains at once. To improve capital efficiency and reduce systemic risk, cross-chain margin sharing and net settlement become especially necessary. Yet blockchains differ enormously in consensus mechanism, finality-confirmation time, and smart-contract execution environment; how can a unified, secure, and efficient cross-chain clearing standard be established among these heterogeneous systems? This requires not only securing cross-chain message passing but also designing a global clearing algorithm that can tolerate asynchronous execution and partial network partitions. The absence of such a standard remains the main obstacle preventing on-chain financial networks from optimizing capital efficiency globally.

33.5 Disciplinary contributions and epistemological legacy

The book's disciplinary contribution lies not in analyzing an emerging market as such but in the way the extreme institutional design of perpetual futures exposes blind spots long obscured within traditional microstructure theory. The institutional infrastructure of traditional markets—designated market makers, central-counterparty clearing, and market-closing conventions—masks several deep contradictions. Once perpetual futures strip away these protective institutional layers, the essential features of microstructure become visible under extreme conditions, giving rise to five epistemological legacies.

The classic framework established by O'Hara (1995) [6] divides volatility into two categories: informational volatility (from fundamental updates) and noise volatility (from liquidity imbalances or irrational behavior). This dichotomy has strong explanatory power in traditional equity and futures markets, where liquidation rules and trading institutions are relatively mild and the frictions they generate are usually subsumed under noise in the broad sense. Perpetual futures, however, break this pattern: high-leverage liquidation engines force selling when a price crosses a trigger threshold, disregarding fundamental information, and periodic funding-rate settlement induces the mechanical repositioning of arbitrage capital at fixed points in time. The mechanical buying and selling pressure this produces [7] neither reflects new information nor stems from trader irrationality but is a direct product of the logic of system design. On this basis, the book expands the information/noise dichotomy into a four-source framework—informational, liquidity, mechanistic, and transmission—making institution-driven volatility explicit as an independent analytical category, and it poses a new question for future research:

to what extent does the institutional design of any emerging market endogenously create additional volatility? Only by separating out institution-driven volatility, as far as it can be identified, can regulators distinguish the healthy cost of price discovery from pure friction that architectural optimization can mitigate.

The classic models of Kyle (1985) [21] and Glosten and Milgrom (1985) [5] tend to treat liquidity as an exogenous background condition, reducing the market maker to a passive provider seeking a static equilibrium between adverse-selection costs and inventory costs; and the designated-market-maker system and stable credit support of traditional markets further obscure the intrinsic fragility of liquidity supply. Analysis of perpetual futures markets, by contrast, shows that liquidity is essentially a commodity produced by market makers, and that its production equation contains a structural flaw: production cost is positively correlated with demand. Under normal conditions, adverse selection and inventory risk are low, and market makers tend to supply liquidity far in excess of demand; once the market shifts into a high-volatility state, risk costs balloon exponentially, and the production logic of liquidity collapses with them. This countercyclical fragility is not a matter of any moral failing or irrational panic on the part of market makers but a structural result of profit maximization under the adverse-selection and inventory-cost structure [5]. It rewrites our understanding of market resilience: the effective way to sustain liquidity during a crisis is not to make moral appeals or issue administrative decrees but to change the underlying risk-return structure of market makers, or to reshape the production logic of liquidity through institutional innovations such as protocol-level liquidity vaults.

The mainstream microstructure literature assumes by default that trading takes place within a complete institutional infrastructure, in which central counterparties, securities depositories, and state regulation form a network of trust [6]; asset security, fairness of matching, and settlement finality are treated as given premises, and the question of trust is left to law and institutional economics. The crises in crypto markets [22] [23]—and especially the systemic failures of centralized exchanges—demonstrate, however, that in an environment lacking state backing, trust is a scarce, central variable that can break down at any moment. The book therefore turns trust from an abstract institutional problem into a concrete engineering problem and proposes the concept of the trust stack: it decomposes a financial transaction into five layers—asset custody, matching and execution, clearing and settlement, data feeds, and governance decisions—and explores at each layer the trade-off frontier between trust-minimization and performance-maximization (Chapters 28–29). This provides a new first principle for financial-system design: the designer no longer faces a binary choice between wholesale centralization and wholesale decentralization but independently seeks the optimal architectural solution at each layer; and future microstructure theory must place trust architecture alongside liquidity and volatility as a foundational dimension for assessing market quality.

