

FINANCIAL MARKET INFRASTRUCTURE

EXCHANGE CONCEPTS

PART IV – Technology & Infra

A conceptual guide to financial exchange architecture, order books & matching engines.

46	273	10	NO
CHAPTERS	GLOSSARY TERMS	REAL EXCHANGES	MATHS REQUIRED

J. Persson

v1.10.1 · 2026

Exchange Concepts An Introduction
Book Version: 1.10.1
Copyright © Johan Persson
Generated on October 16, 2026
Prepared with LaTeX from Markdown sources for duplex printing. Sources for documentation available at
https://github.com/johan162/EduMatcher

Exchange Concepts

An Introduction for Beginners

Johan Persson <johan162@gmail.com>

Generated on October 16, 2026

Overview

Preface to the Second Edition ix

Preface to the First Edition xi

I Part IV: Technology, Infrastructure, and Market Ecology 1

1 Speed Bumps, Leveling the Playing Field 5

2 The Technology Architecture 7

3 Conformance Testing and Onboarding 11

4 Primary and Secondary Sites, Resilience Architecture 15

5 Load Balancing, Distributing the Work 19

6 Market Data Architecture, How the Market Sees Itself 21

7 Market Data Economics 23

8 Smart Order Routing and Market Fragmentation 25

9 Fixed Income Market Structure: Where the Order Book Model Breaks Down 31

10 Cryptocurrency and Digital Asset Venues 37

11 Latency and Co-location, The Speed Dimension 41

12 Operational Observability and Incident Response 45

13 Corporate Actions, When the Instrument Changes 49

14 Options Mechanics: Exercise, Assignment, and Expiry 51

15 Determinism, Replay, and Persistence, The Exchange Must Not Forget 55

16 Reference Data, The Exchange's Ground Truth 57

Glossary **59**

References **79**

Exchanges own technical references **83**

Back-Cover Text **87**

About the Author **89**

Contents

Preface to the Second Edition ix

Preface to the First Edition xi

I Part IV: Technology, Infrastructure, and Market Ecology 1

1 Speed Bumps, Leveling the Playing Field 5
1.1 What Is a Speed Bump? . . . 5
1.2 IEX: The Most Famous Speed Bump . . . 5
1.3 Speed Bumps in Broader Design . . . 5

2 The Technology Architecture 7
2.1 The Gateway . . . 7
2.2 Binary Protocols: ITCH and OUCH . . . 8
2.3 TCP vs UDP: Why Market Data and Order Submission Use Different Transports . . . 9
2.4 The Matching Engine . . . 9
2.5 The Message Bus . . . 10
2.6 Subscribers . . . 10
2.7 Book Snapshots . . . 10

3 Conformance Testing and Onboarding 11
3.1 Why This Cannot Be Skipped . . . 11
3.2 The Onboarding Sequence . . . 11
3.3 What a Conformance Test Script Actually Checks . . . 11
3.4 Keeping Certification Meaningful: The Drift Problem . . . 12
3.5 Offboarding . . . 12

4 Primary and Secondary Sites, Resilience Architecture 15
4.1 Primary and Secondary Sites . . . 15
4.2 Synchronisation: Keeping the Secondary Current . . . 15
4.3 Failover . . . 16
4.4 The Geographic Trade-Off . . . 18

5 Load Balancing, Distributing the Work 19
5.1 Horizontal Scaling Across Symbols . . . 19
5.2 Gateway Load Balancing . . . 19
5.3 The Unique Order Number Challenge . . . 19
5.4 Monotonic Timestamps Across Sites . . . 20

6 Market Data Architecture, How the Market Sees Itself 21
6.1 Snapshots vs Incremental Updates . . . 21
6.2 Sequence Numbers . . . 21
6.3 Sequencing Scope and Session Boundaries . . . 21

6.4 Gap Recovery and Replay . . . 22
6.5 Conflation . . . 22
6.6 Tick-to-Trade Latency . . . 22
6.7 Top-of-Book vs Depth-of-Book . . . 22

7 Market Data Economics 23
7.1 Two Tiers of the Same Information: SIP vs. Proprietary Feeds . . . 23
7.2 Market Data as Exchange Revenue . . . 23
7.3 The Regulatory Response . . . 24

8 Smart Order Routing and Market Fragmentation 25
8.1 Why Markets Are Fragmented . . . 25
8.2 The National Best Bid and Offer (NBBO) . . . 25
8.3 Locked and Crossed Markets . . . 25
8.4 Smart Order Routing (SOR) . . . 26
8.5 Dark Pools and Hidden Liquidity . . . 26
8.6 Payment for Order Flow (PFOF) . . . 26
8.7 Dark Pool Regulatory Scrutiny . . . 27
8.8 Exchange Fee Models: Maker-Taker and Taker-Maker . . . 27
8.9 Trade Reporting Obligations . . . 28
8.10 Execution Algorithms, Slicing Large Orders . . . 29

9 Fixed Income Market Structure: Where the Order Book Model Breaks Down 31
9.1 Why Bonds Break the Central-Limit-Order-Book Model . . . 31
9.2 The Dealer-Intermediated, Quote-Driven Market . . . 32
9.3 The Two Very Different Halves: Governments vs Corporates . . . 32
9.4 Post-Trade Transparency: TRACE and the End of the Opaque Quote . . . 33
9.5 Pricing Conventions: Yield, Price, and the 32nds . . . 33
9.6 Settlement and the Move to Central Clearing . . . 34
9.7 The Primary Market: Auctions and Syndication . . . 34
9.8 What the Developer Should Carry Away . . . 35

10 Cryptocurrency and Digital Asset Venues 37
10.1 What Carries Over Unchanged . . . 37
10.2 What Does Not: Continuous, Global, 24/7 Trading . . . 37
10.3 What Does Not: Settlement . . . 37
10.4 What Does Not: A Single Regulator . . . 38
10.5 Precision: Why Tick Size Gets Harder, Not Easier . . . 38

11 Latency and Co-location, The Speed Dimension 41
11.1 Why Latency Matters . . . 41
11.2 Co-location . . . 41
11.3 Microwave and Laser Links . . . 42
11.4 Hardware Acceleration . . . 42
11.5 PTP Clock Synchronisation . . . 43

12 Operational Observability and Incident Response 45
12.1 The Four Golden Signals . . . 45
12.2 Core Production Signals . . . 45
12.3 Clock and Time-Synchronisation Health . . . 46
12.4 The Three Pillars: Metrics, Logs, Traces . . . 46
12.5 Service-Level Objectives and Error Budgets . . . 46
12.6 Alerting Design . . . 46
12.7 Runbooks and Recovery Drills . . . 46
12.8 Incident Command and Postmortems . . . 47
12.9 Participant Communication and Regulatory Obligations . . . 47

13 Corporate Actions, When the Instrument Changes 49

13.1 Stock Splits . . . 49
13.2 Dividends . . . 49
13.3 Mergers and Acquisitions . . . 49
13.4 Symbol Changes and Delistings . . . 49
13.5 Why Corporate Actions Matter to Developers . . . 50

14 Options Mechanics: Exercise, Assignment, and Expiry **51**
14.1 A Brief Recap . . . 51
14.2 Exercise: The Holder's Decision . . . 51
14.3 Assignment: The Writer's Obligation . . . 51
14.4 Cash Settlement vs. Physical Delivery . . . 52
14.5 A Worked Example: Pin Risk . . . 52

15 Determinism, Replay, and Persistence, The Exchange Must Not Forget **55**
15.1 Deterministic Replay . . . 55
15.2 Sequence Numbers . . . 55
15.3 Persistence and Recovery . . . 56

16 Reference Data, The Exchange's Ground Truth **57**
16.1 What Reference Data Contains . . . 57
16.2 Why Reference Data Is So Dangerous to Get Wrong . . . 58
16.3 Reference Data as an Exchange Engineering Problem . . . 58

Glossary **59**
A . . . 59
B . . . 60
C . . . 61
D . . . 63
E . . . 64
F . . . 64
G . . . 65
H . . . 66
I . . . 66
K . . . 67
L . . . 68
M . . . 68
N . . . 70
O . . . 70
P . . . 71
Q . . . 72
R . . . 72
S . . . 73
T . . . 75
U . . . 76
V . . . 76
W . . . 76
Z . . . 77

References **79**

Exchanges own technical references **83**
CME Globex Matching Engine & Implied Logic . . . 83
Cboe US Options Complex Book Architecture . . . 83
Intercontinental Exchange (ICE) Order Books . . . 83
NYSE Pillar (Equity Trading Platform) . . . 84
Further Sources Cited in the Text . . . 84
Notes on Sources Not Cited Inline . . . 85

Back-Cover Text **87**

About the Author **89**

Preface to the Second Edition

This book is still for the same person: the developer who gets dropped into an exchange codebase full of words like “NBBO”, “LULD”, “implied in/out”, “ATC”, “SIP”, “MMP”, and quietly wonders whether they have accidentally joined a cult. You have not. You have joined a domain with very old vocabulary, very modern latency budgets, and very expensive failure modes.

The first edition focused on giving the reader a map. This second edition keeps the same promise, almost no code-first deep dives, no equations for their own sake, just concepts that make the system legible, but the map is now wider, denser, and more practical. Compared with the first edition, this edition adds well over a thousand lines of new material and rewrites key chapters that were a bit too brief in the first pass where obviously my aspiration to keep it short sometimes conflicted with the need for completeness and understandability. Some pseudo-code has in addition been added where it helps clarify the mechanics, but the focus is still on concepts, not implementation. The goal is still to give you a mental model that lets you reason about the system, not just translate jargon.

The largest expansion is in the implied-orders material, which has been rebuilt from a short overview into two full chapters placed at the end of Part II, once the order book and the matching engine are familiar. The first is a complete walkthrough: the spread sign convention, a vector model from which every implied price formula follows, quantity, depth and priority rules, tick alignment, market-data publication, a first-generation algorithm, engineering invariants, and self-check exercises with worked solutions. The second follows chains of implied orders across many contract months and shows why every venue must limit how far implication is allowed to reach. The order-book chapter has also been expanded from concept to implementation shape: cache behavior, memory layout, single-threaded matching partitions, pool allocation, and step-by-step insert/cancel/sweep mechanics, including determinism concerns.

Several missing bridge topics are also now included explicitly. Part I this time around closes with listing (including the part black-magic art of IPO valuation) and delisting mechanics (initial and continued standards, cure periods, reverse splits, voluntary delisting, direct listings, and SPAC pathways). Part II adds round-lot vs odd-lot treatment and a dedicated tick-size chapter (Rule 612, historical fraction pricing, Tick Size Pilot, and midpoint rounding). The auction material has been upgraded with auction-only order types, opening/closing cross mechanics, and manipulation-at-the-close context.

Beyond core matching, this edition also adds the operational and structural topics developers eventually run into: benchmark integrity (LIBOR and FX fix case studies), trade busting and clearly erroneous handling, conformance testing and onboarding drift, market-data economics (SIP vs proprietary feeds, locked/crossed markets), fixed-income microstructure, crypto venue differences, and options expiry mechanics (exercise, assignment, settlement type, and pin risk).

So yes, the document is longer now. That is deliberate. The first edition tried to keep every section short; this one tries to keep every section useful when things get messy in production. If the first edition was the “day one” guide, this one aims to be the “month six” guide, the point where you are no longer asking what a term means, but why the system was designed that way and what breaks if it is wrong.

As before, this is an introduction, not a rulebook. Exchange rulebooks and venue technical specifications remain the final authority. But if this second edition helps you move from terminology to reasoning faster, it has done its job. The reference section includes the publicly available reference material many details in this books draws from to be as exact as possible.

Reasonable efforts have been made to ensure accuracy, but, it would be foolish of me to believe that I during some late night/early morning while editing this book didn't swap some bid/ask columns or simply forgot a whole section I had in my notes or my mind.

Enjoy the pleasure of finding things out and the triumph of being sure that the author was proveable wrong! So, any typos, missing signs, bad explanations/grammar or faults otherwise, are mine and mine alone. I have no-one to blame!

If you find any mistakes or have other suggestions, please let me know at *johan162@gmail.com*. I genuinely appreciate it.

Johan Persson, Second Edition, Autumn 2026, Järnboås Socken

Preface to the First Edition

This is a book for anyone who finds themselves having been asked to work on a financial exchange system, and yet, has no financial background. Scary. The codebase is full of words like "bid," "ask," "FOK," "the book," "OCO," "circuit breaker," "drop copy," and "kill switch," and many more. These are not arbitrary names; each one represents a concept that evolved over decades of real-world market operation, regulation, and hard-won lessons about what can go wrong when large amounts of money change hands at high speed. You just happened to be "unlucky" enough to start when all these terms have already been defined. Had you instead started 50-ish years ago the terms did not yet exist and you would have seen them come to life organically.

My goal with this book is simple: give you the conceptual vocabulary so that when someone says "GW01 submitted a GTC iceberg order that triggered a circuit breaker halt," you nod instead of panic. No code. No math. Just the map of the territory your code inhabits.

I have tried to keep things brief: each concept gets only as much space as it needs to be useful, and not a page more. Where I have necessarily oversimplified, you will find pointers at the end to the proper academic treatments (hundreds of pages of the hard stuff) for those rare evenings when you feel brave (or masochistic) enough to dive deeper. I make that promise with a clear conscience, because I spent rather too many evenings, and far too many weekends, working through the classic finance texts listed in the references, which served as the primary sources for this little book. If I am honest, this booklet is really nothing more than my own "lecture notes" from all that reading, tidied up and handed to you so you can skip the long way round. In the reference section you can find micro-reviews of all the books I used and taken inspiration from. You can also discover my own personal favourite!

This is an introduction, not a specification. Real exchanges are far more complex, and their official rulebooks are the final word. Think of this as the document I wish someone had handed me on day one I started to work in this world. I hope it helps you get up to speed faster than I did.

Reasonable efforts have been made to ensure accuracy, but as usual any remaining errors are entirely my fault. If you find mistakes or have suggestions, please let me know at *johan162@gmail.com* I genuinely appreciate it!

Johan Persson, First Edition, Winter 2025/2026, Järnboås Socken

PART I

Part IV: Technology, Infrastructure, and Market Ecology

The engineering stack, distributed-systems concerns, market data, and the broader market ecology in which the exchange operates.

Historic Notes

Early in 1987, an inspector from NASDAQ visited the office of **Thomas Peterffy**, a Hungarian-born engineer who had become a prolific options and stock trader. Peterffy had done something the market had never anticipated: he had wired his NASDAQ terminal directly to an IBM computer, which generated and submitted orders automatically from his pricing models. The inspector's verdict was blunt: the terminal had to be disconnected from the computer, and all orders had to go through the keyboard, typed one by one, just like everyone else's.

Peterffy complied, after a fashion. In a frantic week, he and his engineers built a machine that read the prices off the terminal's screen and physically pressed the keys: a robot typist driven by the same computer. The orders now went "through the keyboard", exactly as required. Peterffy went on to found Interactive Brokers and became a billionaire, and he had glimpsed the future of market infrastructure. Every exchange in the world is now built on exactly the principle NASDAQ once tried to prohibit: automated, computer-generated orders, submitted at machine speed, processed by machine logic, without a human hand in the loop [Christopher Steiner, "How the Nasdaq Got Hacked", *Forbes*, 30 August 2012; adapted from his book *Automate This*, 2012].

Part IV is the technical story of how that future was built. It covers the gateway that receives those automated orders, the matching engine that processes them in microseconds, the market data infrastructure that broadcasts results to thousands of subscribers simultaneously, and the resilience, latency, and operational systems that keep it all running, day after day, from one opening bell to the next.

Part Summary:

Examine the engineering reality of exchanges at scale: deterministic engines, messaging and market-data distribution, resilience patterns, and the fragmented multi-venue environment where routing and latency shape outcomes.

Learning Objectives:

- Understand the roles of gateways, matching engines, buses, and subscribers.
- Evaluate resilience strategies such as replication, failover, and site architecture.
- Explain how market data sequencing, replay, and snapshots preserve correctness.
- Relate routing, venue fragmentation, and latency design to execution quality.

Content:

- Speed Bumps, Leveling the Playing Field
- The Technology Architecture
- Conformance Testing and Onboarding
- Primary and Secondary Sites, Resilience Architecture
- Load Balancing, Distributing the Work
- Market Data Architecture, How the Market Sees Itself
- Market Data Economics
- Smart Order Routing and Market Fragmentation
- Fixed Income
- Cryptocurrency and Digital Asset Venues
- Latency and Co-location, The Speed Dimension
- Operational Observability and Incident Response
- Corporate Actions, When the Instrument Changes
- Options Mechanics: Exercise, Assignment, and Expiry
- Determinism, Replay, and Persistence, The Exchange Must Not Forget
- Reference Data, The Exchange's Ground Truth

CHAPTER 1

Speed Bumps, Leveling the Playing Field

Not all exchange participants operate on equal footing. High-frequency trading (HFT) firms invest heavily in co-location, low-latency network links between venues, and specialised hardware to be a few microseconds faster than competitors. Being faster often means seeing price changes and reacting before others, which can be profitable but is also controversial from a market-structure perspective.

1.1 What Is a Speed Bump?

A **speed bump** is a deliberate artificial delay introduced by the exchange to incoming orders. A fixed delay (for example, 350 microseconds) does **not** erase relative timing differences: if one order is 1 microsecond earlier before the delay, it is still 1 microsecond earlier after the delay. What the speed bump can do is reduce the value of ultra-small latency edges and make some short-horizon race strategies harder, especially when combined with venue rules on cancels, quote updates, and matching behavior.

1.2 IEX: The Most Famous Speed Bump

IEX (Investors Exchange), founded as a company in 2012, began operating as a dark pool in 2013, and became a registered national securities exchange in 2016, introduced the speed bump concept to mainstream exchange operation [6]. IEX routes all orders through 38 miles of coiled fibre-optic cable (housed in a small box called the "magic shoe") before they reach the matching engine. The cable introduces a fixed 350-microsecond delay.

IEX's founders argued in the book *Flash Boys* (Lewis, 2014) [6] that speed advantages primarily benefit HFT firms at the expense of long-term investors. The speed bump was their answer. IEX gained regulatory recognition as a licensed national securities exchange and stimulated substantial regulatory and academic debate about speed bump design, though it has consistently captured a small fraction of total US equity volume (around 2–3%) rather than displacing the established venues [1]. Its significance lies more in having demonstrated that speed bump mechanisms are legally and operationally viable, and in influencing ongoing market structure policy discussions, than in market share.

1.3 Speed Bumps in Broader Design

Speed bumps can also be **asymmetric**: applied to some message types and not others. The common modern design delays *aggressive* orders (those that would trade immediately on arrival) while letting passive orders and cancellations through at full speed. The purpose is to protect liquidity providers from **latency arbitrage**: when the market moves, a fast trader races to hit a market maker's now-stale quote before the market maker can cancel it. Delay the aggressor by a millisecond or two, and the market maker's cancel wins the race. Supporters argue this lets market makers quote tighter and larger, because they are picked off less often; critics argue it gives one class of participant a structural privilege and invites "fading" (quotes that vanish just as you try to trade with them).

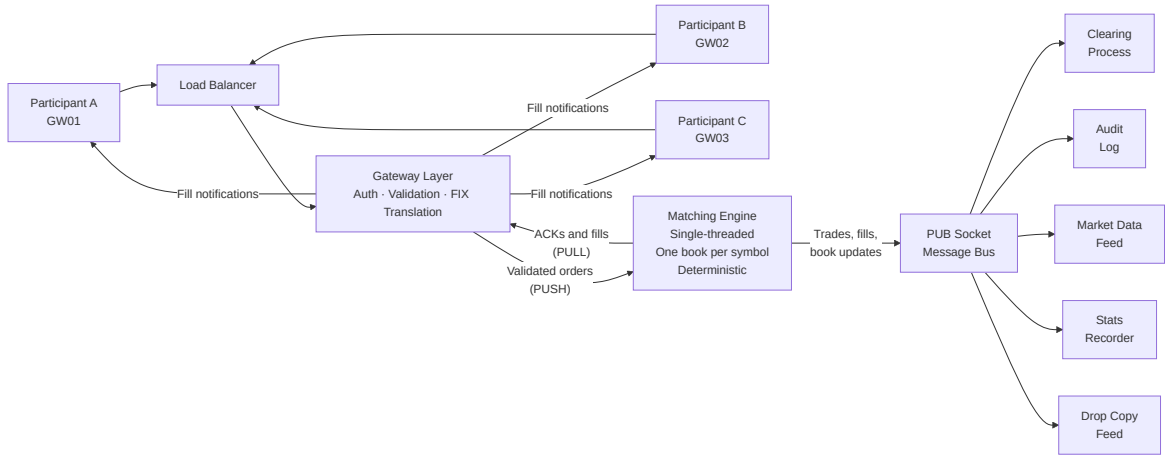
Real examples are no longer rare. Canada's TSX Alpha Exchange introduced a randomised delay of 1 to

3 milliseconds on most incoming orders in 2015, exempting post-only orders. Eurex introduced **Passive Liquidity Protection** in 2019, delaying "all aggressive orders" by 1 millisecond in German equity options and 3 milliseconds in French equity options, and by a product-specific time in certain FX futures, before they can interact with the book [Eurex, *Passive Liquidity Protection*]. The design space is wide: symmetric or asymmetric, fixed or randomised, all messages or only some, and every choice shifts advantage between participants in a way the exchange must be able to defend to its regulator.

CHAPTER 2

The Technology Architecture

Now let us look at how all these concepts are realised in a production technology stack.



2.1 The Gateway

The **gateway** is the participant-facing interface, the "door" through which participants connect to the exchange. Each participant connects through their assigned gateway. The gateway's responsibilities:

- **Authentication:** Verifying the participant's identity and credentials.
- **Session management:** Tracking active connections, handling reconnections.
- **Message translation:** Converting the participant's orders from their format (FIX protocol, binary, or text commands) into the exchange's internal format.
- **Basic validation:** Checking that mandatory fields are present, that the symbol is valid, that quantities are positive.

In many architectures, the gateway performs only structural and session checks, while instrument-specific validation (tick alignment, price collars, and auction/session constraints) is enforced in the matching engine or in a dedicated risk layer fed by reference data. Some venues do cache reference data in gateways and validate there too; the key requirement is consistent enforcement in exactly one authoritative path.

FIX Protocol: The industry-standard format for order submission is FIX (Financial Information eXchange), a text-based protocol developed in the early 1990s. A FIX order message looks something like:

```
8=FIX.4.2|35=D|49=CLIENT1|56=EXCHANGE|11=ORDER001|55=AAPL|54=1|38=100|44=150.30|40=2|59=0
```

(Note: the `|` separator here is a display convention. On the wire, FIX uses the SOH character, ASCII value 1, a non-printable control character, as the field delimiter. Tools and documentation almost always substitute `|` or `^A` for readability.)

All eleven fields in this message have meaning. (A complete message on the wire also carries a few mandatory

housekeeping fields, such as BodyLength (tag 9), MsgSeqNum (34), SendingTime (52), and a CheckSum (10) at the end, omitted here for readability.)

Tag	Name	Value	Meaning
8	BeginString	FIX.4.2	Protocol version
35	MsgType	D	New order (single-sided)
49	SenderCompID	CLIENT1	Who is sending this message
56	TargetCompID	EXCHANGE	Who should receive it
11	ClOrdID	ORDER001	Client's own order identifier (used to track and cancel)
55	Symbol	AAPL	Instrument to trade
54	Side	1	1=Buy, 2=Sell
38	OrderQty	100	Quantity in shares
44	Price	150.30	Limit price
40	OrdType	2	1=Market, 2=Limit, 3=Stop
59	TimeInForce	0	0=DAY, 1=GTC, 2=At the Opening, 3=IOC, 4=FOK, 6=GTD, 7=At the Close

Most exchanges accept FIX for order entry, and many also offer faster binary protocols, described next. (For market data, FIX has a compressed encoding called FAST, "FIX Adapted for STreaming", although most venues now use their own binary formats.) A simplified FIX-inspired text format for internal gateway commands might look like: `NEW|SYM=AAPL|SIDE=BUY|TYPE=LIMIT|QTY=100|PRICE=150.30`.

2.2 Binary Protocols: ITCH and OUCH

FIX is human-readable and widely compatible but inefficient for high-throughput, low-latency environments. Every order message is a text string that must be scanned for field separators and parsed into numeric types. At a million messages per second, this overhead matters.

Production exchanges use **binary protocols** for the critical paths. The most widely deployed are those developed by NASDAQ and adopted or adapted by many exchanges globally:

NASDAQ ITCH (market data) is a binary UDP-based protocol for publishing the full order book feed. Instead of text like `35=D|44=150.30`, an ITCH message is a fixed-width binary structure. The 44-byte ITCH 5.0 trade message, for example, contains a message type (1 byte), a stock locate code and a tracking number (2 bytes each), a timestamp in nanoseconds since midnight (6 bytes), the order reference number (8 bytes), side (1 byte), shares (4 bytes), stock symbol (8 bytes, space-padded), price (4 bytes, an integer with four implied decimal places), and a match number (8 bytes). An ITCH message is 2–5× smaller than its FIX equivalent and requires no text parsing; the fields are at fixed offsets and read directly as integers. ITCH 5.0 is the current version and is publicly documented by NASDAQ. Other venues publish protocols in the same spirit, the London Stock Exchange's MITCH, for example [NASDAQ ITCH 5.0 specification].

NASDAQ OUCH (order submission) is the binary counterpart for order entry. Where FIX is session-based and feature-rich, OUCH is minimal: an "Enter Order" message is a fixed binary layout of a few dozen bytes. OUCH is transmitted over TCP (for reliability) in the same way FIX is, but the compact binary format dramatically reduces serialisation overhead.

CME MDP3 (Market Data Platform 3) is CME Group's binary market data protocol, based on the **SBE (Simple Binary Encoding)** standard. SBE is schema-driven: a protocol specification file defines

every message type's field layout, and code generators produce optimised parsers for multiple languages. The generated parsers decode directly from the raw buffer with no heap allocation, which is critical for deterministic latency. CME's GLOBEX matching engine publishes all market data in MDP3 format [CME Group MDP3 specification].

For exchange developers, the practical implication is that systems dealing with high-volume data, market data normalisation, risk engines, algo trading systems, will almost always use binary protocols on the critical path, with FIX reserved for session management and less latency-sensitive channels.

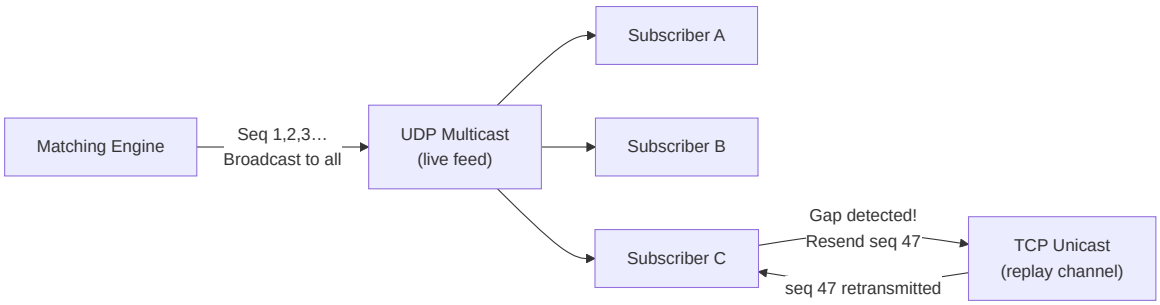
2.3 TCP vs UDP: Why Market Data and Order Submission Use Different Transports

Order submission uses **TCP** (Transmission Control Protocol). TCP guarantees reliable, in-order delivery: if a packet is lost, TCP automatically retransmits it. For order management, this is essential, a lost order acknowledgement or fill notification must eventually arrive. TCP also provides backpressure: if the receiver falls behind, the sender slows down automatically.

Market data uses **UDP multicast**. UDP (User Datagram Protocol) has no reliability guarantee, packets can be lost and are not retransmitted automatically. But UDP's advantages are significant:

- **One-to-many delivery:** A single UDP multicast packet can be received by hundreds of subscribers simultaneously. With TCP, the exchange would need to maintain a separate connection and send a separate copy of each message to each subscriber, which does not scale.
- **No connection overhead:** UDP is connectionless; there is no TCP handshake, no per-connection state.
- **Lower latency:** No TCP acknowledgement flow, no retransmission delay.

How does UDP's unreliability get handled? Through **sequence numbers**. Every ITCH/MDP3 message carries a sequence number. If a subscriber receives message 1000 followed by 1002, it knows 1001 was lost. It requests a retransmission via a separate **TCP unicast recovery channel** (specifically for replay requests). This hybrid design, UDP multicast for the live feed plus a separate recovery path (a TCP replay service, or periodically published snapshots on their own multicast channel), is standard across major exchange market data architectures.



Order submission uses the reverse pattern, TCP from participant to exchange, because every order must be reliably received and acknowledged.

2.4 The Matching Engine

The matching engine is the exchange's core computational component: the software that receives every incoming order, maintains the state of all order books, and executes trades when compatible buy and sell orders can be paired. It sits at the centre of the architecture, receiving from all gateways through a single serialised input queue and publishing results to all subscribers. A single-threaded design, one book per symbol, and deterministic processing are its defining characteristics, each discussed in detail in the *Matching Engine* section above. The gateway is the entry point into this pipeline; the engine is where the actual work happens.

2.5 The Message Bus

The matching engine communicates with all other components through a **message bus**, a pub/sub (publish/subscribe) messaging system. The engine publishes events (trades, order status changes, book snapshots) to topics, and subscribers connect to the topics they care about.

ZeroMQ (ZMQ) is a commonly used messaging library for educational and smaller production implementations. The engine can use a PUSH/PULL socket pair with the gateways (for commands flowing in) and a PUB socket (for market data and events flowing out to subscribers). In high-performance production environments, alternatives purpose-built for financial infrastructure include **Aeron** (developed by Real Logic, used by several exchanges and clearing houses for its deterministic low latency), **Chronicle Queue** (a persistent, low-latency IPC mechanism), and custom UDP multicast implementations. The conceptual pub/sub model is the same regardless of the underlying transport.

Topics on the PUB socket might include: - `trade.executed.AAPL`, a trade happened in AAPL - `order.fill.GW01`, a fill event for Gateway 01's participant - `book.snapshot.AAPL`, current state of the AAPL order book - `session.state`, the trading session changed state

Subscribers filter by topic, receiving only the events relevant to them.

2.6 Subscribers

Any process that connects to the engine's PUB socket and processes events is a **subscriber**. Common subscribers:

- **Clearing process:** Tracks positions and P&L for all participants, generates clearing reports.
- **Stats recorder:** Stores OHLCV (Open, High, Low, Close, Volume) statistics to a database for historical analysis.
- **Viewer/Board:** Display processes that show the current order book, recent trades, and market data on screens.
- **Ticker:** A simplified display of the most recent prices.
- **Audit log:** A complete, immutable record of every event, for regulatory purposes.
- **AI trader / algorithmic participants:** Automated trading strategies that observe market data and submit orders.

The key design principle: subscribers are passive receivers. They observe the market through event streams. They do not write to the order book.

2.7 Book Snapshots

Subscribers that start up mid-session need a way to get caught up on current book state without replaying every event since the start of day. The engine periodically publishes **book snapshots**, a complete current view of all resting orders, aggregated by price level, to the market data feed. A subscriber that misses events can simply wait for the next snapshot. How often snapshots are published is a design choice (the reference exchange in this book publishes one every 500 milliseconds per symbol); real venues range from sub-second to tens of seconds, often on a separate snapshot channel.

Snapshots include: - All bid price levels with total resting quantity - All ask price levels with total resting quantity - The last trade price and quantity - Recent trade history (a rolling window of the last N trades) - The current tick size for the symbol (so subscribers can correctly format prices)

CHAPTER 3

Conformance Testing and Onboarding

The *Technology Architecture* section described the gateway, FIX, and the binary order-entry protocols a participant uses to talk to the exchange. It did not describe how a participant is allowed to start using them in the first place. No serious exchange lets a new participant connect a production gateway and start sending live orders on day one. Between "we have a FIX specification" and "this firm is trading real money" sits a formal process of **onboarding** and **conformance testing**, and it is a substantial engineering surface area in its own right.

3.1 Why This Cannot Be Skipped

A new participant's software has never talked to this specific exchange before. It may misinterpret a field, mishandle a rejection, retry in a way that floods the gateway, or simply have a bug that only manifests against this venue's particular message sequencing. Letting an untested system connect directly to production risks exactly the kind of uncontrolled, unexpected behaviour the *Pre-Trade Risk Controls* and *Knight Capital* sections spent so much effort defending against, except originating from outside the exchange's own deployment process rather than inside it. Conformance testing exists to catch integration problems in a safe environment, before they can touch a real order book, a real counterparty, or real money.

3.2 The Onboarding Sequence

A typical participant onboarding follows a fixed sequence, each stage gating the next:

1. **Membership and legal agreement.** The prospective participant (or its sponsoring broker) signs a membership or market-access agreement with the exchange, establishing the legal relationship, fee schedule, and regulatory obligations described in the *Participants* and *Regulatory Surveillance* sections.
2. **Technical onboarding.** The exchange provisions test credentials, network access to a **certification (or UAT, User Acceptance Testing) environment**, and documentation: the FIX or binary protocol specification, a list of supported order types, and, critically, a **conformance test script**.
3. **Conformance testing.** The participant's system executes a prescribed sequence of test cases against the certification environment. Only once every required test case passes does the exchange issue **production credentials**.
4. **Production go-live**, typically with an initial period of closer-than-normal monitoring (sometimes with temporary, tighter risk-control limits) before the participant is treated as a fully established member of the market.

3.3 What a Conformance Test Script Actually Checks

A conformance test script is a checklist of scenarios the participant's system must demonstrably handle correctly, not just "can it connect." Representative test cases include:

- Submit a standard DAY limit order and correctly process the resulting ACK.

- Submit an order that should be rejected by a pre-trade check (an invalid tick-aligned price, see the *Tick Sizes* section, or a quantity above the configured maximum) and correctly process the REJECTED response rather than assuming the order is live.
- Submit an IOC order against a partially available book and correctly process a PARTIAL fill followed by an implicit cancellation of the remainder.
- Submit a cancel-replace and confirm the participant's system correctly tracks the resulting change in queue priority (see *Price-Time Priority*).
- Handle a simulated disconnect and reconnect cleanly, including correctly requesting any missed drop-copy or execution-report messages by sequence number (see *Drop Copy* and *Determinism, Replay, and Persistence*).
- Correctly interpret a session-state message (see the *Trading Sessions* state machine) and refrain from submitting order types that are invalid in the current state, for example, not submitting a plain DAY order expecting immediate matching while the session is in a scheduled auction phase.

None of this exercises anything conceptually new, every behaviour tested is something this book has already described from the exchange's side. Conformance testing is simply the exchange verifying, mechanically and in advance, that the participant's system has implemented the *other* side of each of those behaviours correctly.

3.4 Keeping Certification Meaningful: The Drift Problem

A certification environment is only useful if passing it reliably predicts correct behaviour in production. This creates an ongoing engineering obligation for the exchange itself: the certification environment's protocol version, reference data (tick sizes, session schedules, risk-parameter defaults), and matching behaviour must be kept synchronised with production. An exchange that certifies participants against a stale UAT environment, one still running last quarter's protocol version or last year's tick-size table, is effectively certifying nothing; a participant could pass certification cleanly and still fail against the live venue on day one. This is the same "keep test and production identical" discipline the Knight Capital section identified as the industry's most important lesson, applied here to the exchange's own certification pipeline rather than to a single participant's deployment.

Test symbols. Certification environments typically trade a small set of dedicated **test symbols** (often deliberately unrealistic-looking tickers, so they can never be confused with a real listed instrument) rather than mirrored copies of real production symbols, both to avoid any risk of test activity leaking into real market data and to give the exchange full control over the synthetic order flow a new participant's system is tested against.

Recertification on exchange-side change. The relationship runs in both directions. When the exchange itself changes something a connected participant depends on, a new order type, a modified FIX tag, an updated risk-check behaviour, affected participants are typically required to recertify against the change in the UAT environment before it is enabled for them in production. This is the direct mitigation for the deployment-verification failure at the heart of the Knight Capital story: a coordinated recertification requirement forces both sides of a protocol change to confirm, in a safe environment, that they agree on the new behaviour before it goes live with real money behind it.

3.5 Offboarding

Onboarding has a mirror image. When a participant's membership ends, whether by choice, by firm closure, or by regulatory action, the exchange must formally revoke gateway credentials, ensure all resting orders for that participant are cancelled (functionally similar to the kill switch behaviour described in *Risk Controls*, but as a permanent rather than temporary state), and confirm final position and clearing reconciliation with the participant's clearing broker. An exchange's participant reference data must always answer, unambiguously, "is this gateway ID currently allowed to trade," and that answer must be correct within the same tight latency and reliability guarantees as every other pre-trade check described in Part III.

> **Key idea:** A protocol specification describes what messages are legal to send. Conformance testing is the exchange's mechanism for confirming, before real money is at risk, that a participant's system actually implements the *other side* of every documented behaviour correctly, and it is only meaningful if the certification environment itself is kept faithfully in sync with produc-

tion. Onboarding and offboarding are gateway-side reference-data problems with the same rigour requirements as any other exchange configuration.

CHAPTER 4

Primary and Secondary Sites, Resilience Architecture

A financial exchange is critical infrastructure. If it goes down unexpectedly, due to hardware failure, software bug, network outage, or even a power cut, the consequences are severe: participants cannot trade, prices cannot be discovered, and confidence in the market is damaged. For this reason, exchanges operate with **redundancy**, duplicate systems at separate locations.

Two metrics define the resilience targets:

RPO (Recovery Point Objective): The maximum amount of data loss acceptable. For a matching engine, RPO is effectively **zero**, no committed trade or order acknowledgement can ever be lost. A trade that was confirmed to a participant and then lost in a crash would represent a contractual failure with regulatory and legal consequences.

RTO (Recovery Time Objective): The maximum acceptable downtime. Different systems have different RTOs: a batch reporting system might tolerate hours; a matching engine must recover in **seconds to minutes**. NYSE and NASDAQ publish target failover times; for critical components, modern exchanges target sub-minute RTO. On 8 July 2015, NYSE halted trading in all its listed symbols from 11:32am to 3:10pm, more than three and a half hours, after a software update introduced a configuration incompatibility between its customer gateways and its trading systems. Trading in NYSE-listed stocks continued on other venues throughout, but for the exchange itself the outage was widely considered an unacceptable recovery time, and a benchmark against which subsequent resilience investments were made.

4.1 Primary and Secondary Sites

A **primary site** (also called the **primary** or **production site**) is where the live matching engine runs. All orders are processed here, all trades happen here.

A **secondary site** (also called the **backup**, **disaster recovery site**, or **secondary**) is an identical system at a geographically separate location. Its purpose: if the primary site becomes unavailable, the secondary site can take over and continue matching.

Real exchanges take this very seriously. NASDAQ's US equity matching engines run in a data centre in Carteret, New Jersey; NYSE's in Mahwah, New Jersey; CME Globex's in Aurora, Illinois; and each of them maintains a disaster-recovery site in a geographically separate location, ready to take over if the primary is lost.

4.2 Synchronisation: Keeping the Secondary Current

For the secondary to be able to take over instantly, it must be an exact replica of the primary's state at all times. This means every order, trade, and state change on the primary must be reflected on the secondary with minimal delay.

Approaches include:

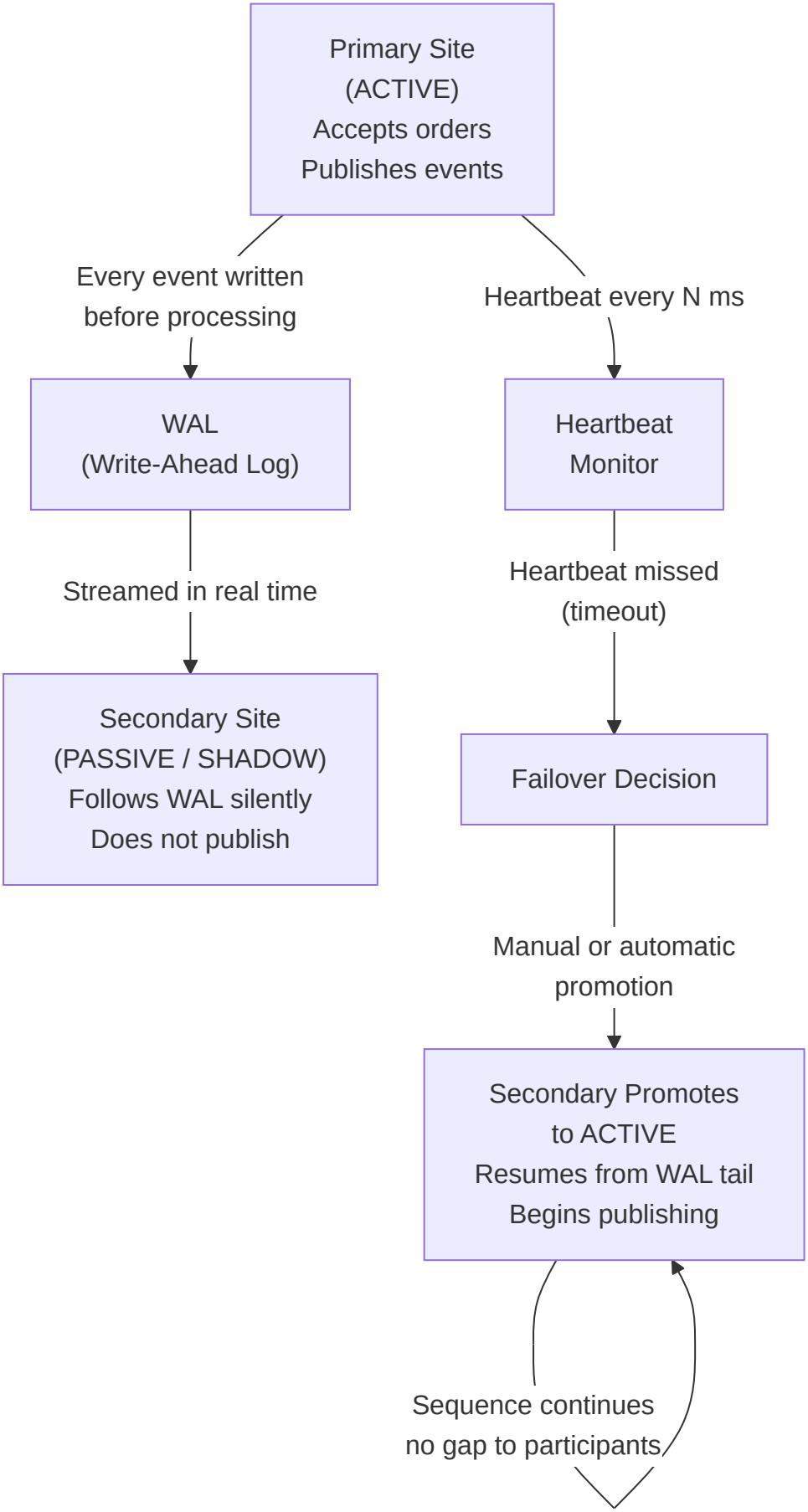
State replication: The primary engine sends every order event to the secondary before or immediately after processing it. The secondary maintains an identical copy of the order book, only it does not publish results or accept new orders from participants, it merely follows along silently.

Log-based replication: Every incoming message is written to a durable **write-ahead log (WAL)** before being processed. The secondary replays from the log. If the primary fails, the secondary catches up to the log's tail and takes over. This is architecturally similar to how database replication works (PostgreSQL's streaming replication, for example, uses this approach).

Consensus protocols: In the most robust designs (used by clearing houses and critical infrastructure), a consensus algorithm like **Raft** or **Paxos** ensures both the primary and secondary agree on every event before it is considered "committed." Nothing is published to participants until both sites confirm they have the event. This guarantees zero data loss on failover but adds latency.

4.3 Failover

Failover is the act of switching from the primary to the secondary. It may be:



- **Automatic (active-passive):** The system detects the primary has stopped responding (via heartbeat monitoring) and automatically promotes the secondary to primary, alerting participants. The challenge:

detecting a true failure vs. a temporary network partition without making hasty decisions, the **split-brain problem**: if both sites think the other is dead, both might try to become primary simultaneously, resulting in two active engines accepting conflicting orders.

The standard distributed-systems solutions are **fencing** (issuing a command that forces the suspected-failed node to stop operating before the secondary takes over, sometimes called **STONITH: Shoot The Other Node In The Head**) and **quorum requirements** (requiring the agreement of a majority of nodes before any single node can assume the primary role). Without one of these mechanisms, split-brain in a matching engine produces a catastrophic situation: two independent order books diverging while both believe they are the authoritative state.

- **Manual (operator-controlled):** A human operator monitors the primary and initiates failover when a problem is confirmed. This is slower (seconds or minutes, not milliseconds) but avoids false failovers.
- **Active-active:** Both sites process orders simultaneously, and a consensus mechanism ensures they produce identical results. Extremely complex but enables zero-downtime failover.

  True active-active matching engines are **rare in production exchanges** because distributed consensus introduces latency, and maintaining deterministic total ordering across two simultaneously-processing sites is a fundamentally hard problem. Most production exchanges prefer active-passive or shadow replication rather than active-active, accepting a brief failover window in exchange for simplicity and predictable latency.

4.4 The Geographic Trade-Off

Placing the secondary site far from the primary protects against geographically localised disasters (earthquakes, floods, building fires). But distance means higher network latency, making synchronisation slower.

A common compromise: a "near" site (same city or region, connected by dark fibre) provides fast synchronisation and fast failover, plus a "far" site (different country or continent) for true disaster recovery.

CHAPTER 5

Load Balancing, Distributing the Work

A single exchange may serve dozens of gateways simultaneously, each handling orders from hundreds of participants. As trading volume grows, a single matching engine process may become a bottleneck.