Most existing microstructure theory is built on the assumption that the market has opening, closing, and non-trading cycles: the closed period is regarded as a buffer for digesting information; the open and the close carry special price-discovery functions; and concepts such as overnight risk, opening gaps, and closing-price manipulation form the core vocabulary of traditional analysis. Even though the foreign-exchange market

achieved near-around-the-clock trading well before crypto markets did [24], its weekend breaks and the relatively limited microstructure research devoted to it left this assumption largely unshaken for a long time. Perpetual futures markets combine around-the-clock trading with high leverage and the absence of an expiration date, producing microstructure features different from those of the foreign-exchange market: the lack of a closed-period buffer gives price shocks a stronger long-memory effect over time, and volatility clustering is more pronounced than in traditional markets; around-the-clock pricing, meanwhile, fills the pricing vacuum that traditional assets face during the Asian or European sessions, making the global price-discovery process more continuous. This finding calls for redefining the dynamic relationship between information arrival and price adjustment, for exploring market makers' cross-time-zone inventory-management strategies in the absence of time boundaries, and for designing new dynamic risk-control systems for continuous markets—the traditional time-slice-based circuit-breaker mechanism no longer applies here.

The objects of analysis in traditional financial theory are usually anchored to real assets or their direct derivatives: a stock represents ownership of a company; a commodity future carries an obligation of physical delivery; and even complex derivatives have definite expiration dates and exercise conditions. These rigid constraints force prices to converge over time, and traditional arbitrage models and pricing formulas are essentially organized around the expiry constraint; CFDs and total return swaps, although they likewise provide synthetic exposure, are limited by bilateral counterparty risk and finite market depth (see Section 33.1.1). Perpetual futures, by contrast, create a new class of financial instrument—one with no expiration date, no physical delivery, and price anchoring achieved purely through the funding rate's endogenous game-theoretic mechanism [7]. What the trader trades is nothing but pure exposure to price movements, and its economic nature and risk characteristics pose new challenges to traditional derivatives-pricing theory. The book thus establishes the economics of synthetic exposure as a new analytical category, one that calls for building an entirely new theoretical framework: the anchoring efficiency and failure conditions of the funding rate across different states await systematic study; and when this mechanism extends to domains such as foreign exchange and commodities, its reshaping effect on the global risk-transmission network and the distribution of liquidity must also be assessed. The economics of synthetic exposure thus provides the core theoretical framework for understanding the digitalized financial networks of the future.

33.6 Recommendations for different readers

The book's argument now nears its close. This section offers readers of different backgrounds concrete paths of action that build on this book.

What the book offers academic researchers is not a closed set of conclusions but an open research agenda. The unresolved problems listed in Section 33.4—such as the causal identification of four-source volatility, the formal modeling of the liquidity reflexivity triangle, and the critical threshold of AI-strategy homogenization—can all serve as starting points for independent research projects. Hypotheses that merit priority testing include whether the share of mechanistic volatility in total volatility grows nonlinearly

as the leverage ratio rises; whether protocol-level liquidity vaults accelerate rather than relieve liquidity drains under extreme conditions; and whether the anchoring efficiency of the funding rate still holds once it is extended across asset classes. Recommended data sources include the high-frequency order-book data of Kaiko and CoinMetrics, the on-chain metrics of Glassnode, and the public liquidation logs of the various protocols. The book's analytical frameworks—the four-source volatility decomposition, the state-machine model, and trust-stack theory—combined with classical microstructure tools such as market-making quote models [25] and informed-trading models [26], are not confined to crypto markets and can serve as general-purpose theoretical tools for the microstructure analysis of emerging markets.

The book offers practitioners three directly applicable decision frameworks. For market makers, the triple-cost decomposition of the profit equation (adverse selection, inventory, and institutional cost), combined with the four-source identification of volatility, forms the basis for dynamically adjusting quoting strategies under normal and stressed conditions. Specifically, when the share of mechanistic volatility rises, one should give priority to reducing depth rather than widening the spread. For exchange architects and protocol designers, the five-layer decomposition of the trust stack provides an engineering framework for evaluating architectural choices: prioritize trust-minimization at the custody layer, prioritize performance at the matching layer, and adopt a gradual rather than cliff-edge design at the clearing layer. For investors, the five-dimensional market-quality model offers a structured method for evaluating trading venues, while the state-machine perspective is a reminder that, in assessing market resilience, one must distinguish normal-state performance from stressed-state performance.

The book offers regulators tools in three directions. Architectural guarantees of asset segregation and information disclosure are the most direct regulatory instrument: requiring trading platforms to implement verifiable asset segregation at the custody layer is low in cost and measurable in effect. Calibrating leverage limits, by contrast, must rest on microstructure analysis: the book's four-source volatility decomposition shows that the higher the share of mechanistic volatility in a market, the lower the optimal leverage ceiling should be. Embedded compliance—encoding regulatory requirements into the execution logic of smart contracts—offers a technological path for shifting from ex post punishment to real-time monitoring, and its core logic is to use the architectural transparency of on-chain infrastructure to reduce the cost of regulation.