5.1 Horizontal Scaling Across Symbols

The most natural scaling approach for an exchange is to distribute different symbols across different matching engine instances. The AAPL book and the MSFT book are completely independent, they can run on different servers, different processes, even in different data centres.

This is called **partitioning by symbol**. It scales well: adding a new server lets you handle more symbols, without requiring coordination between servers (since no order crosses symbol boundaries except through combo order handling at a higher level).

5.2 Gateway Load Balancing

Multiple gateway processes can run in parallel, each accepting connections from different participants. Gateways forward orders to the appropriate matching engine instance based on symbol. This allows gateway capacity to scale independently of matching engine capacity.

A **load balancer** sits in front of the gateways, distributing incoming participant connections across the available gateway processes. Participants typically do not know which specific gateway process they are connected to, the load balancer makes this transparent.

5.3 The Unique Order Number Challenge

This is where scaling creates a genuinely complex problem, and it is worth understanding in detail.

Every order must have a **unique identifier**, a number or string that unambiguously identifies it across the entire system, forever. The order ID appears in trade records, audit logs, regulatory reports, and clearing messages. It must be:

1. **Unique:** No two orders should ever have the same ID.
2. **Monotonically increasing:** A later order should always have a higher (or at least not lower) ID than an earlier order. This property is valuable for sequencing, for detecting missed events, and for regulatory audit trails that can be sorted by order arrival.
3. **Globally consistent across sites:** If the primary site generates order ID 1000 and then the secondary takes over, the secondary must not also generate an order ID 1000 for a different order.

The straightforward solution: A single, central counter. Every time an order arrives, atomically increment the counter and assign the result. Order IDs are 1, 2, 3, 4… This is perfectly monotonic and unique.

The problem: A single central counter is a single point of failure. If that counter's database is unavailable, no new orders can be accepted, the entire exchange halts. It is also a bottleneck, every order, from every

gateway, must go through the same counter.

Solution 1: Pre-allocated ranges. The primary site and secondary site each hold a pre-allocated range of IDs. The primary uses 1–1,000,000; the secondary uses 1,000,001–2,000,000. If the primary uses up its range and requests more, the allocation manager assigns the next block. If the primary fails and the secondary takes over, the secondary continues from its own unused range. No IDs collide, and no coordination is needed at order submission time.

The challenge: if the primary used IDs 1–50,000 before failing and the secondary starts from 1,000,001, there is a large gap in the ID sequence. This is not strictly a problem (the IDs are still unique and monotonically increasing) but it can confuse monitoring systems that expect dense sequences.

Solution 2: Site-prefixed IDs. Each site adds a prefix to its generated IDs: `P-000001` for primary, `S-000001` for secondary. IDs are unique because the prefixes differ. This is simple but makes IDs slightly longer, and "comparing" IDs (which one came first?) requires stripping the prefix.

Solution 3: Time-based IDs. Compose the ID from the nanosecond timestamp plus a site identifier plus a sequence number within that nanosecond. For example: `{timestamp_ns}{site_id}{seq}`. This is monotonically increasing (later timestamps = higher IDs), unique per site (different site IDs), and unique within a nanosecond (the sequence counter handles simultaneous orders). Used in various forms by high-volume systems. The challenge: the IDs become very large numbers.

Solution 4: UUIDs. A **UUID (Universally Unique Identifier)** is a 128-bit random or time-based identifier. UUID v4 is randomly generated; UUID v1 is based on time and network address. UUIDs are practically guaranteed unique without coordination. The drawback: they are not sequential, and they are large (32 hex characters). Sorting by UUID does not give you arrival order.

In practice: Exchange systems typically use a hybrid approach, a system like pre-allocated ranges for transaction ordering (where strict monotonicity matters), and UUIDs for order IDs in internal databases (where uniqueness is paramount and ordering can be handled separately by timestamp). A common approach uses UUID-based order IDs combined with monotonically increasing nanosecond timestamps to achieve both uniqueness and chronological ordering.

5.4 Monotonic Timestamps Across Sites

The nanosecond timestamp on each order must be monotonically increasing, not just within one machine, but across the entire system. If the primary is generating timestamps and then the secondary takes over, the secondary cannot start generating timestamps smaller than the last one the primary generated.

This requires clock synchronisation (NTP or PTP), monotonic enforcement via a wrapper that guarantees strictly increasing values, and site coordination when the secondary takes over.

An important nuance: in practice, deterministic sequencing identifiers matter more than absolute wall-clock timestamp precision. Production exchange systems frequently tolerate small amounts of clock drift and same-timestamp collisions, because the canonical ordering of events is defined by **sequencing infrastructure**, the order in which the single-threaded engine processes messages from its queue, not by the wall-clock value of the timestamp. Timestamps are primarily for auditability and observability; the sequencing of the input queue defines what actually happens. The monotonic enforcement described here prevents the most obvious failures (a clock jumping backward and assigning a "past" timestamp to a new order), but it is the engine's input queue, not the clock, that ultimately determines priority.

CHAPTER 6

Market Data Architecture, How the Market Sees Itself

The order book and trade data produced by the matching engine are published as **market data**, real-time feeds that participants and their systems consume to make trading decisions. Market data architecture is a significant engineering domain in its own right.

6.1 Snapshots vs Incremental Updates

A **snapshot** is a complete picture of the current book state: all resting price levels, their quantities, the last trade price, recent trade history. An **incremental update** (also called a **delta**) describes only what changed since the previous message: an order was added, a fill occurred, or a level was removed.

Sending full snapshots is wasteful, most of the book did not change. Production systems send incremental updates continuously and periodic full snapshots (e.g., every second or every 5 seconds) so that subscribers who missed messages can resynchronise.

6.2 Sequence Numbers

Every market data message carries a **sequence number**, a monotonically increasing integer. Subscribers track the last received sequence number. If a message with sequence number 1000 arrives followed by 1002, the subscriber knows message 1001 was lost (a **gap**). The subscriber must either request a replay of message 1001 or wait for the next full snapshot to resynchronise.

Without sequence numbers, a subscriber has no reliable way to detect data loss. A subscriber that silently misses a cancellation event would have a stale view of the book, showing a resting order that no longer exists, and any trading decisions based on that view could be incorrect.

6.3 Sequencing Scope and Session Boundaries

For production feeds, sequence semantics must be explicit in protocol documentation:

- **Scope:** Is sequencing global, per symbol, per channel, or per partition?
- **Reset policy:** Do sequence numbers reset at start-of-day, or continue indefinitely?
- **Wrap behaviour:** What happens at integer limits?
- **Recovery anchor:** Should clients request replay from a sequence number, timestamp, or both?

Without these rules, two clients can implement different recovery logic and both appear correct in normal conditions, then diverge under packet loss or after a session restart.

6.4 Gap Recovery and Replay

Exchanges provide a **replay channel** or **retransmission service**: a subscriber can request resending of specific message sequence numbers it missed. Some exchanges use **multicast** delivery for the live feed (one packet sent to all subscribers simultaneously) and **unicast TCP** for retransmission (individual resend requests).

The replay buffer typically holds the last N seconds of messages, enough to cover a brief network hiccup. A subscriber disconnected for several minutes may not be able to replay from the live feed and must wait for the next full snapshot.

6.5 Conflation

When the exchange is under high load, market data messages may queue behind each other. **Conflation** is the process of merging consecutive updates for the same price level into a single message, reducing message volume at the cost of losing intermediate states. A subscriber using conflated data may not see every individual order addition and cancellation, it only sees the net effect on each price level. This is acceptable for display purposes but not for analytical systems that need complete tick data.

6.6 Tick-to-Trade Latency

Tick-to-trade latency is the time from when a market data message leaves the exchange to when a trading decision based on that message results in an order being submitted and received by the exchange. It is a critical performance metric for market makers and HFT firms. Reducing tick-to-trade latency requires optimising every component in the path: network delivery, data parsing, decision logic, order serialisation, and order routing.

6.7 Top-of-Book vs Depth-of-Book

Some participants subscribe only to **top-of-book** data (best bid and ask prices and quantities only, Level 1). Others subscribe to **depth-of-book** (multiple price levels, Level 2). Depth data requires more bandwidth and processing but enables more sophisticated analysis of near-term price pressure.

> **Key idea:** Market data is not just a convenience, it is the primary input to every trading algorithm and market maker. Its correctness, latency, and sequencing directly affect the quality of market participants' decisions. Exchange developers must treat market data publishing with the same rigour as the matching engine itself.

CHAPTER 7

Market Data Economics

The *Market Data Architecture* section explained how market data is technically produced and distributed, snapshots, incremental updates, sequence numbers, conflation. This section covers something quite different: market data as a business, one that has become a major, and increasingly contested, share of exchange revenue, and a recurring subject of regulatory intervention.

7.1 Two Tiers of the Same Information: SIP vs. Proprietary Feeds

In US equities, every trade and quote from every registered exchange is consolidated into a single public feed, the **Securities Information Processor (SIP)**, introduced briefly in the *Smart Order Routing* section's discussion of trade reporting. The SIP is the feed that ultimately reaches retail brokerage apps and financial news tickers, and its content and pricing are governed by the exchanges collectively, under SEC oversight.

Separately, each exchange also sells its own **proprietary direct feed**, NASDAQ TotalView, NYSE OpenBook, CME's MDP3 (already introduced in the *Technology Architecture* section), and equivalents at every other venue. A direct feed typically offers two advantages over the SIP: it carries **more information** (full depth-of-book rather than just top-of-book, and historically, and controversially, sometimes odd-lot and auction-imbalance data not present on the SIP at all), and it arrives with **lower latency**, because it is generated and published directly by the venue's own matching engine rather than being aggregated across venues by a separate SIP processor, a step that necessarily adds processing time.

That latency difference is not academic. Recall from the *Latency and Co-location* section that a market maker whose information is even microseconds stale is exposed to adverse selection. A firm subscribing only to the SIP is, by construction, always looking at the market slightly later than a firm subscribing to every exchange's direct feed and its own co-located infrastructure. This means the "public" price and the "fastest available" price are not quite the same thing, and which one a market participant sees depends directly on how much market-data infrastructure they have bought. This is one of the more concrete ways market structure creates a real, monetised speed hierarchy on top of information that is nominally public.

7.2 Market Data as Exchange Revenue

Exchanges earn revenue from three broad sources: **trading fees** (the maker-taker and taker-maker structures described in *Smart Order Routing*), **listing fees** (described in *Listing and Delisting Mechanics*), and **market data and connectivity fees** (direct feed subscriptions, co-location rack space, and cross-connect charges). Over the past two decades, data and connectivity revenue has grown to represent a substantial and increasingly important share of major exchange groups' overall profit, in some cases growing faster than trading-fee revenue, even as headline trading fees themselves have compressed under competitive pressure from the fragmentation described in Part IV.

This creates a structural tension worth naming explicitly: the exchanges that jointly govern and price the *public* SIP are the same companies competing to sell *proprietary* feeds that outperform it. A pricing or content decision that makes the SIP more competitive with proprietary feeds directly reduces the commercial appeal, and the revenue, of the venue's own premium product. Critics (institutional investors, retail-broker

associations, and academic market-structure researchers) have long argued this is an inherent conflict of interest; exchanges counter that data fees fund the substantial technology investment described throughout this Part, and that customers voluntarily choose to pay for a faster or richer product.

7.3 The Regulatory Response

The SEC's **Market Data Infrastructure** rulemaking (finalized 2020, with phased implementation through the following years) was a direct response to this tension. Its core changes: expanding the *content* of the SIP itself to include some previously proprietary-only data (round-lot depth-of-book and, over time, more odd-lot detail, connecting back to the odd-lot discussion in the *Order: The Fundamental Unit* section of Part II), and restructuring the *governance* of how SIP revenues are shared among exchanges and how new data products must be competitively justified. The rule was contested through litigation by exchange operators and remains a live area of regulatory and industry disagreement as of this writing; developers building market-data consumers should treat the exact content and cost of the SIP as something that continues to evolve, rather than a fixed baseline.

Europe's parallel debate under MiFID II centres on the cost and availability of **consolidated tape** data (a SIP-equivalent that, unlike in the US, never existed as a single mandated entity across EU trading venues; the 2024 review of MiFIR finally set up a process to appoint consolidated tape providers, starting with bonds and then equities) and on capping the fees venues can charge for market data, reflecting the same underlying concern: that a market structure built on competing venues needs some guarantee that the *public* picture of the market remains genuinely accessible, and not merely a degraded version of a product every serious participant is expected to pay to upgrade past.

> **Key idea:** Market data is not a free byproduct of trading, it is priced, tiered, and increasingly central to exchange profitability. The gap between the public SIP and an exchange's own proprietary direct feed is a real latency and content gap with real economic consequences for anyone trading against it, and the governance of that gap, who sets SIP content and pricing, when a venue that also sells competing data can be trusted to do so fairly, is an active and unsettled regulatory question, not a solved problem from decades past.

CHAPTER 8

Smart Order Routing and Market Fragmentation

Modern equity markets are not a single exchange. In the US, there are over a dozen registered stock exchanges plus dozens of alternative trading venues, all trading the same stocks. This fragmentation has profound implications for participants and exchange systems.

8.1 Why Markets Are Fragmented

Regulatory choices drive fragmentation. In the US, **Regulation NMS** (National Market System), adopted by the SEC in 2005 and phased in through 2007, prohibited venues from trading through a better displayed price on another exchange, which meant a new venue's quotes could not be ignored, creating strong incentives for new venues to compete with NYSE and NASDAQ. In Europe, the original **MiFID** (2007) had a similar effect by abolishing national rules that had concentrated trading on each country's incumbent exchange. Today a heavily traded US stock trades on every exchange at once, no single venue typically handles more than about a fifth of its volume, and a large share (in recent years close to half) executes off-exchange in dark pools and broker internalisers.

8.2 The National Best Bid and Offer (NBBO)

In the US, regulators require that market participants be offered the best available price across all exchanges. The **National Best Bid and Offer (NBBO)** is the highest available bid price and the lowest available ask price, computed in real time across all exchanges. If AAPL's best bid is \$150.30 on NYSE and \$150.31 on NASDAQ, the NBBO bid is \$150.31. A market sell order must be executed at the NBBO price or better, a broker cannot route it to NYSE at \$150.30 when \$150.31 is available elsewhere.

8.3 Locked and Crossed Markets

Because the NBBO is assembled from independent quotes published by a dozen or more venues, each with its own network path and processing delay, the quotes a smart order router sees are never perfectly synchronised. Two related, slightly disorienting conditions follow directly from this:

A **locked market** occurs when the best bid on one venue equals the best ask on another, for example, NASDAQ shows a bid of \$150.30 at the same moment CBOE shows an ask of \$150.30. Under normal price-time priority within a single book this could never happen (a marketable order would simply cross and trade), but across two independent venues it can persist for a brief window because neither venue's matching engine can see the other's book directly.

A **crossed market** is more extreme: the best bid on one venue is actually *higher* than the best ask on another, for example, NASDAQ bidding \$150.32 while CBOE offers \$150.30. To a newcomer reading raw market data this looks like a bug, a bid above an ask should be arbitraged away instantly, but it is a real, if fleeting, artefact of fragmentation and propagation latency: an order to buy at \$150.32 on NASDAQ and sell at \$150.30 on CBOE would in fact be a risk-free arbitrage, and this is exactly the kind of few-microsecond opportunity that latency arbitrageurs (see the *Latency and Co-location* section) are built to capture. Regulation NMS requires

exchanges to have policies to avoid deliberately displaying locked or crossed quotes, and persistent locking or crossing by a single venue can trigger regulatory scrutiny, but transient locks and crosses lasting a few microseconds during periods of fast quote updates are a normal, expected feature of a fragmented market, not evidence of a broken system. Developers building smart order routers or market-data consumers should treat brief locked/crossed conditions as routine input to handle gracefully, not as an error state to reject.

8.4 Smart Order Routing (SOR)

A **Smart Order Router (SOR)** is software that determines how to route an order across multiple venues to achieve the best overall execution. For a large buy order, the SOR might: 1. Route 200 shares to IEX (cheapest in fees). 2. Route 500 shares to NASDAQ (deepest at the best ask). 3. Route 300 shares to CBOE (additional depth at the next price level). 4. Hold back remaining shares pending fills.

The SOR must evaluate venue fees, available depth, likely price impact, and speed simultaneously. A naive router that always sends everything to the same exchange would often leave better prices on the table elsewhere.

In real implementations, SOR logic also incorporates:

- **Protected-quote obligations** (for example, US trade-through constraints)
- **Queue position probability** (expected fill likelihood if posted passively)
- **Venue toxicity signals** (adverse-selection risk by venue and time of day)
- **Hidden-liquidity interaction** (dark midpoint opportunities versus lit certainty)

These constraints are why production SORs are stateful optimisation systems rather than simple price sorters.

8.5 Dark Pools and Hidden Liquidity

A **dark pool** is a trading venue, often operated by a bank or broker-dealer, where orders are not displayed in a public order book. Participants submit orders to the dark pool and are matched against other participants' dark pool orders, typically at the midpoint of the NBBO. The trade is only publicly reported after it occurs.

Why would anyone use a dark pool? A large institutional investor trying to buy 1 million shares knows that showing their intention in a lit (public) order book would immediately move prices against them as other participants react. By routing to a dark pool, they hide their intent and may receive fills at better prices with less market impact.

The trade-offs: dark pools offer less market impact for large orders but provide no pre-trade transparency (you cannot see what orders are resting, or whether a counterparty exists at all). **Lit markets** (public order books on regulated exchanges) offer full pre-trade transparency but more information leakage.

The existence of dark pools explains why the displayed order book is not the whole market. At any moment, significant liquidity may be available in dark venues that is invisible to standard Level 2 data.

> **Key idea:** When working on exchange software, be aware that "the market" is larger than "the exchange." Smart order routing, fragmented liquidity, and dark pools are the reality in which the exchange you are building operates. Features like the NBBO, order routing decisions, and market data aggregation all exist because of fragmentation.

8.6 Payment for Order Flow (PFOF)

Payment for Order Flow (PFOF) is a practice in which retail brokers sell their clients' order flow to wholesale market makers, large firms such as Citadel Securities, Virtu Financial, and G1 Execution Services, rather than routing orders to lit exchanges. The market maker pays the broker a per-share fee for the right to execute the orders internally.

Why do market makers pay for retail flow? Retail orders have low **adverse selection risk**, retail investors are statistically less likely to be trading on superior information than institutional investors or HFT firms. A market maker can fill a retail order at or near the NBBO and earn the spread with relatively low risk of being "picked off." The retail flow is essentially a stream of low-risk, profitable execution opportunities.

Why do brokers accept PFOF instead of routing to exchanges? The payment from market makers subsidises the broker's operations, enabling the zero-commission trading that many retail brokers now offer. Platforms like Robinhood derive a significant fraction of their revenue from PFOF.

The controversy. PFOF supporters argue that retail orders receive **price improvement**, fills at prices better than the NBBO bid or ask, because market makers compete for the flow. Critics argue that by routing to market makers rather than lit exchanges, retail orders never contribute to public price discovery; the market maker captures the profit that would otherwise benefit the investor through tighter spreads; and the broker has a structural conflict of interest (paid to route to a market maker, not to find genuinely best execution).

PFOF is effectively **banned in the United Kingdom**, where the regulator concluded in 2012 that it conflicts with a broker's duty to its clients, and the **European Union**'s 2024 review of MiFIR bans it EU-wide, with the ban applying from 30 June 2026. In the US, the SEC proposed significant restrictions (including an "Order Competition Rule" that would have forced many retail orders into auctions) in December 2022, but faced substantial industry opposition, and formally withdrew those proposals in June 2025. PFOF therefore remains legal in the US. Developers building exchange or brokerage infrastructure should monitor this area, as the regulatory position keeps changing.

8.7 Dark Pool Regulatory Scrutiny

Dark pools are legal but have attracted significant regulatory attention when operators have misrepresented how they work. In 2014, the New York Attorney General's office and the SEC brought cases against major dark pool operators:

Barclays agreed to pay $70 million in 2016 to settle allegations that it misled clients about the presence of high-frequency traders in its dark pool (LX Liquidity Cross), claiming to offer protection from HFT while actually allowing aggressive HFT firms to trade against institutional clients [SEC v. Barclays, 2016].

Credit Suisse paid $84.3 million to settle similar allegations about its CrossFinder dark pool in the same period.

These cases established that dark pool operators have affirmative obligations of transparency to their clients about pool membership and execution policies, not just about prices. For exchange developers, the lesson is that any venue offering dark or non-displayed liquidity must have defensible, auditable, and accurate representations of its matching rules and participant population.

8.8 Exchange Fee Models: Maker-Taker and Taker-Maker

Understanding how exchanges charge for trading is essential for anyone building SOR logic, because fee differences between venues directly affect routing decisions.

Maker-taker is the dominant fee model among US equity exchanges. It works as follows:

- **Makers** (participants who post resting limit orders, providing liquidity) receive a **rebate**, the exchange pays them a small amount per share, historically around $0.0020–$0.0030.
- **Takers** (participants who submit aggressive orders that execute against resting orders) pay a **fee**, historically up to $0.0030 per share, the maximum "access fee" Regulation NMS allowed for stocks priced at $1 or more.
- The exchange retains the difference as its revenue.

This model incentivises liquidity provision: market makers are paid to quote, and the payment compensates partly for adverse selection risk. NYSE Arca and NASDAQ use maker-taker structures. A SOR routing a large aggressive order that sweeps through multiple levels pays taker fees on every share executed: for a million-share order at $0.0030 per share, 1,000,000 × $0.0030 = **$3,000**. That is small relative to the value of the shares, but on a routing desk handling billions of shares a month it is real money, and it is often comparable to the spread a patient order might have saved, which is why fee comparison between venues is a significant input to routing decisions. (In September 2024 the SEC adopted amendments cutting the access-fee cap to $0.0010 per share and introducing a half-penny tick for many heavily traded stocks; the rule survived a court challenge in October 2025, but its implementation timetable has since been under review, so check its current status before hard-coding any fee.)

Taker-maker (sometimes called the **inverted model**) reverses the incentives: takers are paid a rebate and makers are charged a fee. This sounds counterintuitive, but it attracts aggressive order flow from participants who want to execute immediately and are willing to pay to provide that flow to the maker side. Cboe BYX and Nasdaq BX are examples of inverted venues. Inverted venues are often used for orders in highly liquid symbols where the maker-taker economics of the dominant venues create distortions.

Zero-fee models: Some venues, particularly in the EU, charge neither makers nor takers a per-trade fee, instead monetising through subscription data fees, co-location charges, or flat access fees. Aquis Exchange operates on a subscription model.

For SOR logic: a venue with a large rebate for makers may be preferred for posting passive orders even if its spread is fractionally wider, because the rebate income offsets the spread cost. This creates visible patterns in routing decisions that can appear irrational without understanding the fee structure.

Best execution is the regulatory obligation for brokers and investment firms to take all reasonable steps to achieve the best possible outcome for their clients when executing orders. "Best" is not simply the highest price or lowest cost in isolation, regulators define it as the best overall result considering price, execution costs, speed, likelihood of execution, market impact, and other relevant factors.

In the EU, best execution is mandated by MiFID II, which requires firms to maintain an order execution policy, disclose it to clients, and regularly monitor whether it actually delivers the best result. In the US, the SEC's duty of best execution (codified in FINRA Rule 5310 for broker-dealers) has similar intent.

Best execution is the regulatory foundation that makes smart order routing necessary. Without a best execution obligation, a broker could route all orders to the venue that pays the highest PFOF kickback, regardless of the execution quality. Best execution compliance creates the legal obligation to have and use a SOR that genuinely seeks the best available outcome for the client, and to document that process.

For exchange developers, best execution manifests in several observable ways: exchanges must be fast (slow fills mean worse prices), transparent (firms need accurate data to compare venues), and competitive on fees. An exchange that is consistently expensive or slow will be deprioritised by SOR systems fulfilling their best execution duty.

8.9 Trade Reporting Obligations

An exchange's matching engine does not operate in regulatory silence. Every executed trade must be reported to regulators and/or public reporting facilities within strict time limits. For exchange developers, these obligations manifest as mandatory downstream subscribers that cannot be missed or delayed.

US equity markets: Trades on registered exchanges are reported automatically by the exchange to the **Securities Information Processor (SIP)**, which consolidates all exchange trades into the public tape. Off-exchange trades (from dark pools, internalised retail flow, or OTC transactions) must be reported by broker-dealers to a **FINRA Trade Reporting Facility (TRF)** within 10 seconds of execution, which then publishes them to the consolidated tape.

EU markets under MiFID II: Investment firms must report every trade to a regulator via an **Approved Reporting Mechanism (ARM)** within T+1, and publish the trade to the market via an **Approved Publication Arrangement (APA)** as close to real time as technically possible (within one minute for shares; publication of very large or illiquid trades may be deferred, in some cases by days). Large exchanges typically operate their own APAs and ARMs as part of their data services.

Derivatives under EMIR (EU) and Dodd-Frank (US): Derivatives trades must be reported to a **Trade Repository (TR)** (called a swap data repository in the US); DTCC and ICE operate some of the largest. Both counterparties must report, or one must be designated to report on both sides.

For exchange developers, these obligations mean the clearing and audit systems must produce reports in multiple formats to multiple regulatory recipients within multiple latency windows, without impacting the matching engine's performance. The reporting infrastructure is a first-class engineering component, not an afterthought.

8.10 Execution Algorithms, Slicing Large Orders

An individual investor buying 100 shares submits a single order. An institutional investor buying 5 million shares cannot. Sending a single 5-million-share market order would sweep through every level of the book, move the price dramatically, and fill at disastrous average prices. Instead, institutions break large orders into many small pieces and execute them over time, typically using standardised **execution algorithms** (called **algos**).

Understanding execution algorithms matters for exchange developers because they generate the vast majority of order flow in real markets. What looks like continuous random order activity in the book is largely the output of algos executing institutional orders. Algos also interact directly with exchange features: smart order routing, dark pool access, iceberg orders, and closing auctions are all used by algos as tools.

VWAP (Volume-Weighted Average Price) algorithm. The most common benchmark for institutional execution. The goal: execute the large order throughout the day such that your average price is close to, or better than, the day's overall VWAP. A VWAP algo estimates the expected volume profile of the stock throughout the day (typically heavier at open and close, lighter mid-day) and participates proportionally, sending more orders during high-volume periods. Performance is measured by comparing the average execution price to the day's VWAP; beating VWAP is good, lagging it is bad.

TWAP (Time-Weighted Average Price) algorithm. Simpler than VWAP: divide the total quantity evenly across equal time slices. If you need to buy 1 million shares over 2 hours with 1-minute slices, send approximately 8,333 shares per minute regardless of volume. TWAP is predictable and easy to verify but leaves performance on the table relative to VWAP in markets with a non-uniform volume profile. Used when the trader wants mechanical simplicity or when the stock is illiquid and has no reliable volume profile.

Implementation Shortfall (IS) / Arrival Price. More sophisticated than VWAP or TWAP. The benchmark is the *decision price*, the market price at the moment the investment decision was made (the "arrival price"). The algorithm minimises the difference between this theoretical price and the actual average execution price. IS algos trade faster when the price is moving away from you (urgency increases to avoid further shortfall) and slower when the price moves in your favour. They adapt dynamically to market conditions rather than following a fixed schedule. IS is often preferred by quantitative funds whose models predict short-term price moves, they want to execute before the predicted move happens, not passively over the full day.

Percentage of Volume (POV) / In-Line. Participate at a fixed percentage of the market's traded volume, for example, "be 10% of whatever the market trades." If the market trades 100,000 shares in an interval, the algo sends 10,000. This limits market impact (you are never the dominant force in the market) and adapts naturally to varying liquidity throughout the day.

> **Key idea:** The vast majority of exchange order flow originates from execution algorithms, not from humans pressing buttons. Understanding how these algos work helps explain patterns visible in market data, clustering of activity near the open and close (VWAP/ATC algos), evenly-spaced order arrivals (TWAP), and bursts of activity when prices move (IS algos increasing urgency).

CHAPTER 9

Fixed Income Market Structure: Where the Order Book Model Breaks Down

Every chapter so far has quietly assumed a particular kind of market: many participants trading a modest number of fungible instruments, meeting in a single continuous central limit order book, matched by price-time priority, with a tight two-sided quote available almost all the time. That model fits equities and listed futures extremely well. It fits most of the **bond market** hardly at all, and a developer who carries the CLOB mental model unmodified into fixed income will build the wrong system. This chapter exists to inoculate against that, by explaining *why* bonds are different and *what structure* they use instead.

The stakes are not marginal. The global bond market is larger than the global equity market, and the U.S. Treasury market alone, the debt of the U.S. government, is the single largest and most important securities market in the world, the risk-free asset against which nearly everything else is priced. Yet for most of its history the Treasury market has had no central limit order book that a retail investor could see, and the corporate bond market still largely does not. Understanding why is understanding a genuinely different market structure.

9.1 Why Bonds Break the Central-Limit-Order-Book Model

Three facts about bonds, none of which is true of a typical stock, jointly dismantle the assumptions behind a continuous order book.

1. Instrument proliferation. A company has, in the ordinary case, *one* class of common stock. That same company may have *dozens* of distinct bonds outstanding, each a separate instrument with its own coupon, maturity date, seniority, covenants, and callability. A single large issuer, or a government, can have hundreds of individual bond lines. Apple has one ticker for its equity; it has issued many separate bond CUSIPs across different maturities and coupons. So where equities concentrate all interest in a handful of instruments per issuer, bonds *shatter* it across a vast number of instruments, each with only a small slice of the total interest.

2. Infrequent trading. The direct consequence of proliferation is that most individual bonds trade rarely, many go days or weeks between trades, and a large fraction of outstanding bonds do not trade on any given day at all. A great deal of the bond market is **buy-and-hold**: insurers, pension funds, and central banks buy bonds to hold to maturity for the income, not to trade them. An instrument that trades a handful of times a month cannot sustain a continuous two-sided order book; there is simply not enough flow to keep resting bids and offers meaningfully populated.

3. No continuous two-sided liquidity to rest. A central limit order book presumes that at almost every instant there are participants willing to post firm resting bids and offers. For an infrequently traded bond there are not. There is no standing crowd; liquidity has to be *summoned* when a trade is wanted, rather than *found* sitting in a book. This single fact is the hinge of the entire chapter: if liquidity must be summoned on demand rather than read off a book, the market's structure reorganises around a completely different primitive, the **dealer** and the **quote request**, rather than the order and the match.

> **Key idea:** Equities concentrate interest into few instruments with continuous resting liquidity, which is exactly what a central limit order book is built to match. Bonds fragment interest across

a huge number of rarely traded instruments with no continuous resting liquidity, which is exactly what a central limit order book cannot serve. The bond market therefore did not adopt the CLOB; it built a dealer-intermediated, quote-driven structure instead.

9.2 The Dealer-Intermediated, Quote-Driven Market

In place of a matching engine crossing anonymous orders, the classic bond market is organised around **dealers** (also called **market makers** in this context, typically the fixed-income desks of large banks) who stand ready to buy bonds into their own inventory and sell bonds out of it. A customer who wants to trade a bond does not post an order to a book and wait; they *ask a dealer for a price*. The dealer, drawing on their inventory, their read of the market, and their willingness to take the position, quotes a bid and an offer. The customer trades against the dealer's balance sheet, not against another customer.

This is a **quote-driven** market (as opposed to the **order-driven** markets of the rest of this book), and its defining workflow is the **Request for Quote (RFQ)**:

1. A customer decides they want to buy (or sell) a particular bond in a particular size.
2. They send an RFQ, historically by phone, now overwhelmingly electronically, to one or several dealers, specifying the bond and the size.
3. Each solicited dealer responds with a firm quote (or declines).
4. The customer compares the responses and trades against the best one, or none.

Contrast every step with the CLOB. There is no anonymous continuous book; there is a bilateral solicitation. There is no price-time priority; the customer chooses among dealer quotes on price (and relationship, and certainty). The "liquidity" is not resting in a data structure; it is manufactured on request from a dealer's balance sheet and appetite. And critically, the dealer takes **principal risk**, when they buy a bond from a customer, they own it, and bear the risk of its price moving until they can sell it on. The dealer's compensation for that risk and that service is the **bid-offer spread**, which in illiquid bonds can be far wider than the penny-wide spreads of a liquid stock.

Electronification, but not "CLOB-ification." It is a common misconception that bond markets are simply "less electronic" than equities and slowly catching up to a CLOB. That is wrong. Bond markets have electronified enormously, platforms such as **MarketAxess**, **Tradeweb**, and **Bloomberg** now intermediate a large and growing share of corporate-bond and government-bond trading, but what they electronified is the **RFQ workflow**, not the order book. An electronic RFQ platform lets a customer fire a request to many dealers at once and receive competing quotes back in seconds, which is a massive efficiency gain over sequential phone calls, but the underlying market structure is still request-and-quote, still dealer-intermediated, still principal-based. The technology accelerated the dealer model; it did not replace it with a matching engine.

All-to-all trading: a partial move toward the book. The one genuine structural evolution is **all-to-all** trading, in which platforms allow *any* participant, not only designated dealers, to respond to an RFQ or provide liquidity. This lets an asset manager who happens to want the exact bond another asset manager is selling trade directly, with the platform as intermediary, rather than each paying a dealer spread. All-to-all blurs the sharp dealer/customer distinction and imports a little of the anonymous-matching flavour of equities, and it has grown significantly, especially in corporate credit. But it operates *alongside* the dealer RFQ model rather than replacing it, and the deepest, largest, or most illiquid trades still route to dealers who can commit balance sheet.

9.3 The Two Very Different Halves: Governments vs Corporates

"Fixed income" spans a spectrum from the most liquid security on earth to instruments that trade twice a year, and the market structure differs sharply along it.

Government bonds (U.S. Treasuries), the liquid end. Treasuries are the exception that partly proves the rule. Because they are homogeneous, enormous in size, and central to the whole financial system, the most recently issued, **on-the-run**, Treasuries are liquid enough to support something much closer to the order-driven markets of this book. The **interdealer** market for on-the-run Treasuries does run on **central limit order books** with price-time priority, operated by platforms such as **BrokerTec** and **Dealerweb**, and increasingly populated by the same principal-trading and high-frequency firms that trade equity futures. This

is why the *Latency and Co-location* chapter's world reaches into Treasuries at all: the on-the-run interdealer market is a genuine low-latency CLOB. But step one issue back to the **off-the-run** bonds (everything not most-recently-issued), and liquidity thins dramatically and the market reverts to the dealer/RFQ structure. So even within Treasuries, the CLOB model holds only for a thin, liquid sliver at the front; the vast bulk trades like the rest of fixed income.

Corporate bonds, the illiquid end. Corporate bonds sit firmly in the dealer-intermediated, RFQ world, for all the reasons in the previous section, magnified: enormous instrument proliferation, mostly buy-and-hold holders, and long gaps between trades for any given CUSIP. There is essentially no continuous central limit order book for a typical corporate bond, and building one would be pointless, there is not enough continuous interest to populate it. Electronic corporate-bond trading is RFQ and, increasingly, all-to-all, layered on top of the dealer model.

Municipals, mortgage-backed, and the rest occupy various points on the same spectrum, but the organising principle is identical: the more homogeneous and heavily traded the instrument, the closer it gets to an order book; the more fragmented and buy-and-hold, the more purely it lives in the dealer/RFQ world.

9.4 Post-Trade Transparency: TRACE and the End of the Opaque Quote

An order-driven market is *pre-trade* transparent almost by definition: the order book *is* a public display of resting supply and demand, and (in equities) consolidated tapes publish trades in near real time. A quote-driven bond market has no book to display, and for most of its history it had strikingly little transparency of any kind. A customer asking a dealer for a price had no reliable, independent way to know where that bond had recently traded, which left the dealer's spread almost entirely to the dealer's discretion and the customer's negotiating skill.

The decisive reform in the U.S. was **TRACE** (the **Trade Reporting and Compliance Engine**), introduced by the NASD (now FINRA) in **2002**, which for the first time required dealers to *report* their corporate-bond transactions and made much of that post-trade information public. TRACE did not create a central limit order book, and it did not make the market order-driven; it created **post-trade transparency**, so that a customer, or a regulator, could see the prices at which bonds had actually traded. The documented effect was a measurable narrowing of bid-offer spreads and a reduction in customers' transaction costs, particularly for smaller trades, precisely because information asymmetry between dealer and customer shrank. TRACE is the fixed-income analogue of the consolidated tape: not a matching mechanism, but a transparency mechanism, and it is the reason a modern bond RFQ can be checked against recent trade prints at all.

For the developer, TRACE and its equivalents are a reminder that "transparency" in a quote-driven market is a *reporting* problem, not a book-display problem: the system's obligation is to capture, disseminate, and archive completed trades accurately and promptly, not to publish a live book that does not exist.

9.5 Pricing Conventions: Yield, Price, and the 32nds

Bonds also differ from equities at the granular level of how a price is even *expressed*, and a system handling bonds must model these conventions natively, exactly the kind of reference-data detail the *Reference Data* chapter warned is dangerous to get wrong.

Price and yield are two views of the same trade. A bond can be quoted as a **price** (a percentage of face value, so a bond at "98.50" costs $985 per $1,000 of face) or as a **yield** (the annualised return an investor earns holding it to maturity). Price and yield move inversely and are linked by the bond's coupon and maturity through a standard present-value calculation. Different instruments are conventionally quoted different ways, some markets quote price, some quote yield, some quote a **spread** to a benchmark bond, and a system must convert correctly among them.

Clean price vs dirty price. Because bonds pay periodic coupons, a buyer partway between coupon dates owes the seller the interest **accrued** since the last coupon. The **clean price** is the quoted price excluding accrued interest; the **dirty price** (or **invoice price**) is what the buyer actually pays, clean price plus accrued

interest. Quoting is conventionally in clean price; settlement is in dirty price. A system that conflates the two will misprice every trade between coupon dates, i.e., almost all of them.

Fractional quoting in 32nds. U.S. Treasuries are, by long convention, quoted in **32nds of a point** rather than in decimals, a price shown as "99-16" means 99 and 16/32, i.e., 99.50% of face, and finer gradations are quoted in halves or quarters of a 32nd, and on some platforms in eighths of a 32nd (that is, in 256ths of a point). This is a direct descendant of the fractional-quoting history the *Tick Sizes and Fractional Ticks* chapter described for equities before decimalisation, except that Treasuries never decimalised, and a system parsing or displaying Treasury prices must handle the 32nds convention explicitly rather than assuming decimal prices.

9.6 Settlement and the Move to Central Clearing

Settlement timing. U.S. Treasuries have long settled on a next-business-day basis (**T+1**). U.S. corporate and municipal bonds settled on a **T+2** basis until the market-wide U.S. move to **T+1** in **May 2024** (the same shortening that applied to equities), so post-2024 most U.S. cash bond settlement is next-day. As always with settlement cycles, the exact convention is jurisdiction- and instrument-specific and periodically shortened, verify against current rules for the market you are building for.

Central clearing, arriving late to Treasuries. A striking structural fact is that, unlike equities and listed derivatives, a large portion of the U.S. Treasury market has historically **not** been centrally cleared: many trades settled bilaterally between counterparties, without a central counterparty novating and guaranteeing them in the way the *Clearing and Settlement* chapter described as standard elsewhere. The clearing house that does exist for Treasuries, the **Fixed Income Clearing Corporation (FICC)**, a DTCC subsidiary, historically cleared only a fraction of the market. Regulators, concerned about the systemic risk of an enormous, systemically central market resting on bilateral, uncleared exposures (an echo of the LTCM and Archegos lessons about opaque bilateral risk), moved to **mandate central clearing** of a broad set of Treasury cash and repurchase (repo) transactions. The SEC adopted rules to this effect in **December 2023**, with a **phased implementation** that has since been subject to extension; treat the specific compliance dates as live and changeable, and verify current deadlines rather than relying on any date printed here. The direction of travel is unambiguous, however: the Treasury market is being pushed toward the central-clearing model that the rest of this book treats as the norm, precisely because its size makes bilateral, uncleared counterparty risk a systemic concern.

9.7 The Primary Market: Auctions and Syndication

The *primary market* mechanics of the *Raising Capital* chapter also look different in fixed income.

Government issuance by auction. Governments issue debt through **auctions**. The U.S. Treasury sells new securities via a **single-price (uniform-price, or "Dutch") auction**: competitive bidders submit the yield they are willing to accept and the quantity they want; the Treasury fills bids from the lowest yield upward until the issue is sold; and, crucially, *all* winning bidders pay the *same* price, the highest accepted yield (the "stop-out" yield), regardless of the yield they individually bid. Non-competitive bidders (typically smaller investors) agree in advance to accept whatever that clearing yield turns out to be, guaranteeing them an allocation. The single-price format, and the rules constraining how much of an issue any one bidder may take, are not arbitrary; they are the direct product of a manipulation scandal, described in the historic note below.

Corporate issuance by syndication. Corporate bonds are typically brought to market by a **syndicate** of investment banks (analogous to IPO underwriters) who build a book of investor demand and price the new issue, rather than by open auction. The mechanics resemble the equity underwriting of the *Raising Capital* chapter more than they resemble a Treasury auction.

> **Historic Note: The Salomon Brothers Treasury Auction Scandal, 1991**
>
> The rules that govern U.S. Treasury auctions today, in particular the strict limit on how much of a single issue any one bidder may acquire, exist because of a deliberate abuse of the old rules by one firm. In 1991, **Salomon Brothers**, then a dominant force in government-bond trading, repeatedly violated the Treasury's rule limiting any single bidder to **35%** of an auction. Salomon submitted bids in the names of customers *without their authorisation*, using them to acquire far more than

35% of certain issues, and in the May 1991 two-year note auction it thereby gained control of a dominant share of the issue. Cornering the supply of a specific note let Salomon **squeeze** the market: participants who had sold the note short and needed to buy it back to deliver found that almost the entire supply was controlled by one firm, and were forced to pay artificially high prices. When the conduct came to light, the consequences were severe: Salomon paid roughly **$290 million** in fines and penalties, then among the largest ever imposed on a securities firm; its senior management resigned; and the firm was saved from potential collapse only when its largest shareholder, **Warren Buffett**, stepped in as interim chairman and pledged the firm's cooperation to regulators to preserve its ability to trade government debt. In the aftermath, the Treasury and its regulators overhauled the auction process, tightening bidder-identity and bidding rules and, over the following years, moving from the old multiple-price format toward the **single-price (uniform-price) auction** now standard, tested from 1992 and adopted for all marketable Treasury securities by **1998**, a format that reduces both the incentive and the ability to game the auction. The episode is the fixed-income counterpart to the manipulation cases elsewhere in this book: it shows that a primary-market mechanism is as much a target as a secondary-market benchmark, and that auction rules, like matching rules, are written in the scar tissue of specific abuses. [U.S. Securities and Exchange Commission and U.S. Department of the Treasury actions against Salomon Brothers, 1991–92; Joint Report on the Government Securities Market (Department of the Treasury, SEC, and Board of Governors of the Federal Reserve System), January 1992.]

9.8 What the Developer Should Carry Away

If you move from an equities or futures exchange to a fixed-income venue or platform, the mental adjustments are specific and large:

- The core primitive is the **RFQ and the dealer quote**, not the order and the match. Your central workflow is soliciting and comparing quotes, not maintaining a crossing book, except in the narrow on-the-run interdealer Treasury niche, where a genuine CLOB does apply.
- **Liquidity is summoned, not resting.** Do not model a bond market as a sparsely populated order book; model it as a network of dealers from whom quotes are requested on demand.
- **Transparency is a post-trade reporting obligation** (TRACE-style), not a live-book display, because there is no live book to display for most instruments.
- **Pricing conventions are richer and must be modelled natively**: price *and* yield, clean *and* dirty price, 32nds for Treasuries, spread quoting, day-count and accrued-interest calculations, all of which are reference-data and calculation correctness problems of exactly the kind the *Reference Data* chapter flagged as high-consequence.
- **Instrument reference data dominates.** With hundreds of bonds per issuer, each with its own coupon, maturity, seniority, and call schedule, the instrument master is far larger and more intricate than an equity venue's, and errors in it mispricing accrued interest or a call date corrupt every trade in that bond.

Key idea: Fixed income is the largest securities market in the world and almost none of it looks like the central-limit-order-book, price-time-priority, continuous-matching model that the rest of this book describes. Because bonds fragment interest across a vast number of rarely traded instruments with no continuous resting liquidity, the market is dealer-intermediated and quote-driven, organised around the request-for-quote and the dealer's balance sheet, with a genuine order book surviving only in the thin, liquid on-the-run Treasury interdealer niche. Electronification accelerated the RFQ workflow rather than replacing it with a matching engine; transparency arrived as post-trade reporting (TRACE) rather than a public book; pricing carries its own conventions (yield, clean/dirty, 32nds); and central clearing is only now being extended to the Treasury market it had long bypassed. Carry the CLOB model into bonds unmodified and you will build for a market that does not exist.

CHAPTER 10

Cryptocurrency and Digital Asset Venues

Every chapter so far has quietly assumed a particular model of "exchange": a regulated entity, operating within defined trading sessions, settling trades some fixed number of days later through a central counterparty, under the supervision of a single national regulator. Cryptocurrency and digital asset venues share a great deal of underlying mechanics with that model, order books, price-time priority, matching engines, are recognisably the same ideas, but they also break several of the traditional model's assumptions simultaneously. Because a growing number of developers move between traditional and digital-asset exchange systems over the course of a career, it is worth making those differences explicit rather than leaving them implicit.

10.1 What Carries Over Unchanged

The core matching concepts from Part II transfer directly. A crypto exchange's order book is still two sorted lists of resting orders; it still resolves priority by price and then arrival time; limit, market, and IOC orders behave exactly as described earlier in this book; and a single-threaded, deterministic matching engine (see *The Matching Engine*) is just as valuable here as anywhere else, arguably more so, since many digital asset venues have historically had thinner regulatory audit-trail requirements than traditional exchanges, making internal determinism and replay the operator's main defence against disputed trades.