Students are advised to choose one of the following reading paths, according to their interests, as an entry point. Derivatives-pricing track: Chapter 10 (the game theory of funding rates) → Chapter 13 (price discovery) → Chapter 17 (arbitrage risk), with an emphasis on mechanism design and information economics. Market-microstructure track: Chapter 19 (the market maker's profit equation) → Chapter 20 (the endogeneity of liquidity) → Chapters 22–24 (volatility), with an emphasis on the application of classical microstructure theory in a new market. System-design track: Chapter 28 (the trust stack) → Chapter 29 (clearing architecture) → Chapter 31 (AI infrastructure), with an emphasis on engineering and governance design. After exploring a particular topic in depth, one can gradually expand into adjacent areas through the cross-references between chapters.

References

[1] Shiller, R. J. (1993). *Macro markets: Creating institutions for managing society's largest economic risks*. Clarendon Press.

[2] CoinGlass. (2026). *2025 crypto derivatives market annual report*. https://www.coinglass.com/learn/2025-annual-report-en

[3] Alexander, C., Choi, J., Park, H., & Sohn, S. (2020). BitMEX bitcoin derivatives: Price discovery, informational efficiency, and hedging effectiveness. *Journal of Futures Markets*, *40*(1), 23–43.

[4] Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, *135*(2), 293–319.

[5] Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100.

[6] O'Hara, M. (1995). *Market microstructure theory*. Blackwell.

[7] He, S., Manela, A., Ross, O., & von Wachter, V. (2024). Fundamentals of perpetual futures. *arXiv*. https://arxiv.org/abs/2212.06888

[8] Duffie, D., Gârleanu, N., & Pedersen, L. H. (2005). Over-the-counter markets. *Econometrica*, *73*(6), 1815–1847. https://doi.org/10.1111/j.1468-0262.2005.00639.x

[9] Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*, *6*(1), 58–73.

[10] Hurwicz, L. (1972). On informationally decentralized systems. In C. B. McGuire & R. Radner (Eds.), *Decision and organization* (pp. 297–336). North-Holland.

[11] Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, *130*(4), 1547–1621. https://doi.org/10.1093/qje/qjv027

[12] Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098

[13] Adrian, T., & Shin, H. S. (2010). Liquidity and leverage. *Journal of Financial Intermediation*, *19*(3), 418–437. https://doi.org/10.1016/j.jfi.2008.12.002

[14] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*(2), 579–625. https://doi.org/10.1111/1468-0262.00418

[15] Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56. https://doi.org/10.1016/s1386-4181(01)00024-6

[16] Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *Review of Financial Studies*, *27*(8), 2267–2306. https://doi.org/10.1093/rfs/hhu032

[17] Hasbrouck, J. (1995). One security, many markets: Determining the contributions

to price discovery. *The Journal of Finance*, *50*(4), 1175–1199. https://doi.org/10.1111/j.1540-6261.1995.tb04054.x

[18] Cong, L. W., Li, X., Tang, K., & Yang, Y. (2023). Crypto wash trading. *Management Science*, *69*(11), 6427–6454. https://doi.org/10.1287/mnsc.2021.02709

[19] Grossman, S. J., & Miller, M. H. (1988). Liquidity and market structure. *The Journal of Finance*, *43*(3), 617–633.

[20] Bollerslev, T., Litvinova, J., & Tauchen, G. (2006). Leverage and volatility feedback effects in high-frequency data. *Journal of Financial Econometrics*, *4*(3), 353–384.

[21] Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335.

[22] Capponi, A., & Jia, R. (2021). The adoption of blockchain-based decentralized exchanges. *arXiv*. https://arxiv.org/abs/2103.08842

[23] Zetzsche, D. A., Buckley, R. P., Arner, D. W., & Föhr, L. (2019). The ICO gold rush: It's a scam, it's a bubble, it's a super challenge for regulators. *Harvard International Law Journal*, *60*(2), 267–315. https://doi.org/10.2139/ssrn.3072298

[24] Lyons, R. K. (2001). *The microstructure approach to exchange rates*. MIT Press.

[25] Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217–224.

[26] Easley, D., & O'Hara, M. (1987). Price, trade size, and information in securities markets. *Journal of Financial Economics*, *19*(1), 69–90. https://doi.org/10.1016/0304-405x(87)90029-8