10.2 What Does Not: Continuous, Global, 24/7 Trading

Traditional exchanges are built around the **trading session state machine** described in Part II: pre-open, opening auction, continuous trading, closing auction, close, GTC orders persisted overnight. Most cryptocurrency venues have no such structure at all. Trading is continuous, 24 hours a day, 365 days a year, with no defined open or close and, historically, no opening or closing auction establishing an official daily reference price (some larger venues have begun introducing scheduled auctions, but it is far from universal). There is no overnight batch window, no "GTC orders reloaded at start of day," because there is no start of day. This single difference cascades: features described earlier in this book as anchored to the official close, the static price collar's reference price, end-of-day P&L, the settlement price used for margin, all need a different anchor (often a rolling time window, such as a trailing 24-hour or 8-hour mark price) when there is no closing auction to provide one.

10.3 What Does Not: Settlement

The *Clearing and Settlement* section described a model with three distinct steps, matching, clearing through a novating CCP, and settlement, deliberately separated in time (T+1 in current US equities) to allow risk management to happen in between. Digital asset venues diverge from this in two quite different ways depending on their design:

Centralised exchanges (CEXs), such as the major custodial trading platforms most retail and institutional crypto activity flows through, typically operate their matching engine much like a traditional exchange, but settlement between the trade and the underlying blockchain is decoupled: a trade executed on the exchange

is recorded as an internal ledger entry between the exchange's own custodial accounts, not as a blockchain transaction in real time. Moving the actual asset onto or off the blockchain (a deposit or withdrawal) is a separate, much slower operation, subject to blockchain confirmation times rather than the exchange's own matching latency. Critically, there is typically no independent, regulated CCP standing between counterparties the way DTCC or a CME-style clearing house does in traditional markets (see *Clearing and Settlement*); the exchange itself is commonly both the matching venue and the de facto custodian and counterparty for its users' balances.

Decentralised exchanges (DEXs) dispense with a matching engine and an order book altogether for many products, replacing it with an **automated market maker (AMM)**: a smart contract holding a pool of two assets and pricing trades algorithmically from the pool's current ratio, most commonly via a *constant product* formula, `reserve_A × reserve_B = k`, where `k` is held constant. A trade that adds some quantity of asset A to the pool must remove enough of asset B to keep the product constant, and the price a trader receives is determined entirely by the size of their trade relative to the pool's depth, not by matching against another participant's resting order at all. There is no bid, no ask, and no price-time priority in the traditional sense; "depth," in the sense introduced in the *Order Book* section of Part II, is replaced by pool size, and a trade *is* its own settlement, executed atomically as a single blockchain transaction with no separate clearing step, because there is no intermediary institution left to clear through.

10.4 What Does Not: A Single Regulator

Part I described how equity markets operate under a specific regulator (the SEC in the US, the FCA and ESMA-governed regimes in Europe) whose rules are, in practice, an engineering specification. Digital asset markets have no equivalent single authority: oversight is split across securities regulators, commodities regulators, money-transmission licensing regimes, and, for many offshore venues, jurisdictions with comparatively little oversight at all. A digital asset venture operating across multiple countries may face materially different, sometimes contradictory, regulatory obligations depending on where its users and its corporate entities are located, a sharp contrast with the relatively unified regulatory picture (SEC/Reg NMS domestically, MiFID II across the EU) that the rest of this book has been able to describe with reasonable confidence.

10.5 Precision: Why Tick Size Gets Harder, Not Easier

The *Tick Sizes and Fractional Ticks* section explained why prices are stored as integer tick counts rather than floating-point decimals. Digital assets make the same lesson more demanding rather than less: Bitcoin is commonly denominated down to one hundred-millionth of a unit (the "satoshi"), and many Ethereum-based tokens use eighteen decimal places, precision far beyond what a standard 32- or 64-bit integer can represent exactly at realistic price and quantity ranges. Exchange systems handling these assets typically require wider fixed-point or arbitrary-precision integer arithmetic throughout the matching and ledger path, the same "never use floating point for money" principle from Part II, just with a much larger number of ticks between whole units.

> **Historic Note: FTX, November 2022**
>
> The clearest illustration of what goes wrong when a venue collapses the separation between matching, custody, and independent clearing is the collapse of **FTX**, at the time one of the world's largest cryptocurrency exchanges, in November 2022. FTX was simultaneously the operator of a customer-facing trading platform and, through an affiliated trading firm, a major participant trading against its own exchange's customers, with customer deposits allegedly commingled with and used to fund that affiliated firm's activities. When a liquidity crunch triggered a wave of customer withdrawal requests, FTX was unable to meet them, filed for bankruptcy within days, and its founder was later convicted of fraud. There was no independent CCP standing between FTX's customers and the risk they were actually taking, no novation of the kind DTCC or CME Clearing provide in the traditional markets described in the *Clearing and Settlement* section, so when the exchange itself failed, customer assets were directly exposed, precisely the concentration of counterparty risk that the introduction of central clearing was designed to eliminate in traditional markets after events like the 2008 Lehman collapse discussed earlier in this book. For exchange system developers, the FTX case is the digital-asset-era companion to Barings and Knight Capital:

a reminder that the institutional separation of duties this book describes throughout, matching engine, independent risk layer, independent custodian, independent clearing house, is not bureaucratic overhead, it is the specific structure that prevents any single point of failure from taking the whole system down with it. [US Department of Justice, *United States v. Samuel Bankman-Fried*, 2023 conviction; US Securities and Exchange Commission and Commodity Futures Trading Commission civil complaints, December 2022.]

Key idea: Digital asset venues prove that the order book, price-time priority, and deterministic matching concepts in this book are more general than the regulated-exchange context they are usually taught in, but they also show, sometimes at great cost, what is lost when the surrounding institutional structure, session boundaries, independent clearing, unified regulation, is missing or deliberately collapsed into a single entity.

CHAPTER 11

Latency and Co-location, The Speed Dimension

Speed is not incidental to electronic markets, it is a primary competitive dimension. Understanding why latency matters, and what mechanisms exchange systems use to minimise it, is essential background for exchange developers.

11.1 Why Latency Matters

At any moment, the same stock is trading on multiple exchanges simultaneously. If the price of AAPL changes on NYSE, that information takes time to propagate to NASDAQ. During that window, even if only 50 microseconds, participants who know the price has changed on NYSE can trade on NASDAQ before NASDAQ's quotes adjust. This is called **latency arbitrage**.

Market makers must continuously update their quotes faster than latency arbitrageurs can act. A market maker whose quotes are 10 microseconds stale will find themselves adversely selected on the stale side. The entire market structure is shaped by this dynamic: venue design, technology choices, and the physical geography of data centres all aim to reduce or equalise latency.

Reuter's Pigeons: The First Latency Arbitrage Infrastructure (1850)

In 1850 the European telegraph network had a hole in it: the Prussian line ended at Aachen and the Belgian–French line began at Brussels, roughly 120 kilometres apart. Market-moving information, above all, closing stock prices, crossed the gap by train, taking many hours. Paul Julius Reuter's solution was a fleet of carrier pigeons flying the Aachen–Brussels route in around two hours, giving his subscribers Brussels prices from Berlin (and vice versa) many hours before anyone relying on the railway. The service lasted about a year before the telegraph gap was closed, at which point Reuter moved to London and built the news agency that still bears his name. The structure of the trade is precisely the one the rest of this chapter describes: identify the slowest link in the information path between two markets, replace it with a faster physical medium, and monetise the interval before the infrastructure catches up. Pigeons, dedicated dark fibre, microwave towers, and millimetre-wave links are one continuous story separated only by technology.

References: Donald Read, *The Power of News: The History of Reuters* (Oxford University Press, 1992), ISBN: 978-0198207689

11.2 Co-location

Co-location is the practice of placing a participant's trading servers physically inside the same data centre as the exchange's matching engine. Even tens of kilometres of metro fibre typically add hundreds of microseconds of round-trip latency in real paths. A co-located server in the same facility can reduce this to low single-digit microseconds end-to-end on optimised networks.

Exchanges provide co-location as a service: they sell rack space in their data centres to participants who want the lowest possible latency. NYSE's co-location facility in Mahwah, New Jersey, and NASDAQ's in Carteret, New Jersey, host servers for hundreds of trading firms. Eurex runs co-location at its Frankfurt data centre.

11.3 Microwave and Laser Links

The speed of light in fibre-optic cable is approximately 70% of the speed of light in a vacuum, light travels somewhat slower through glass than through air. For data links between distant locations (Chicago to New York, London to Frankfurt), microwave and millimetre-wave radio links are faster than fibre because radio waves travel at nearly the speed of light through air.

The economics of this advantage are extreme. **Spread Networks** spent a reported $300 million or so to build a new, nearly straight fibre route between Chicago and New York, opened in 2010, cutting round-trip latency from roughly 14.5–16 milliseconds on existing routes to about 13 milliseconds. The project was a breakthrough, and within a few years it had been overtaken by microwave networks with round trips of around 8.5 milliseconds, faster than even a perfectly straight fibre could ever be, because signals in air travel about 1.5 times faster than light in glass.

Multiple firms have since built dedicated microwave networks between major financial data centres. On the Chicago-to-New York route (roughly 1,150 km as the crow flies), a microwave signal takes about 4 milliseconds *one way*, against well over 6 milliseconds through the best fibre, a decisive advantage in latency-sensitive trading. (A useful rule of thumb for developers: light in fibre covers about 200 km per millisecond, or 5 microseconds per kilometre.) Experimental laser links (which travel through air but require line-of-sight and clear weather) can be even faster.

11.4 Hardware Acceleration

For the most latency-sensitive operations, software running on general-purpose CPUs is too slow. Exchange systems and trading firms use:

FPGAs (Field-Programmable Gate Arrays): Reconfigurable hardware chips that can implement logic in dedicated circuits rather than software. An FPGA can parse a market data message and generate an order response in tens of nanoseconds, far faster than any software path.

Kernel bypass: Standard network communication on Linux involves the operating system kernel, data travels from the network card through kernel buffers to user-space. This adds microseconds of latency. Kernel bypass technologies (DPDK, OpenOnload) allow applications to communicate directly with the network hardware, eliminating kernel overhead.

Specialised trading network adapters (SmartNICs): General-purpose Ethernet cards are designed for throughput, not latency. A dedicated class of network interface cards, sometimes called **SmartNICs** or **ultra-low-latency NICs**, is built specifically for trading environments. The most widely deployed in financial markets are the **AMD Solarflare** cards (originally from Solarflare Communications, acquired by Xilinx, then by AMD), along with **NVIDIA Mellanox ConnectX** adapters.

What distinguishes these cards from commodity NICs:

- **Onload / kernel bypass stack built into the card driver.** Solarflare's *OpenOnload* and *TCPDirect* technologies implement a full user-space TCP/IP stack that runs inside the application process, entirely bypassing the Linux kernel for every send and receive. A packet sent with OpenOnload never touches the kernel, it goes directly from application memory to the NIC hardware. Round-trip latency for a network message drops from the 20–50 microseconds typical of a standard Linux kernel path to roughly 1–5 microseconds.
- **Hardware timestamping with nanosecond precision.** When a packet arrives at the NIC, the card itself records the arrival time in hardware, before any software has seen the packet. This timestamp is far more accurate than a software timestamp applied by the operating system (which may be delayed by scheduling jitter, interrupt coalescing, or kernel overhead). Hardware timestamps are important for two reasons: regulatory compliance (audit trails require accurate order-arrival times), and latency measurement (calculating the precise round-trip time of a market data message requires knowing exactly when each packet arrived at the wire, not when software noticed it).
- **Kernel-free packet capture.** Solarflare's *SolarCapture* allows applications to capture every packet at line rate, 10 Gbps, 25 Gbps, or 100 Gbps, without packet loss and with hardware timestamps on each frame. Trading firms use this for post-trade analysis and regulatory audit trails without impacting the

latency of the live trading path.

- **CPU offload.** Certain networking operations, checksum calculation, TCP segmentation, flow steering, are offloaded to the NIC hardware entirely, freeing the CPU to run trading logic rather than network stack operations.

A typical co-location deployment for a market making firm might use Solarflare or Mellanox adapters for all connections to the exchange gateway, running OpenOnload or ConnectX's RDMA (Remote Direct Memory Access) stack in the application. The same physical server that sends thousands of orders per second also hardware-timestamps every incoming market data packet, providing a complete, nanosecond-accurate audit trail of every event the system observed and every action it took.

For developers building exchange infrastructure rather than trading clients, the significance is slightly different: the exchange's own gateway servers and matching engine servers commonly use the same class of hardware to minimise the latency of accepting, processing, and acknowledging orders, and to generate the authoritative hardware timestamps that appear in the regulatory audit trail.

Deterministic latency: In high-frequency systems, not just average latency but **worst-case latency** matters. A response that is usually 5 microseconds but occasionally spikes to 500 microseconds due to garbage collection or OS scheduling jitter is unacceptable. Systems are designed with deterministic, bounded latency using techniques like memory pre-allocation, real-time OS scheduling, and CPU pinning.

11.5 PTP Clock Synchronisation

When events across multiple machines must be compared, all machines must share a common, accurate time source. **PTP (IEEE 1588 Precision Time Protocol)** synchronises clocks across a network to sub-microsecond accuracy, far more precise than NTP (which typically achieves milliseconds over the internet and, at best, tens to hundreds of microseconds on a well-engineered local network). Regulators now require this kind of precision: under MiFID II, trading venues and firms engaged in high-frequency trading must keep their business clocks within 100 microseconds of UTC, with timestamps recorded to at least microsecond granularity. Exchange systems and co-location facilities use PTP or dedicated hardware timing signals (GPS-disciplined oscillators) to ensure that timestamps on events from different machines are directly comparable.

> **Key idea:** In exchange infrastructure, latency is not just a performance metric, it determines who gets to trade at what price. Every architectural decision (data structures, networking, hardware) has latency implications. Understanding this context helps you make better design choices even in components not directly in the critical path.

CHAPTER 12

Operational Observability and Incident Response

Production readiness depends not only on matching correctness, but on fast, reliable detection and recovery when something degrades. An exchange that matches perfectly but cannot tell when it is sick, or cannot recover quickly when it is, is not production-grade. Observability and incident response are first-class engineering features, and in many jurisdictions they are also regulatory obligations.

12.1 The Four Golden Signals

A useful framework for deciding *what* to measure is the set of **four golden signals** popularised by site reliability engineering. Each maps directly onto an exchange:

- **Latency:** how long operations take, for example order ACK and fill time. Measure successful and failed requests separately, a fast rejection is not the same as a fast fill.
- **Traffic:** demand on the system, for example orders per second, cancels per second, and market-data messages published.
- **Errors:** the rate of failed operations, for example rejects by reason code.
- **Saturation:** how full the system is, for example ingress queue depth, CPU and cache pressure, and WAL write latency. Saturation is the leading indicator: it rises *before* latency and errors degrade.

The value of the framework is that it turns an open-ended "what should we monitor?" into four concrete questions that every component, gateway, matcher, market-data publisher, must answer.

12.2 Core Production Signals

Building on the golden signals, exchanges track and alert on:

- End-to-end order ACK/fill latency (p50/p95/p99/p99.9)
- Ingress queue depth and backlog age per gateway and per symbol partition
- Market-data publish latency and subscriber lag
- Sequence-gap and retransmission rates on each market-data channel
- Replay buffer utilisation and replay hit/miss ratios
- Reject rates by code (risk, syntax, session, throttling)
- Order-to-trade and cancel-to-fill ratios per participant
- Business-level events: halts, MMP and LULD firings, extreme auction imbalances
- Failover readiness (replication lag, last durable offset, heartbeat health)

These metrics should be broken out by venue session state (auction, continuous, halt), because acceptable baselines differ by phase. A p99 latency that is healthy during continuous trading may be meaningless during an auction uncross.

12.3 Clock and Time-Synchronisation Health

An exchange is a time machine: every order is sequenced, every event is timestamped, and the audit trail is only as trustworthy as the clocks that produced it. **Clock health** is therefore a first-class signal unique to exchange infrastructure.

Operators monitor the **offset** between each host's clock and a reference (via PTP or NTP), the drift rate, and the synchronisation source's own health. A gateway whose clock has drifted from the matching engine can produce timestamps that appear out of order, corrupting sequencing logic and regulatory reporting. Because clock degradation is silent, it must be alerted on directly rather than inferred after the fact.

12.4 The Three Pillars: Metrics, Logs, Traces

Mature observability rests on three complementary data types:

- **Metrics** are aggregated numeric time series (the signals above). They are cheap to store and ideal for dashboards and alerting, but they cannot explain *why* a single order behaved oddly.
- **Logs** are structured, timestamped records of discrete events. In an exchange they overlap heavily with the **audit trail**, every accept, reject, fill, and cancellation, and must be structured (machine-parseable fields, not free text) to be searchable under pressure.
- **Traces** follow a single request across components. By attaching a **correlation ID** (typically the order ID) to every message, an operator can reconstruct an order's full path, gateway → risk checks → matcher → market-data publish, and locate exactly where latency or an error was introduced.

Metrics tell you *that* something is wrong, traces tell you *where*, and logs tell you *why*.

12.5 Service-Level Objectives and Error Budgets

To distinguish "degraded" from "normal", a venue defines **Service-Level Indicators (SLIs)**, the precise metrics it cares about, and **Service-Level Objectives (SLOs)**, the target values for those metrics, for example "99.9% of order ACKs within 200 microseconds during continuous trading."

The gap between an SLO and 100% is the **error budget**: the amount of degradation the venue is willing to tolerate over a period. Error budgets convert reliability into an explicit, shared currency, when the budget is healthy, teams can ship changes faster; when it is exhausted, the priority shifts to stability. Because acceptable performance differs by session phase, SLOs are usually defined per phase rather than as a single daily number.

12.6 Alerting Design

More monitoring is not automatically better. Poorly designed alerting produces **alert fatigue**, where operators learn to ignore a noisy channel and miss the one alert that mattered.

Good practice favours **symptom-based alerts** (page when participants are affected, for example fills are slow) over **cause-based alerts** (page on every elevated CPU reading), and separates **paging** alerts (wake a human now) from **ticketing** alerts (review during business hours). Critically, the monitoring system must itself be monitored: a **dead-man's-switch** or **heartbeat** alert fires when expected signals *stop arriving*, catching the dangerous case where the exchange looks healthy only because telemetry has silently failed.

12.7 Runbooks and Recovery Drills

A professional venue maintains tested runbooks for:

- Gateway degradation and targeted traffic shedding
- Symbol-partition isolation (quarantine one partition without halting all trading)
- Primary-to-secondary failover with deterministic replay verification
- Market-data incident handling (gap storms, replay saturation, stale snapshots)

Runbooks should include operator decision thresholds, not just command sequences. Drills should be rehearsed under load and during controlled session simulations, not only in staging idle conditions. **Determinism and**

replay (covered later in this part) are themselves incident-response tools: the ability to replay an exact input sequence lets operators reproduce a failure offline while live trading continues, and the **kill switch** and **mass cancel** controls (from the risk and compliance part) are the operator's primary levers for containing a runaway participant.

12.8 Incident Command and Postmortems

When an incident exceeds routine handling, an explicit structure prevents chaos. Incidents are assigned a **severity level** (for example SEV1 for a trading-impacting outage down to SEV3 for a contained, non-customer-facing fault), which drives escalation and communication expectations.

A defined **incident command** model assigns clear roles, an **Incident Commander** who coordinates and decides, and a separate **communications lead** who keeps participants and stakeholders informed, so that responders are not simultaneously fixing and explaining. Afterwards, a **blameless postmortem** records the timeline, measures **time-to-detect (MTTD)** and **time-to-recover (MTTR)**, and produces concrete corrective actions that feed back into runbooks, alerts, and tests. The goal is organisational learning, not individual blame; engineers who fear punishment hide the very information that prevents recurrence.

12.9 Participant Communication and Regulatory Obligations

An exchange outage is not a private engineering matter. Participants need timely, accurate notice through **status pages** and **market notices** so they can manage their own risk, and silence during an incident erodes trust faster than the outage itself.

In many jurisdictions, incident handling is mandated. In the US, **Regulation SCI (Systems Compliance and Integrity)** requires covered market infrastructure to maintain capacity and resilience, to report significant systems issues to the regulator, and to conduct business-continuity testing. Observability and incident response are therefore compliance functions as much as engineering ones: the venue must be able to *demonstrate*, with evidence, that it detected, communicated, and resolved an incident appropriately.

> **Key idea:** Low latency is valuable, but operational predictability is what keeps markets open during stress. A venue is publishable and production-grade only when observability and incident response, signals, tracing, SLOs, alerting, runbooks, incident command, and regulatory communication, are engineered as first-class features rather than added after the first outage.

CHAPTER 13

Corporate Actions, When the Instrument Changes

An exchange does not just serve static instruments. Companies whose shares trade on an exchange undergo **corporate actions**, events that change the structure of the instrument itself. These events have significant operational consequences for every component of an exchange system.

13.1 Stock Splits

A **stock split** divides each existing share into multiple new shares, reducing the price proportionally. If AAPL trades at $200 and does a 4-for-1 split, each shareholder receives 4 shares for every 1 they held; the price adjusts to approximately $50. The total market capitalisation is unchanged.

Exchange implications: - All open limit orders must be adjusted: an order to buy 100 shares at $200 becomes an order to buy 400 shares at $50. - Historical price data must be adjusted (or marked as pre-split) to avoid false apparent price changes. - Tick size may change (a lower price may use a different tick increment). - Reference data (symbol metadata) must be updated.

A **reverse split** works in the opposite direction: multiple shares are consolidated into one, and the price rises proportionally. A company trading at $0.50 might do a 10-for-1 reverse split to bring the price to $5, typically to meet exchange listing requirements.

13.2 Dividends

When a company declares a dividend, shares trade **cum-dividend** (with dividend entitlement) up to the **ex-dividend date**, after which they trade **ex-dividend** (without dividend entitlement). The price typically drops by approximately the dividend amount on the ex-dividend date.

Open orders spanning the ex-dividend date may require handling, some exchanges cancel all open orders; others adjust prices.

13.3 Mergers and Acquisitions

When a company is acquired, its shares may be converted to shares in the acquirer, cash, or a combination. The target company's shares are eventually **delisted**, removed from trading on the exchange. All open orders in the target must be cancelled. Open positions must be settled.

13.4 Symbol Changes and Delistings

Companies change their ticker symbols (rebranding, mergers). Systems that track positions and orders by symbol must handle symbol remapping without losing continuity. **Delistings** require orderly unwinding of all open orders and positions in the symbol.

13.5 Why Corporate Actions Matter to Developers

Corporate actions are among the most operationally complex events an exchange system handles. They require coordination across: the order book engine (cancel or adjust open orders), the clearing system (adjust positions and cost bases), the market data system (update reference data), the audit trail (record the adjustment events), and downstream applications that may have cached the old instrument parameters.

A developer who underestimates corporate action complexity will eventually face a bug report like: "after the split, the order that was resting at $200 for 100 shares is now resting at $200 for 100 shares instead of $50 for 400 shares." The split was never propagated to the open order book.

> **Key idea:** Instruments are not static. Every system that touches order, position, or price data must be prepared to handle corporate action adjustments. Reference data management is a discipline in its own right in production exchange systems.

CHAPTER 14

Options Mechanics: Exercise, Assignment, and Expiry

Options have appeared throughout this book, the OCC as an options CCP in *Clearing and Settlement*, Cboe as the largest US options exchange group, straddles and strangles as combo-order examples in Part II, the VIX as an index derived from option prices, but always as a supporting reference rather than a subject in its own right. This section closes that gap by covering what actually happens to an option contract at the end of its life, which is a distinct set of mechanics from anything else already covered.

14.1 A Brief Recap

An **option** is the right, but not the obligation, to buy (a **call**) or sell (a **put**) an underlying instrument at a fixed **strike price** on or before a specified **expiry date**. The option's seller (the "writer") takes on the corresponding *obligation* if the buyer chooses to exercise. **American-style** options can be exercised any time up to expiry; **European-style** options (including most cash-settled index options) can only be exercised at expiry itself. This distinction matters directly for what follows: only American-style contracts create exercise, and therefore assignment, risk before expiry day arrives.

14.2 Exercise: The Holder's Decision

Exercise is the option holder invoking their right, instructing their broker to convert the option into the underlying transaction it represents: buying the underlying at the strike (for a call) or selling it at the strike (for a put). A rational holder exercises only when the option is **in-the-money**, that is, when doing so is more favourable than the current market price (a call struck at $150 is in-the-money whenever the underlying trades above $150).

Automatic exercise. Rather than requiring every holder to submit an explicit exercise instruction, the OCC automatically exercises any expiring equity option that is in-the-money by as little as $0.01 (measured against the OCC's closing price for the underlying), unless the holder has filed a specific **contrary instruction** telling it not to. This threshold, deliberately set at a single cent, exists precisely because leaving it to manual action would mean some holders forget to exercise a position that is unambiguously profitable, a similar spirit to the pre-trade safety-net logic seen elsewhere in this book, just applied to an investor's own inaction rather than to a runaway order.

14.3 Assignment: The Writer's Obligation

When a holder exercises (or is auto-exercised), someone on the other side of that contract must fulfil the obligation. Because options are cleared through the OCC exactly as described in *Clearing and Settlement*, no option writer is directly paired with a specific holder, the OCC's novation means every writer's counterparty is the OCC itself, and every holder's counterparty is also the OCC. This means the OCC must select, from among all the writers of that identical option series, which one (or which several, if the exercised quantity is

smaller than any single writer's position) is actually obligated to deliver.

This selection process is **assignment**. The OCC assigns the exercise notice to a specific clearing member (a broker), and that broker then allocates the assignment among its own clients who hold short positions in that option series, using an allocation method disclosed in advance, commonly either **random allocation** or **FIFO** (the earliest-opened short position is assigned first), echoing the same FIFO fairness principle already familiar from *Price-Time Priority*, just applied to which existing position gets the obligation rather than to which new order gets a fill. A writer of an option genuinely does not know, until assignment notices are processed, whether they specifically will be the one required to perform, only that *someone* holding the aggregate short position will be.

14.4 Cash Settlement vs. Physical Delivery

As reference data already noted in Part IV (*Reference Data, The Exchange's Ground Truth*), every option contract specifies its **settlement method**:

Physical delivery is standard for individual equity options: exercise and assignment result in an actual transfer of shares, one standard US equity option contract corresponds to 100 shares, at the strike price. A writer who is assigned on a call must deliver 100 shares per contract; if they do not already own them (a **covered call**), they must acquire them in the market, potentially at a materially worse price than the strike that triggered the assignment.

Cash settlement is standard for most broad index options (SPX on the S&P 500, and by extension products like the VIX options already introduced in Part II), since physically delivering "the S&P 500" is not a meaningful operation. Instead, the difference between the strike and the index's official settlement value (itself a carefully defined, deterministically computed print, echoing the closing-price integrity discussion in the *Indexes* section of Part II) is simply paid in cash from the assigned writer to the exercising holder, via the OCC.

14.5 A Worked Example: Pin Risk

Pin risk is the practical hazard that follows directly from automatic exercise operating on a one-cent threshold. Suppose a trader has written (sold) one call option, strike $150.00, physically settled, 100-share multiplier, and the underlying closes on expiry Friday at exactly **$150.03**.

Because $150.03 is more than one cent above the $150.00 strike, the option is in-the-money and is automatically exercised, assuming the holder filed no contrary instruction. The writer, if assigned, must deliver 100 shares at $150.00 regardless of where the stock opens the following Monday. If the writer did not already hold the shares (an uncovered position) and the stock gaps up over the weekend to $155.00 on unrelated news, the writer must buy 100 shares at $155.00 to deliver them at the $150.00 strike, a loss of $500 on a position that, at Friday's close, looked like it was only three cents in-the-money.

The genuinely uncomfortable part of pin risk is that the writer does not know, at the moment trading closes on expiry Friday, whether they will actually be assigned, assignment depends on the final settlement print (which can differ slightly from the last continuous-trading price if the closing auction, see Part II, produces a different value) and on which specific writers the OCC's allocation process selects. A position that appears to close just barely out-of-the-money, and therefore safe, can still be exercised if the official settlement value lands a cent on the wrong side of the strike, and a position that looks assigned can turn out not to be. This uncertainty is sharpest for the underlying stocks and index products involved in **triple witching**, the quarterly expiry of stock options, index options, and index futures simultaneously, already flagged in the *Indexes* section of Part II as a source of unusual volume and volatility, for exactly this reason: a large number of positions across multiple related instruments are all resolving their exercise and assignment status at the same closing print.

> **Key idea:** Exercise and assignment are a distinct lifecycle stage that sits after an option trade is matched and cleared but before it is truly finished, closer in spirit to the corporate-action adjustments described elsewhere in this Part than to an ordinary trade. Automatic exercise at a one-cent threshold, OCC-mediated assignment to writers who do not know in advance whether they specifically will be selected, and the choice between cash and physical settlement all have to be correctly modelled by any system that carries option positions through to expiry, and pin risk

is the direct, sometimes costly, consequence of getting the timing or the settlement-price precision wrong.

CHAPTER 15

Determinism, Replay, and Persistence, The Exchange Must Not Forget

These three properties, determinism, replay, and persistence, are the engineering foundations that allow an exchange to be trusted over long periods. They are closely related and each depends on the others.

15.1 Deterministic Replay

A matching engine is **deterministic** if, given the same initial state and the same ordered sequence of input messages, it always produces exactly the same outputs: the same fills, the same cancellations, the same book state, and the same sequence of events.

Determinism matters because:

Auditing. Regulators must be able to reconstruct what happened in the market at any past moment. If the exchange keeps a complete ordered log of every input message, a regulator (or an internal audit team) can replay that log and recreate the exact state of the book at any point. Non-determinism would make replay unreliable, replaying the log might produce different results from the original run.

Debugging. When a bug is discovered in exchange software, the developer needs to reproduce the exact sequence of events that triggered it. A deterministic engine with a complete message log can be replayed in a test environment to reproduce the bug precisely.

Disaster recovery. If the primary matching engine crashes, the secondary can recover by replaying the input log from the last known good snapshot. The state it reconstructs will be identical to the primary's last state.

The key implication for software design: the matching engine must have no hidden sources of non-determinism. Random number generators (if used), system clock reads, and any external state must be eliminated from the matching path, or be made deterministic inputs from the log. In practice, this means the `now` timestamp used for order sequencing is passed in as an input parameter, not read from the system clock mid-execution.

> **Key idea:** The matching engine's determinism is a design requirement, not an implementation detail. Every decision that could introduce non-determinism, reading the clock, using a hash map with randomised iteration order, calling OS functions, must be carefully controlled.

15.2 Sequence Numbers

Exchange systems use **sequence numbers** pervasively. Every message on every channel carries a monotonically increasing sequence number. This enables:

- **Gap detection:** A subscriber that receives sequence 1001 after 999 knows it missed 1000.
- **Replay:** A subscriber can request retransmission of specific sequence numbers.
- **Ordered reconstruction:** In disaster recovery, replaying messages in sequence-number order guarantees correct reconstruction.

- **Practical exactly-once semantics through duplicate detection:** True exactly-once delivery is notoriously difficult in distributed systems. Production exchanges achieve it in practice by sequencing every message and detecting duplicates at the receiver: if a message with sequence number N has already been processed, a second arrival of the same N is discarded. This gives the effect of exactly-once processing without requiring distributed coordination, each receiver maintains the last processed sequence number and silently discards retransmissions.

Sequence numbers are distinct from order IDs. An order ID identifies an order throughout its lifetime; a sequence number identifies a message in a specific channel at a specific moment. The same order generates multiple messages (new, partial fill, fill, cancel), each message has its own sequence number.

15.3 Persistence and Recovery

A production exchange cannot lose state across restarts or failures. This requires deliberate persistence design.

Write-ahead logging (WAL): Before processing any input message, the engine writes it to a durable, append-only log. The log is the source of truth. The in-memory state (the order book) is a derived, transient representation. If the engine crashes, it recovers by replaying the WAL from a recent **snapshot**, a complete dump of the book state at a known point, and then replaying subsequent WAL entries.

Snapshots: Taking a full snapshot of the order book periodically reduces the amount of WAL that must be replayed on recovery. A snapshot taken every 5 minutes means recovery replays at most 5 minutes of WAL entries regardless of how long the engine has been running.

Warm restart: A restart where the engine begins from a snapshot and replays recent WAL entries, typically completing in seconds. Distinguished from a **cold start** (starting from the beginning of the day, replaying all entries) which takes longer.

GTC persistence: Good-Till-Cancelled orders survive session boundaries. They must be persisted at end-of-day and reloaded at start-of-day. This requires a separate persistence layer beyond the WAL, typically a structured file or database holding the open GTC order set with all their parameters.

Failover: When the primary engine fails, the secondary takes over. For failover to be seamless, the secondary must have been receiving and processing all input messages in parallel (shadow mode), maintaining an up-to-date copy of the book state. When the primary fails, the secondary's warm state is already current; it only needs to begin accepting new inputs and publishing outputs.

> **Key idea:** The exchange cannot lose a single order or trade. Persistence is not optional. Every production exchange system is built around the assumption that hardware will fail, software will crash, and the system must recover to exactly the state it had before the failure.

CHAPTER 16

Reference Data, The Exchange's Ground Truth

Every component of an exchange system depends on a common set of facts about the instruments it handles: what symbols exist, what their tick sizes are, what their trading hours are, and dozens of other parameters. This is called **reference data** (also called **static data** or **instrument master data**), and it is the foundational layer that every other component reads from.

Reference data is not glamorous. It does not move prices or fill orders. But a disproportionate fraction of real exchange outages, including some very expensive ones, originate from bad reference data, not from bugs in the matching engine. A matching engine with perfect logic will produce wrong results if given incorrect tick sizes, wrong price scales, or stale contract specifications.

16.1 What Reference Data Contains

For each tradeable symbol, the exchange maintains:

Identity - **Symbol / Ticker:** The exchange-assigned code (AAPL, ESH4, EURUSD). - **ISIN (International Securities Identification Number):** A 12-character global identifier for the underlying security, independent of which exchange it trades on. AAPL has one ISIN regardless of whether it trades on NYSE, NASDAQ, or a European exchange. - **CUSIP (US) / SEDOL (UK):** Alternative identification systems used in clearing and settlement. - **Full name and issuer:** Apple Inc., the US Treasury, the underlying company.

Price parameters - **Tick size:** The minimum price increment. For AAPL on NASDAQ, $0.01. For E-mini S&P 500 futures on CME, 0.25 index points. - **Tick decimals / price scale:** The number of decimal places in a price. Used to convert between human-readable display prices and the integer tick counts used internally. - **Price currency:** USD, EUR, GBP, JPY. An exchange may list instruments denominated in different currencies. - **Contract multiplier (derivatives):** For futures and options, the dollar value of one price unit. The E-mini S&P 500 has a multiplier of $50, a 1-point move in the index is worth $50 per contract.

Quantity parameters - **Minimum lot size:** The smallest tradeable quantity. In some Asian markets, equities trade in lots of 100 or 1000 shares. - **Quantity increment:** Orders must be whole multiples of this quantity.

Lifecycle and schedule - **Trading status:** ACTIVE, SUSPENDED, HALTED, DELISTED, PRE-IPO. - **Session schedule:** When pre-open, continuous, and close auction periods begin and end. Different instruments on the same exchange may have different session schedules (bonds vs equities, for example). - **Expiry date (derivatives):** Futures and options expire. The matching engine must stop accepting new orders and settle open positions on the expiry date. - **Last trading day:** For futures, the last day orders can be submitted. - **Settlement method (derivatives):** Cash settlement (the difference between the final price and the original trade price is exchanged in cash) or physical delivery (the underlying asset is actually delivered).

Risk parameters - **Price collars:** Static and dynamic collar band percentages specific to this instrument. - **Circuit breaker thresholds:** How much the price must move in a window before a halt is triggered. - **Position limits:** Maximum position any single participant may hold.

16.2 Why Reference Data Is So Dangerous to Get Wrong

Because reference data is read by every component, matching engine, gateway, clearing, market data, surveillance, a single wrong value can corrupt all of them simultaneously.

Tick size error: If the reference data says the tick size for AAPL is $0.10 instead of $0.01, the matching engine will reject all orders priced in odd cents as "tick-misaligned." The exchange will appear broken, even though the matching engine itself is working correctly.

Contract multiplier error: If a futures contract's multiplier is loaded as 500 instead of 50, every P&L calculation, every margin requirement, every risk check is ten times too large. Participants' risk limits will be breached on positions that should be well within bounds.

Expiry date error: If a futures contract's expiry is recorded as one day late, the matching engine will continue accepting orders for a contract that has already expired. Trades that should be impossible will be committed.

Session schedule error: If the closing auction is scheduled to begin 30 minutes earlier than intended, orders will stop matching during normal continuous trading for 30 minutes.

All of these have happened in real exchanges. Reference data errors have caused trading halts, regulatory interventions, and significant financial losses.

A prominent real-world example: on **8 July 2015, NYSE halted all trading for approximately 3.5 hours**. The cause was a software configuration mismatch introduced during a gateway update deployed the previous evening. The new gateway software version was incompatible with the version running on some internal systems, causing connectivity failures that propagated into a halt of the entire venue. Trading migrated to NASDAQ and other venues, demonstrating both the fragility of reference data and system configuration, and the resilience of the fragmented US equity market structure as a whole, since the market continued to function on other venues while NYSE was dark [NYSE Inc. statement, July 2015; SEC review].

16.3 Reference Data as an Exchange Engineering Problem

For developers, reference data has several important architectural properties:

It changes rarely but critically. Most reference data for a given instrument is stable for months. But corporate actions (the *Corporate Actions* section of Part IV) change it: a stock split can move a stock into a different tick-size band, any split or reverse split changes the prices and quantities of resting orders, and a name change can alter the symbol. These changes must be propagated atomically, all components must switch to the new values at the same moment, not over a period of minutes.

It is loaded at startup and cached aggressively. The matching engine reads tick sizes and price scales for every order it processes. If it had to query a database for each order, latency would be unacceptable. Reference data is loaded into memory at engine startup and cached. The flip side: if a cached value is stale, every order processed during the stale period is affected.

Version control and audit matter. Reference data changes must be versioned and audited. Regulators may ask: "What were the circuit breaker parameters for AAPL at 2:43pm on the day of the halt?" If reference data is overwritten rather than versioned, that question cannot be answered.

> **Key idea:** Reference data is the configuration that all exchange software depends on. Treat it with the same rigour as code changes, version controlled, tested before deployment, applied atomically, and auditable after the fact. A wrong tick size is just as dangerous as a bug in the matching loop.

Glossary

The following definitions are written to be concise and standalone, useful as a quick reference after reading the document.

A

Automated Collar Expansion (ACE): This book's name for rules that, when an instrument's likely price lies outside its protective collar, pause only that instrument, keep collecting orders, and widen the collar in published steps until the indicative auction price fits, instead of either blocking a genuine repricing or trading at an absurd price. Real venues use their own names (Nasdaq's Halt Cross price protections, NYSE's auction-collar extension after LULD pauses, Xetra's extended volatility interruption). See *Price Collars*, Part III.

Adverse Selection: The risk a market maker faces when their quote is hit by a counterparty who has superior information about where prices are heading, causing the market maker to trade at a price that will quickly move against them. A market maker who is repeatedly adversely selected will lose money even while quoting at "correct" prices. MMP exists specifically to limit adverse selection damage.

Aggressive order: (also called a **taker**), an incoming order that immediately trades against resting orders. It "sweeps" the book.

Aggressor / Taker: The participant whose incoming order triggered a match by crossing the spread to meet a resting order. The aggressor "takes" liquidity that was already available. Contrast with *Maker*, the participant whose resting order was already in the book.

Assignment: The process by which an options clearing house (the OCC, in the US) selects which option writer(s) must fulfil the obligation created when a holder exercises. Assignment passes from the OCC to a clearing member and from the clearing member to a specific client, typically via random or FIFO allocation. See *Exercise*, *Pin Risk*.

Automated Market Maker (AMM): A smart-contract-based pricing mechanism used by decentralised cryptocurrency exchanges in place of an order book. Prices are set algorithmically from the ratio of two assets held in a liquidity pool, commonly via a constant-product formula, rather than by matching a buy order against a sell order. See *Decentralised Exchange (DEX)*.

Automatic Exercise: The OCC's practice of exercising any in-the-money option on a holder's behalf at expiry (by a threshold as small as $0.01 in-the-money) unless the holder has filed a contrary instruction. See *Exercise*.

Alternative Trading System (ATS): A trading venue that matches orders but is not a registered national securities exchange. Includes ECNs and dark pools. Subject to lighter regulation than exchanges in some jurisdictions.

Arbitrage: The simultaneous purchase and sale of the same (or equivalent) asset on different markets to profit from a price difference. Pure arbitrage is risk-free (buy cheap here, sell expensive there at the same moment); in practice, most arbitrage involves some timing or execution risk. Arbitrage is the mechanism by which prices on different venues are kept consistent, arbitrageurs quickly eliminate price discrepancies, making markets more efficient. **Latency arbitrage** is a specific form exploiting the time it takes for price changes

to propagate between venues.

Ask / Offer: The price at which a seller is willing to sell. The best (lowest) ask is the top of the sell side of the order book. The collection of all active asks forms the sell side of the book.

ATC (At-The-Close): An order TIF valid only during the closing auction.

ATO (At-The-Open): An order TIF valid only during the opening auction.

Auction / Call Auction / Uncross / Fixing: A trading mechanism where orders are collected over a period and then matched simultaneously at a single equilibrium price. Contrasts with continuous matching where orders match one at a time as they arrive.

Auction Imbalance: The excess of buy orders over sell orders (or vice versa) at the indicative uncross price during an auction. Published by exchanges to encourage offsetting orders before uncrossing.

Audit Trail: The complete, immutable, time-ordered record of every event in the exchange system, every order, fill, cancellation, and rejection, with precise timestamps. Required for regulatory reporting and replay-based recovery.

B

Big Bang: The 27 October 1986 deregulation of the London Stock Exchange, which abolished fixed commissions, ended the broker/jobber separation, permitted outside and foreign ownership of member firms, and replaced floor trading with the screen-based SEAQ system, all on a single day. The UK counterpart to the US's 1975 "May Day." See *May Day*.

Backwardation: A futures market in which near-dated contracts are more expensive than later-dated ones, so a calendar spread priced as near minus far is positive. The opposite of *contango*.

Basis: In derivatives and futures markets, the price difference between two related instruments, most commonly between the spot/cash price and a futures price on the same underlying, or between two futures contracts with different expiry dates. In the implied-order examples in this document (the *Implied Orders and Synthetic Liquidity* chapter of Part II), the spread price of −$2 between the January and February crude oil contracts is the basis between those two contract months: January is $2 cheaper than February. The basis is not fixed; it changes continuously as supply and demand for the near and far months evolve independently. **Basis risk** is the risk that the expected relationship between two instruments diverges unexpectedly, a spread trader who expected the two legs to move together may find they do not, leaving an unintended net exposure. See *Implied Order*, *Calendar Spread*.

Bear Market: A sustained decline of 20% or more from a recent market peak. Originates from the image of a bear swiping its paws downward. Contrast with *Bull Market*.

Best Ask: The lowest resting sell price in the order book, the price at which a buyer can trade immediately.

Best Bid: The highest resting buy price in the order book, the price at which a seller can trade immediately.

Best Execution: The regulatory obligation for brokers and investment firms to take all reasonable steps to achieve the best possible result for their clients when executing orders. "Best" considers price, execution costs, speed, likelihood of execution, and market impact simultaneously, not just the quoted price in isolation. Mandated by MiFID II in the EU and the SEC's duty of best execution in the US. Best execution is the regulatory foundation that makes smart order routing (SOR) necessary: brokers must demonstrate their routing decisions genuinely serve clients' interests, not just the broker's economics.

Bid: The price at which a buyer is willing to buy. The best (highest) bid is the top of the buy side of the order book.

Bid-Ask Imbalance: A measure of depth asymmetry: bid_depth / (bid_depth + ask_depth), computed within a symmetric price window around the mid. Values near 1 indicate heavy buying interest; values near 0 indicate heavy selling interest. Used as a short-term price direction signal.

Blue Chip: A large, financially sound, well-established company, the most prestigious tier of the equity market. The term originates from poker, where blue chips represent the highest denomination.

Bond: A standardised debt instrument. The issuing company borrows money and promises to pay periodic

interest (coupon payments) and return the face value (principal) at maturity. Bondholders are creditors, not owners, and are paid before equity holders in a bankruptcy. Bonds trade on secondary markets after issuance, allowing investors to exit before maturity. The financial term "bond" traces its roots back to the 14th century from the Old English word bindan and band, which eventually evolved into the standard English word "bind". The word was originally used to describe something that literally ties, fastens, or holds things together. Because a financial bond fundamentally operates as a formal, legally binding contract where an issuer promises to repay a loan with interest, the term stuck.

Book-Building: The customary way a conventional IPO is priced: the issuer and its underwriters publish a price range, collect indications of interest from institutional investors during a roadshow, revise the range, and set the final offer price the evening before trading starts. A process of price discovery, not a legal requirement.

Book Depth: The total volume (quantity of shares) resting near the top of the order book. A deep book absorbs large orders without significant price impact. A thin (or shallow) book moves quickly on even moderate order flow.

Book Snapshot: A periodic complete view of the current order book state (all price levels and their quantities) published by the exchange for subscribers who need to get up to speed.

Borrow Rate: The annualised fee charged to a short seller for borrowing shares from a lender. Easy-to-borrow large-cap stocks typically have borrow rates near zero. Hard-to-borrow, heavily-shorted, or illiquid stocks can carry rates of 5–50% or more per annum, applied daily to the short position. Rising borrow rates signal increasing short interest and decreasing availability of lendable shares, and are themselves a market signal.

Bracket Order: A strategy combining an entry order with both a take-profit LIMIT and a stop-loss STOP, typically implemented as an OCO pair.

Broker: An intermediary who executes orders on behalf of clients without trading for their own account. The word traces to Old French/Norse roots meaning someone who opens a cask and sells by the cup, a retail intermediary.

Bull Market: A sustained rise of 20% or more from a recent market trough. Originates from the image of a bull thrusting its horns upward. Contrast with *Bear Market*.

Buy-In: A remedy for failed settlement: if a seller cannot deliver securities, the buyer purchases them in the open market at the seller's expense.

C

Calendar Spread: A derivatives strategy involving the simultaneous purchase and sale of contracts on the same underlying instrument but with different expiry dates, for example, buy the March futures contract, sell the June futures contract. Used to profit from expected changes in the price difference (the "spread") between contract months. On CME and most futures venues it is also a listed instrument with its own order book, priced as near month minus far month; *buying* the spread means buying the near month and selling the far month, so the buyer profits when that difference rises. Spread prices are routinely negative. Requires cross-symbol coordination in the matching engine. See *Implied Order*.

Call auction: Synonym for auction, "call" refers to the fact that all participants are "called" to submit orders before a single price is determined.

Capital Appreciation: An increase in the market price of a share (or other asset) above the price paid. The primary way equity investors generate returns.

Capital Gain: The profit realised when an asset is sold for more than it was purchased for.

Cascade Cancel: When a COMBO order fails (any leg is cancelled or rejected), all remaining unfilled legs of the combo are automatically cancelled. Previously executed fills are not reversed.

CCP (Central Counterparty Clearing House): An entity that stands between buyer and seller in a trade, becoming the legal counterparty to both, eliminating direct counterparty credit risk.

Circuit Breaker: An automatic trading halt triggered when price movement exceeds a threshold within a time window, designed to prevent cascading automated trading crashes.

Clearly Erroneous Execution (CEE): A trade whose price deviates from a reference price by more than a published, price-tiered percentage threshold, making it eligible for cancellation (busting) or price adjustment under exchange rules. Reviewed execution-by-execution within a short time window (commonly 30 minutes); multi-venue events require coordinated rulings across exchanges. Distinct from a trade that is merely unwelcome to the participant who sent it, the test is an objective price deviation, not the sender's intent. See *Trade Busting and Clearly Erroneous Trades*, Part III.

Conformance Testing: The process by which a prospective exchange participant's trading system is verified, against a prescribed test script in a certification (UAT) environment, to correctly implement the exchange's protocol and expected behaviours before being granted production credentials.

Contango: A futures market in which later-dated contracts are more expensive than near-dated ones (common when the cost of storing the underlying is high), so a calendar spread priced as near minus far is negative. The opposite of *backwardation*.

Continued Listing Standards: The ongoing, ordinarily automated compliance thresholds (minimum bid price, market value of public float, shareholder count) a listed company must keep satisfying after its initial listing, distinct from the one-time *Initial Listing Standards*. Breach triggers a deficiency notice and cure period before involuntary delisting.

Crossed Market: A condition where the best bid on one trading venue is higher than the best ask on another venue for the same instrument, a transient artefact of fragmentation and quote-propagation latency across independent venues, and a theoretical risk-free arbitrage opportunity for whoever can act on it fastest. See *Locked Market*.

Cure Period: The window (commonly 180 days, sometimes extended) an exchange gives a listed company to regain compliance with a continued listing standard after a deficiency notice, before involuntary delisting proceedings begin.

Clearing: The process of confirming and recording a trade, and (via a CCP) guaranteeing its settlement.

Co-location: Placing a participant's trading servers physically inside the same data centre as the exchange's matching engine, minimising network round-trip latency to microseconds.

Collar Band / Price Collar: A price filter that rejects orders whose submitted price deviates too far from a reference price. Two distinct variants: the *static price collar* (compares against the last close) and the *dynamic price collar* (compares against the last trade). See those entries for details.

Combo Order: A single instruction to execute orders in multiple instruments simultaneously. Each component is a leg.

Conflation (Market Data): Merging consecutive market data updates for the same price level into a single message under high-load conditions. Reduces message volume at the cost of losing intermediate states.

Continuous Trading: The normal trading mode where orders match as soon as they arrive and a compatible counterparty is present. Contrasts with auction mode.

Contract Multiplier: For futures and options, the dollar (or other currency) value of a one-unit price move per contract. The E-mini S&P 500 has a multiplier of $50, a 1-point move in the index is worth $50 per contract. Essential reference data: a wrong multiplier affects every P&L and margin calculation.

Corporate Action: An event affecting a company's share structure, stock splits, dividends, mergers, delistings. Requires adjustments to all open orders, positions, and reference data in exchange systems.

Counterparty Risk: The risk that the other party to a trade defaults before settlement is complete. Cleared via CCP novation, which replaces bilateral counterparty risk with exposure to the CCP.

Coupon: The periodic interest payment on a bond. Named after the physical coupon strips that bond holders historically detached and redeemed for payment.

Crossing the Spread: Submitting an order at a price that matches an existing resting order, triggering an immediate fill.

Cum-Dividend: The period during which a share trades *with* entitlement to a declared dividend, the buyer receives the upcoming payment. A share trades cum-dividend from the dividend announcement up to (but not including) the ex-dividend date. After the ex-dividend date the share trades *ex-dividend* and the buyer is not

entitled to the payment. The price typically drops by approximately the dividend amount on the ex-dividend date.

CUSIP: A 9-character identifier for US and Canadian securities used in clearing and settlement. Analogous to ISIN.

Custody: The safekeeping of securities on behalf of investors by a broker or custodian bank, with records maintained at a central securities depository.

D

Dark Fibre: Unused optical fibre cable already laid in the ground but not yet carrying data traffic ("lit"). Trading firms and data providers lease dark fibre from infrastructure companies and run their own optical transceivers, creating private, dedicated network links between financial data centres. Dark fibre links are used between major exchange co-location sites (e.g., NYSE in Mahwah and NASDAQ in Carteret) to reduce cross-venue latency for arbitrage and hedging strategies.

Decentralised Exchange (DEX): A cryptocurrency trading venue that executes trades directly via smart contracts on a blockchain, with no central operator holding custody of participant assets and, commonly, no traditional order book, matching instead via an *Automated Market Maker*. Contrast with a *Centralised Exchange (CEX)*, which operates a conventional matching engine and custodies participant assets itself.

Decimalization: The switch from fractional dollar price quoting (eighths, then sixteenths) to decimal cent quoting in US equity markets, phased in between August 2000 and April 2001. Quoted spreads on liquid stocks roughly halved as a direct result. The direct historical motivation for Rule 612's sub-penny quoting floor. See *Tick Sizes and Fractional Ticks*, Part II.

De-SPAC: The merger transaction in which a Special Purpose Acquisition Company (SPAC) combines with a private operating company, resulting in the operating company becoming publicly traded under the (renamed) shell's listing. See *SPAC*.

Direct Listing: A method of going public in which a company's existing shares begin trading on an exchange via the opening auction mechanism, with no new shares issued, no underwriter guarantee, and no pre-set offer price. Contrast with a traditional underwritten IPO. Pioneered at scale by Spotify (2018) and Slack (2019), both on NYSE.

Dark Pool: A private trading venue where orders are not displayed in a public order book. Participants submit orders and are matched against other participants' orders, typically at the midpoint of the NBBO, with the trade only reported publicly after it occurs. Used by large institutional investors to trade without revealing their intentions and causing adverse price movement. Contrast with *lit markets* (public exchanges) where resting orders are visible to all. See *Regulatory Surveillance*.

DAY: An order TIF valid only for the current trading session.

Delivery versus Payment (DvP): The settlement principle that the transfer of securities and cash happen simultaneously and conditionally, neither is released without the other. Eliminates principal risk in settlement.

Depth (Market Depth): The total quantity of resting orders available at each price level. Deep markets can absorb large orders with minimal price impact.

Depth-of-Book (Level 2): Market data showing multiple price levels and their resting quantities, rather than only the best bid and ask.

Designated Market Maker (DMM): A firm assigned by an exchange to maintain fair and orderly markets in specific securities, with both privileges and obligations. NYSE's term; Nasdaq calls them Market Makers.

Determinism (Exchange): The property that given the same initial state and the same ordered input messages, the matching engine always produces identical outputs. Essential for audit trail replay and disaster recovery.

Discounted Cash Flow (DCF): A valuation method that estimates a company's future free cash flows and discounts each one back to today's money at a rate reflecting risk and the time value of money. For young, fast-growing firms most of the result usually sits in the *terminal value*, the lump that stands for all years beyond the forecast.

Dilution: The reduction in existing shareholders' ownership percentage when new shares are issued. Selling 20% of a company to new investors reduces existing owners to 80%.

Dividend: A cash payment made by a company to its shareholders, distributed from retained profits. Not guaranteed, the board of directors decides whether, when, and how much to pay. Companies that pay regular dividends are typically mature, profitable businesses with stable cash flows. High-growth companies often pay no dividend, choosing to reinvest profits in expansion. The key dates are: the *declaration date* (when the board announces the dividend), the *ex-dividend date* (buyers on or after this date do not receive the payment), the *record date* (the register of shareholders eligible to receive payment), and the *payment date* (when the cash is distributed).

Drop Copy: A real-time copy of all order lifecycle events for a specific participant, sent to a designated third party such as a clearing broker or risk system.

DTCC (Depository Trust and Clearing Corporation): The central clearing and settlement infrastructure for US equity markets. Its subsidiary DTC is the US central securities depository.

Dynamic Price Collar (Reference Price Filter): A pre-trade risk control that rejects any order whose submitted price deviates more than a configured percentage from the most recent traded price. Protects against algorithms that walk the price incrementally through a series of small trades. Contrast with *Static Price Collar*, which uses the last close as the reference. Together, the two collars provide overlapping protection against both sudden errors and gradual drift.

E

ECN (Electronic Communication Network): An electronic trading system that automatically matches buy and sell orders, typically with narrower spreads than exchanges but less regulatory oversight.

Equilibrium Price (Auction Price): The single price at which the maximum volume can trade in a call auction.

Equity: Ownership stake in a company. Equity holders (shareholders) have a residual claim on the company's assets and earnings after all creditors are paid.

ETF (Exchange-Traded Fund): A basket of securities (shares, bonds, commodities) that itself trades as a single share on an exchange.

Ex-Dividend Date: The date from which a stock trades without entitlement to a declared dividend. Buyers on or after this date do not receive the dividend.

Execution Algorithm (Algo): Software that breaks a large institutional order into smaller pieces and executes them over time to minimise market impact and achieve a target average price. Common types include VWAP (match the day's volume-weighted average price), TWAP (distribute uniformly over time), Implementation Shortfall (minimise the gap between decision price and execution price), and POV/In-Line (participate at a fixed percentage of market volume). Execution algorithms generate the majority of order flow on modern exchanges.

Exercise: An option holder's invocation of their right to buy (call) or sell (put) the underlying at the strike price. American-style options may be exercised any time up to expiry; European-style options only at expiry. See *Automatic Exercise*, *Assignment*.

F

Face Value (Bond): The principal amount of a bond, the amount the issuer promises to repay at maturity, and the basis on which coupon payments are calculated.

Failover: The process of switching from a failed primary system to a secondary (backup) system.

Fat Finger: Industry slang for an accidental trading error caused by mistyping an order parameter (quantity, price, etc.).

FIFO (First-In, First-Out): Within a price level, orders are filled in the order they arrived. Earlier arrivals have priority.

Fill: An execution, a match between a buy and sell order resulting in a trade. Partial fills consume part of an order; a complete fill consumes all of it.

FOK (Fill-Or-Kill): An order that must be completely filled immediately or cancelled entirely.

FIX (Financial Information eXchange): The industry-standard protocol for electronic order submission, used by most exchanges and brokers globally.

Flat: Having no open position in a symbol (net quantity = 0). Neither long nor short.

Floor Broker: Historically, a broker who executed trades on the physical exchange floor. The term persists in regulations and informal speech.

FPGA (Field-Programmable Gate Array): Reconfigurable hardware chips that can implement market data parsing or order routing in dedicated logic circuits, achieving nanosecond-level latency.

Front Running: Acting on advance knowledge of a pending client order to trade for one's own account before executing the client's order, then profiting from the price movement the client's order causes. Illegal under market manipulation regulations and a serious breach of fiduciary duty. Front running is detected by surveillance systems that flag cases where a participant's own trades immediately precede large client executions in the same direction.

Futures Contract: A standardised agreement to buy or sell a specified quantity of an underlying asset (commodity, index, currency, interest rate) at a predetermined price on a specified future date. Obligations are legally binding. Futures are exchange-traded, margined, and marked-to-market daily. Unlike equity ownership, futures do not represent ownership of an underlying asset, they represent an obligation about a future transaction.

G

Greenshoe (Over-Allotment Option): An option letting IPO underwriters sell up to 15% more shares than the base deal. If the price sags after listing they buy shares back in the market to support it; if it rises they exercise the option and buy the extra shares from the issuer at the offer price.

Gap (Market Data): A missing sequence number in a market data feed, indicating a lost message. Subscribers must request retransmission or resynchronise from the next full snapshot.

Gateway: The participant-facing interface to the exchange. Handles authentication, session management, and message translation.

Going Long: Owning an asset in expectation that its price will rise, so it can be sold later at a profit. You are "long" the asset. The term originates in physical commodity trading: a merchant who had purchased durable goods, grain, spices, metal, cloth, and was holding them in a warehouse was described as "long" in those goods. The word captured both *possession* (you own something tangible) and *duration* (durable goods last a long time and can be held through seasons or years). Being long in modern finance carries exactly the same meaning: you hold the asset and are exposed to its price over time. A long position profits when the price rises and loses when it falls.

Going Short / Short Selling: Selling an asset you do not currently own, typically by borrowing it from another party, with the intention of buying it back later at a lower price and returning it, pocketing the difference. The term originates in physical commodity trading: a merchant who had sold goods they did not yet possess, committing to deliver more than their warehouse held, was "short" of the goods. The same word family as "falling short," "short of supplies," or "shortage", deficient, insufficient, lacking. In the early forward markets of Amsterdam and London, a short seller would sell forward (promise future delivery), then purchase the goods in the market before delivery, hoping prices had fallen. Isaac Le Maire's 1609 campaign against VOC shares is the earliest recorded large-scale short selling operation. Today, a short position profits when the price falls (you buy back cheaper than you sold) and loses when it rises, in theory without limit, since a rising price has no ceiling.

GTC (Good-Till-Cancelled): An order TIF that persists across trading sessions until filled or cancelled.

GTD (Good-Till-Date): An order Time-In-Force variant that keeps an order active until it fills, is explicitly cancelled, or reaches a specified expiry date, whichever comes first. On the expiry date the order is automat-

ically cancelled. Provides the persistence of GTC without the open-ended risk of an order remaining active indefinitely. Requires the exchange scheduler to track each GTD order's expiry and issue the cancellation at the right time.

H

Heap: A data structure that gives O(1) access to the minimum (or maximum) element. Used in matching engines to efficiently maintain best price access.

High-Frequency Trading (HFT): Automated trading strategies that execute very large numbers of orders at extremely high speeds, typically holding positions for milliseconds or seconds rather than minutes or days. HFT firms invest heavily in low-latency infrastructure (co-location, microwave links, FPGAs) to gain speed advantages. Strategies include latency arbitrage, market making, and statistical arbitrage. Controversial because speed advantages can disadvantage slower participants; the subject of the IEX speed bump.

I

Iceberg Order / Hidden Order / Reserve Order: An order that shows only a small visible peak while hiding a large reserve. Each time the visible portion fills, the order replenishes from the reserve.

Imbalance: The difference between total buy and sell quantity at the equilibrium price. Expressed as a side (BUY or SELL) and a quantity.

Immediate-Or-Cancel (IOC): An order that fills as much as possible immediately and cancels any unfilled remainder. Never rests in the book.

Implementation Shortfall (IS) / Arrival Price Algorithm: An execution algorithm that minimises the difference between the *decision price* (the market price when the investment decision was made) and the actual average execution price. IS algos trade faster when the price moves against the position (urgency increases to avoid further shortfall) and slower when price moves favourably. More adaptive and sophisticated than VWAP or TWAP; preferred by quantitative funds whose models predict short-term price moves.

Implied Generation: A count of how many real orders an implied quote combines, minus one. A first-generation implied quote is built from two real orders; a second-generation quote uses a first-generation implied quote as an ingredient and so combines three. Venues restrict how many generations they compute because the work grows explosively. See *Multi-Layer Implied Orders and the Combinatorial Wall*, Part II.

Implied Matching: A technique used in derivatives markets where the exchange combines resting spread and outright orders (in different books) into a synthetic order in a third book, and lets incoming orders trade against it, all contributing orders executing together. For example: a resting spread order to buy March/sell June combined with a resting outright June bid creates an implied March bid. Implemented on CME Globex, ICE, and (as "synthetic matching") Eurex. Implied matching creates no new liquidity: it makes existing orders reachable from another book.

Implied Order: An order derived by the matching engine from a combination of existing spread and outright orders. Not submitted by any participant, computed from pre-existing orders and published in the relevant outright or spread book. When an implied order matches, all contributing orders execute simultaneously and atomically. If any contributing order is cancelled or filled by other means, the implied order shrinks or disappears in the same processing step. Implied orders never act as the aggressor, and at the same price real orders trade before them.

Implied IN: In CME's terminology, an implied order in the *spread* book created from orders in the outright books. Example: a January bid at \$74.40 and a February ask at \$76.60 together imply a Jan/Feb spread *bid* at $74.40 - 76.60 = -\$2.20$.

Implied OUT: In CME's terminology, an implied order in an *outright* book created from a spread order plus an outright order. Example: a January ask at \$75.00 and a Jan/Feb spread *bid* at −\$2.00 together imply a February *ask* at $75.00 - (-2.00) = \$77.00$.

Inactivation (Quote): Automatically cancelling or suspending the surviving leg of a two-sided quote the moment the other leg is filled. The market maker must re-quote both sides afresh within their re-quoting

window. Used in conservative quote refresh policies (Eurex-style). Protects the market maker from being left with a stale unhedged quote after their inventory position has changed.

Indicative Uncross Price: During an auction accumulation period, the price at which uncrossing *would* occur if matching happened at that moment. Recalculated after every order arrival and published as a signal to help participants decide whether to adjust their orders before the actual uncross.

Information Asymmetry: A situation where some market participants have access to information (faster data, better analytics) that others do not, giving them a trading advantage. Auctions reduce information asymmetry by letting all participants submit orders before any execution occurs.

Informed Flow: Order flow that originates from participants who have an informational advantage about near-term price direction, for example, participants who have seen relevant news before others. When a market maker is repeatedly hit by informed flow, they accumulate inventory at prices that are about to move against them. Informed flow is the root cause of adverse selection.

Initial Listing Standards: The quantitative (minimum share price, public float value, shareholder count) and qualitative thresholds a company must meet before an exchange will admit its shares to trading for the first time. Distinct from *Continued Listing Standards*, which apply on an ongoing basis thereafter.

Initial Margin: Collateral deposited when opening a position to cover potential losses. Calculated by the CCP based on the instrument's historical volatility and the position size.

Insider Trading: Trading on material, non-public information, for example, buying shares before a publicly unannounced merger is announced. Illegal in all major jurisdictions. The exchange's audit trail is a key source of evidence in insider trading investigations, as regulators can reconstruct exactly who traded, when, and at what prices relative to when news became public.

Instrument / Symbol: The specific financial product being traded. Each instrument has its own order book.

Internaliser: A broker-dealer that matches client orders internally against its own inventory or other clients' orders, rather than routing to an exchange. A form of OTC secondary trading.

IOC (Immediate-Or-Cancel): A time-in-force (and order type) where an order sweeps available liquidity at the limit price and any unfilled remainder is immediately cancelled. Never rests on the book

IPO (Initial Public Offering): The first sale of a company's shares to the general public, after which the shares are listed and traded on a stock exchange.

ISIN (International Securities Identification Number): A 12-character global identifier for a security, independent of the exchange it trades on. A single company has one ISIN regardless of where its shares are listed.

ITCH (NASDAQ TotalView-ITCH): NASDAQ's binary, message-based market data feed protocol for order book dissemination. ITCH publishes sequenced events such as add order, modify/replace, cancel/delete, executions, and auction/trade messages so subscribers can reconstruct the full limit order book deterministically from the stream. Official specification: NASDAQ TotalView-ITCH 5.0.

K

Kernel Bypass: Technology that allows a network application to communicate directly with network hardware, bypassing the operating system kernel and reducing latency from microseconds to sub-microsecond.

Kill Switch: An emergency control that immediately cancels all resting orders and quotes for a specific participant AND marks their gateway as inactive, preventing new order submission until the participant reconnects and re-authenticates. Triggered by the participant (emergency self-halt), the exchange operator, or automatically on gateway disconnect. Mandatory under most exchange regulations including MiFID II. Contrast with *Mass Cancel*, which cancels orders but leaves the gateway active.

L

Latency: The time delay between an action and its effect. In exchanges, latency is measured in microseconds or nanoseconds. Low latency is a competitive advantage.

Latency Arbitrage: Trading that profits from the brief period between when a price changes on one venue and when that change propagates to another venue.

Layering: A market manipulation practice: placing large orders to create a false impression of market depth, then cancelling them once the desired price movement occurs.

LCH: A major international clearing house, part of the London Stock Exchange Group. Clears interest rate swaps, equities, foreign exchange, and other products.

Leg: One component order within a combo order or spread strategy.

Leg Risk / Legging Risk: The risk that one leg of a combo order fills while another does not, leaving an unintended single-instrument position.

Level 1 Data: The best bid and ask prices and quantities only.

Level 2 Data (Market Depth): The full order book, all price levels and their quantities.

Limit Order: An order to trade at a specified price or better. May rest in the book if not immediately matched.

Limit Up-Limit Down (LULD): A US regulatory circuit breaker mechanism that pauses trading in an individual stock if its price moves outside a percentage band (the "price band") calculated from a reference price. The bands are tighter for liquid large-cap stocks and wider for smaller, more volatile ones. Trades outside the band are not allowed. If the best bid or offer sits exactly at the band (a "limit state") and does not recover within 15 seconds, a 5-minute trading pause begins. Implemented across all US equity trading venues.

Liquidity: The ease with which an asset can be bought or sold without significantly affecting its price. High liquidity = tight spread, large depth, easy trading.

Locate: The process of finding and reserving a source of borrowable shares before executing a short sale. Under US Regulation SHO, broker-dealers must have reasonable grounds to believe shares can be borrowed before accepting a short sale order. "Failing to locate" before shorting is called naked short selling, which is illegal in most jurisdictions.

Lock-Up: A contractual agreement, typically 180 days, under which insiders and pre-IPO investors promise not to sell their shares after an IPO. It is a contract with the underwriters, not a legal rule, and its expiry can put pressure on the share price.

Locked Market: A condition where the best bid on one venue equals the best ask on another venue for the same instrument, a transient artefact of fragmentation and quote-propagation latency. See *Crossed Market*.

Long Position: Owning a positive quantity of shares. You profit if the price rises. Expressed as a positive number P&L ledger.

Lot: The standard unit of quantity for an instrument.

M

Maintenance Margin: The minimum collateral balance that must be maintained in a margin account. Falling below this level triggers a margin call.

Maker: The participant whose resting order was already in the book when a fill occurred. Contrast with *Aggressor/Taker*.

Margin Call: A demand from a clearing house or broker for a participant to deposit additional collateral to restore their margin account to the required level.

Mark-to-Market: The daily revaluation of positions at current market prices, with cash payments made to reflect daily gains and losses. The mechanism by which variation margin is calculated.

Market capitalisation (market cap): The total market value of all a company's outstanding shares, calculated as share price multiplied by the total number of shares in existence. If Apple has approximately 15.4 billion shares outstanding and each trades at $190, Apple's market cap is roughly $2.9 trillion [1]. Market cap is the most widely used shorthand for a company's size. When rankings refer to "the world's largest exchange by listed market cap," they are summing the market caps of every company listed there.

Market Fragmentation: The distribution of trading volume for a single instrument across multiple competing venues, exchanges, dark pools, ATSs, and internalisers. Modern equity markets are highly fragmented; in the US, AAPL may trade simultaneously across a dozen registered exchanges plus numerous alternative venues. Fragmentation improves competition and reduces fees but complicates order routing and best-execution analysis. Regulatory frameworks like Regulation NMS (US) and MiFID II (EU) were introduced partly in response to fragmentation.

Market Impact / Slippage: The adverse price movement caused by a large order sweeping through multiple price levels.

Market Impact Estimation: The process of computing, before order submission, the average price you would pay (or receive) by sweeping through available book levels for a given order size. Derived from the cumulative depth profile. Used by execution algorithms to size order slices and choose between aggressive execution and passive accumulation.

Market Maker: A professional participant obligated to continuously provide both buy and sell prices, ensuring liquidity for other participants.

Market Maker Protection (MMP): An exchange-provided mechanism that automatically cancels all of a market maker's resting quotes when fills arrive at a rate exceeding a configured threshold within a rolling time window. Protects market makers from being rapidly picked off by participants with superior information (adverse selection). After MMP fires, the market maker enters a protection period before re-quoting. Parameters (fill count, time window) are configured per market maker per symbol.

Market Order: An order that attempts immediate execution against available liquidity at the best obtainable prices. It prioritises execution certainty over price certainty, but execution is still subject to exchange state, available liquidity, price collars, and risk controls. A market order submitted during a halt, against an empty book, or rejected by a pre-trade risk check will not fill.

Mass Cancel: A participant-initiated message type that cancels multiple resting orders simultaneously, all orders for a symbol, all orders of a given type, or all orders system-wide. The gateway connection remains active after a mass cancel; the participant can submit new orders immediately. Contrast with *Kill Switch*, which inactivates the gateway and requires re-authentication.

Matching Engine: The core software that manages order books and matches buy and sell orders to produce trades.

Maturity (Bond): The date on which a bond's issuer must repay the face value to the bondholder.

May Day: 1 May 1975, the date the SEC abolished fixed brokerage commissions on the NYSE, ending nearly two centuries of fixed-rate trading and creating the discount-brokerage industry. The US counterpart to the UK's 1986 "Big Bang." See *Big Bang*.

Microwave Link: A radio communications link using microwave frequencies. Between financial data centres, microwave links provide lower latency than fibre-optic cables because radio waves travel faster through air than light through glass.

Mid Price: The average of the best bid and best ask: (best_bid + best_ask) / 2.

Midpoint Peg Order: An order whose price continuously tracks the current mid price (the average of the best bid and best ask). Executes only when a counterparty arrives whose order is also willing to trade at the midpoint. The mid is a *better* price for both parties than the quoted bid or ask, the buyer pays less than the ask, the seller receives more than the bid; the spread is shared rather than fully paid. Provides price improvement at the cost of uncertain execution timing: there is no guarantee a counterparty will appear. Common in dark pools; also supported on IEX on the lit market. Never competes in the regular price-time priority queue, maintained as a separate midpoint matching layer.

MMP Threshold: The fill count and time window that together trigger Market Maker Protection. For

example, "5 fills within 1 second." If this threshold is exceeded, the exchange automatically cancels all of the market maker's resting quotes. The threshold is negotiated per market maker, per symbol, and finding the right value is a key parameter of market making strategy.

Monotonic: Strictly non-decreasing. A monotonic sequence is one where each value is equal to or greater than all previous values. A strictly monotonic sequence is one where each value is strictly greater.

N

Naked Short Selling: Selling shares short without first locating and arranging to borrow them. The seller delivers nothing at settlement, creating a "fail to deliver." Illegal in most jurisdictions under short sale regulations (Regulation SHO in the US). Distinguished from covered short selling, where shares have been located and borrowed.

Nanosecond: One billionth of a second (10 seconds). Modern matching engines record timestamps at nanosecond precision.

NBBO (National Best Bid and Offer): The best available bid and ask prices aggregated across all US exchanges, as required by Regulation NMS. Brokers must offer clients execution at or better than the NBBO.

NAV (Net Asset Value): The total value of a fund's holdings divided by the number of shares outstanding. For mutual funds, NAV is computed daily using the official **closing price** of each holding, one reason the closing price matters so much.

Novation: The legal process by which a CCP replaces the original buyer-seller relationship: the CCP becomes buyer to the seller and seller to the buyer, eliminating bilateral counterparty risk.

NTP (Network Time Protocol): A protocol for synchronising computer clocks over a network. Basic NTP achieves millisecond accuracy; higher-precision variants achieve sub-millisecond.

O

OCC (Options Clearing Corporation): The central clearing house for equity options in the US.

Odd Lot: An order or execution for a quantity that is not a whole multiple of the standard round lot size for that instrument (commonly 100 shares for US equities). Historically excluded from consolidated NBBO and tape data in US equities, a practice being revised under the SEC's market-data-infrastructure rules. See *Round Lot*.

OCO (One-Cancels-Other): Two linked orders where filling either automatically cancels the other.

OHLCV: Open, High, Low, Close, Volume, the five standard summary statistics recorded for each trading session (or any time period). Open: the first traded price. High: the highest traded price. Low: the lowest traded price. Close: the last traded price (the official closing price in a closing auction). Volume: total shares traded. OHLCV data forms the basis of almost all price charts and historical market analysis.

Open Outcry: The historical method of trading on exchange floors, where traders shouted bids and asks and used hand signals in a sunken circular area called a pit. Largely replaced by electronic trading.

Options: Contracts that give the buyer the right, but not the obligation, to buy (a **call option**) or sell (a **put option**) an underlying asset at a specified price (the **strike price**) before or on a specified date (the **expiry date**). The seller of an option takes on the corresponding obligation. Options are priced using models (Black-Scholes being the most famous) that account for time to expiry, volatility, interest rates, and the distance between the current price and the strike. Cboe is the world's largest options exchange.

Order: An instruction to buy or sell a quantity of an instrument at a specified price or better.

Order Book: The central data structure of a matching engine, all resting buy and sell orders for a symbol, organised by price.

Order ID: A unique identifier assigned to each order upon submission. Must be unique across the entire system.

OTC (Over-the-Counter): Trading that occurs directly between two parties without going through an

exchange. OTC trades are less transparent and subject to different regulations than exchange-traded transactions.

OUCH (NASDAQ OUCH): NASDAQ's high-performance binary order entry protocol used by member firms and low-latency trading systems to submit, modify, cancel, and manage orders directly at the matching engine. OUCH is session-based and message-oriented, with exchange acknowledgments and execution/cancel responses that let clients track the full order lifecycle deterministically. Official specification: NASDAQ OUCH 5.0.

P

P&L (Profit and Loss): The financial gain or loss on a trading position.

Passive Order: See *Resting Order*

Payment for Order Flow (PFOF): A practice in which retail brokers receive payment from wholesale market makers in exchange for routing client order flow to those market makers for internalisation rather than to lit exchanges. Market makers pay for retail flow because retail orders have lower adverse selection risk. Effectively banned in the UK since 2012, and banned EU-wide from 30 June 2026 under the 2024 MiFIR review. Legal but controversial in the US, where it subsidises zero-commission brokerages but raises concerns about conflicts of interest and reduced price discovery. See *Best Execution*, *Internaliser*.

Pin Risk: The risk faced by an option writer when the underlying closes at expiry very close to the strike price, leaving the writer uncertain whether automatic exercise and assignment will occur until after markets have closed, with any resulting delivery obligation priced at the strike regardless of where the underlying opens the next session. See *Exercise*, *Assignment*.

Pit: The sunken, octagonal or circular physical trading area on an exchange floor where open outcry trading happened. CME Group operated prominent futures pits in Chicago. Largely dormant today.

Portfolio: A collection of investments held by an investor, typically diversified across asset types (equities, bonds, etc.) to manage risk and return.

Position: A participant's current holding in an instrument, positive (long) if they own more than they have committed to deliver; negative (short) if they have committed to deliver more than they own; zero (flat) if both sides balance. Position tracking is the clearing system's primary responsibility.

Pre-Trade Risk Controls: Checks applied to orders before they reach the matching engine, quantity limits, notional limits, fat-finger filters, credit checks, and rate limiting. The gateway is typically the enforcement layer.

PRE_OPEN: Session phase before the opening auction. Orders are accepted and rest on the book, but no matching occurs.

Presence Obligation: A market maker's contractual requirement to have live two-sided quotes resting in the book for a minimum percentage of the trading session, commonly 85% or more. Prevents market makers from only quoting on easy days and disappearing on volatile ones when liquidity is most needed.

Price Discovery: The process by which the market determines the current fair price of an asset through the interaction of buyers and sellers. A well-functioning exchange enables efficient price discovery: prices reflect all currently available information and update rapidly as new information arrives. Price discovery is one of the three core promises every exchange makes. Poor price discovery, where the traded price diverges significantly from fair value for extended periods, is a sign of market dysfunction.

Price Level: All resting orders at the same price on the same side of the book.

Price-Time Priority: The matching rule: better prices fill first; at the same price, earlier arrivals fill first.

Primary Market: Where new securities are issued and sold for the first time. The company (or government) receives the proceeds.

Primary Site: The active, live data centre running the production matching engine.

Private Company: A company whose shares are not publicly traded, held only by founders, employees, and private investors.

Prospectus: The disclosure document a company must publish before offering securities to the public. In the EU it is governed by the Prospectus Regulation (EU) 2017/1129 and approved by a national competent authority; in the US it forms part of the S-1 registration statement.

Pro-Rata Allocation: An alternative to price-time priority where fills at a price level are distributed proportionally among all resting orders at that level. Common in some futures markets.

Protection Period: The window of time after MMP fires during which a market maker has no active quotes and is expected to assess market conditions before deciding whether and at what prices to re-enter. The duration is typically a few seconds and is defined in the market maker agreement.

PTP (Precision Time Protocol, IEEE 1588): A network clock synchronisation protocol achieving sub-microsecond accuracy. Used in exchange infrastructure to ensure consistent timestamps across distributed systems.

Public Company: A company whose shares are listed on a stock exchange and available to any investor.

Q

Quote: A two-sided order submitted by a market maker containing both a bid (buy) and an ask (sell) simultaneously. Unlike a regular order, the two legs are linked, what happens to one (a fill, a cancellation) triggers a defined response on the other according to the quote refresh policy.

Quote Leg: One side (bid or ask) of a two-sided market maker quote. Filling one leg triggers the quote refresh policy for the other leg.

Quote Refresh Policy: The rule governing what happens to the surviving leg of a two-sided quote when the other leg is filled. Three main variants: (1) cancel both sides immediately, the most conservative, used by Eurex-style inactivation; (2) cancel only the filled side, leave the other active; (3) leave both active. The choice reflects the market maker's balance between speed of re-entry and risk of holding stale inventory.

Quote Stuffing: A market manipulation practice where a participant rapidly submits and cancels a very large number of orders to consume exchange bandwidth, slow down competitors' systems, and create artificial confusion in market data. A form of denial-of-service abuse in electronic markets.

R

Ratchet: The mechanism in a trailing stop that advances the stop price in the favourable direction but freezes it when the market reverses.

Rate Limiting / Throttling: A control that limits the number of orders a participant can submit per unit time. Prevents denial-of-service conditions from algorithmic misfires.

Re-quoting Obligation: A market maker's contractual requirement to post a fresh two-sided quote within a specified maximum delay after either leg of their previous quote is filled or after MMP fires. Failure to re-quote within this window counts against the presence obligation and may trigger penalties. The window is typically measured in milliseconds to low seconds.

Realized P&L: Profit or loss that has been "locked in" through a trade that reduced an open position. Once realized, it cannot be reversed by subsequent price moves.

Reference Data / Instrument Master: The configuration data describing a tradeable instrument, tick size, price scale, session schedule, contract multiplier, trading status, expiry date, position limits. Read by every exchange component; an error propagates system-wide. Treat it with the same rigour as code: version-controlled, tested, applied atomically.

Regulation NMS: A US Securities and Exchange Commission regulation requiring that equity orders be executed at the best available price across all exchanges (the NBBO). A primary driver of US market fragmentation.

Regulation SHO: A US Securities and Exchange Commission rule governing short sales in equity markets. Key requirements: broker-dealers must have a reasonable belief shares can be borrowed before accepting a short sale order (the "locate" requirement); persistent fail-to-deliver situations must be closed out; short sales

must be marked as "short" in order records. Adopted in 2005 to address concerns about abusive naked short selling.

Replay (Exchange): Replaying a recorded sequence of input messages through the matching engine to reconstruct past market states. Requires the engine to be deterministic.

Reserve Refresh Priority: The queue position rule for iceberg replenishment. Most exchanges place a newly replenished iceberg peak at the back of the queue at its price level, equivalent to a brand-new arrival, rather than preserving the original queue position.

Resting Order / Passive Order: An order that has been accepted by the exchange and is waiting in the book for a counterparty.

Retained Earnings: Profits kept within the business rather than distributed to shareholders. A source of self-funding for company growth.

Rolling Window (MMP): The time period over which fills are counted for Market Maker Protection purposes. Fills older than the window are discarded from the count; only fills within the current window matter. For example, a 1-second rolling window means: "if more than N fills have arrived in any sliding 1-second period, MMP fires."

Round Lot: The standard trading unit for an instrument (commonly 100 shares for US equities), used as the basis for distinguishing a normal order from an *Odd Lot*.

Rule 612 (Sub-Penny Rule): The Regulation NMS provision, adopted 2005, prohibiting exchanges, ECNs, and market makers from displaying or accepting quotations in NMS stocks priced at $1.00 or more in increments finer than $0.01. Stocks priced below $1.00 may be quoted in sub-penny increments. See *Tick Sizes and Fractional Ticks*, Part II.

S

Secondary Market: Where existing securities are traded between investors. The issuing company does not receive proceeds. Stock exchanges are secondary markets.

Secondary Site / Backup Site: A standby data centre that can take over exchange operation if the primary site fails.

Securities Lending / Stock Borrow: The temporary transfer of securities from a lender (typically a long-term investor or custodian) to a borrower (typically a short seller) in exchange for collateral and a lending fee (borrow rate). Short sellers must borrow shares before selling short. The lender retains economic ownership (receives manufactured dividends and can recall the shares) but transfers legal title temporarily. Recall risk, the lender demanding shares back at an inconvenient time, is a key risk for short sellers.

S-1 (Form S-1): The registration statement a US company files with the SEC under the Securities Act of 1933 before an IPO. It contains the prospectus, audited financials, risk factors and the *use of proceeds*. Foreign issuers use the equivalent Form F-1.

SEDOL: A 7-character UK security identifier used in settlement and clearing. Analogous to CUSIP.

Self-Match Prevention (SMP): A mechanism that detects when a participant would trade against their own orders and applies a cancellation policy.

Sequence Number: A monotonically increasing counter attached to events (in a drop copy or message feed) to enable detection of missed events and ordered replay.

Session State Machine: The formal model of an exchange's trading day phases (PRE_OPEN, OPENING_AUCTION, CONTINUOUS, CLOSING_AUCTION, CLOSED) and the permitted transitions between them.

Securities Information Processor (SIP): The consolidated, publicly distributed feed aggregating trade and quote data from every registered US exchange into a single tape. Contrast with a *proprietary direct feed*, which an individual exchange sells separately, typically with lower latency and richer content than the SIP. See *Market Data Economics*, Part IV.

Settlement: The final transfer of securities from seller to buyer and cash from buyer to seller. Legally

completes the trade. In the US, equity settlement currently occurs on T+1.

Settlement Method (Derivatives): Whether a futures or options contract settles by cash payment or physical delivery of the underlying asset. Determines post-expiry processing required by the exchange and clearing house.

Shadow Replication / Shadow Mode: An architectural pattern for high-availability exchange systems. The secondary site receives and processes all input messages in parallel with the primary, maintaining an up-to-date copy of the order book, but does not publish outputs or accept participant connections. If the primary fails, the secondary's in-memory state is already current and it can begin accepting inputs and publishing outputs with minimal failover time. More common in production than true active-active matching.

Share / Stock: A single unit of equity ownership in a company. Each share represents a fractional claim on the company's assets, earnings, and votes.

Slippage: The difference between the expected fill price and the actual average fill price, caused by executing against multiple price levels in the book. Large orders in thin books experience more slippage.

Smart Order Router (SOR): Software that evaluates multiple trading venues and routes orders to achieve the best overall execution, balancing price, fees, available depth, and speed.

SmartNIC / Ultra-Low-Latency NIC: A specialised network interface card designed for trading and other latency-sensitive applications. Unlike commodity Ethernet cards, SmartNICs include an on-card kernel bypass networking stack (such as Solarflare's OpenOnload or Mellanox's RDMA), hardware timestamping of incoming packets to nanosecond precision, and CPU offload for networking tasks. Examples include the AMD Solarflare SFN series and NVIDIA Mellanox ConnectX series, widely deployed in exchange co-location environments.

Snapshot (Persistence): A complete dump of the matching engine's state (all resting orders, positions, key data) taken periodically to reduce recovery time by limiting the amount of write-ahead log that must be replayed.

SPAC (Special Purpose Acquisition Company): A shell company with no operating business that completes its own IPO to raise cash held in trust, with the purpose of later merging with a private operating company to take it public without a traditional IPO process. See *De-SPAC*.

Specialist: NYSE's historical term for the designated firm responsible for maintaining fair and orderly markets in a specific stock, including maintaining the physical order book. The modern equivalent is the Designated Market Maker (DMM).

Speed Bump: A deliberate artificial delay applied to all incoming orders to eliminate the advantage of being marginally faster than other participants. Pioneered by IEX.

Split-Brain: A failure mode in distributed systems where two nodes both believe themselves to be the active primary, leading to conflicting actions.

Spoofing: A market manipulation practice: placing orders with no intention of trading to move prices, then cancelling before execution. Illegal under market abuse regulations.

Spread: The difference between the best ask and the best bid. Market makers earn the spread; participants pay it as the cost of immediate trading.

State Machine: A formal model consisting of a finite set of states and explicit rules about which transitions between states are allowed.

Static Price Collar (Fat-Finger Filter): A pre-trade risk control that rejects any order whose submitted price deviates more than a configured percentage from the last official close price. Designed to catch obvious fat-finger errors, an order at $15.00 on a stock trading at $150.00 is almost certainly a decimal error. Because the reference is the previous close, the static collar is stable throughout the trading day. Contrast with *Dynamic Price Collar*, which tracks the most recent trade instead.

Stop Order: A conditional order that is dormant until the market price reaches the stop price, then converts to another order type.

Stop Price: The trigger price on a stop or stop-limit order. The order remains dormant until the market's last trade price reaches or crosses the stop price, at which point it activates and enters the book (as a market

or limit order, depending on type). A buy stop has a stop price above the current market; a sell stop has a stop price below it.

Stop-Limit Order: A stop order that converts to a limit order when triggered. Provides price protection but may not execute if the market gaps.

Stop-Loss: A sell stop order used to automatically exit a losing long position if the price falls to a specified level.

Surplus: The absolute difference between cumulative buy and sell quantity at the equilibrium price, the quantity that cannot be matched.

Sweeping: The process of an aggressive order filling against resting orders at successive price levels.

T

T+1 / T+2 Settlement: Settlement occurring 1 or 2 business days after the trade date. The US moved from T+3 to T+2 in 2017 and to T+1 in 2024.

T+2 Settlement: Settlement occurring two business days after the trade date. The US standard until 2024; most US equities now settle T+1. T+2 remains the standard in many other markets.

Tick: The minimum price movement for an instrument.

Tick Count / Tick-Based Price: A price represented as an integer number of minimum increments, rather than a decimal. Tick counts are exact; floating-point decimals are not.

Tick Size: The value of one tick, the minimum price increment for an instrument.

Tick Size Pilot Program: A US SEC study (October 2016 – October 2018) that widened the tick size to five cents for roughly 1,200 small-capitalisation stocks to test whether coarser ticks improve market-making economics and liquidity for thinly traded names. Results were mixed: quoted spreads and costs for small trades rose, displayed size at the best prices rose sharply, but the hoped-for improvement in overall liquidity and capital formation did not appear, and no permanent wider-tick regime was adopted. See *Tick Sizes and Fractional Ticks*, Part II.

Tick-to-Trade Latency: The time from when a market data message leaves the exchange to when an order based on that message is received by the exchange. A key performance metric for market makers and HFT firms.

Ticker / Ticker Symbol: The abbreviated code identifying a tradeable instrument (e.g., AAPL for Apple Inc.). Named after the mechanical stock ticker machine that printed trade prices on paper tape using telegraph signals in the late 19th and early 20th centuries.

Ticker Tape: The narrow paper strip printed by the stock ticker machine, showing a continuous stream of trade symbols, prices, and volumes. The source of the term "tick" (minimum price movement) and related concepts.

Time-In-Force (TIF): The attribute specifying how long an order remains valid (DAY, GTC, IOC, FOK, ATO, ATC).

Top-of-Book (Level 1): Market data showing only the best bid price, best ask price, and their quantities. The minimum information needed to understand the current market.

Total Addressable Market (TAM): The total revenue available if a company captured 100% of its market. Often narrowed to the serviceable addressable market (SAM) and the share realistically obtainable (SOM). Popular in growth-company IPO pitches, and notorious for being stretched.

Trade: An executed match between a buy and sell order. Also called a fill or execution.

Trailing Stop: A stop order whose trigger price automatically advances in the favourable direction as the market moves, protected by a fixed trail offset.

TWAP (Time-Weighted Average Price Algorithm): An execution algorithm that distributes a large order evenly across equal time intervals, regardless of trading volume. Simple and predictable, but less adaptive than VWAP. Used when a stock has no reliable volume profile or when the trader wants mechanically

consistent participation. Performance is measured against the average of prices during the execution window, not the overall day's VWAP.

Two-Sided Quote: A market maker's simultaneous bid and ask for the same instrument. Both sides must be live at the same time to fulfil the two-sided quoting obligation. A one-sided quote (only bid or only ask) is a contractual breach of market maker obligations.

U

UUID (Universally Unique Identifier): A 128-bit identifier designed to be unique without central coordination. UUID v4 is randomly generated; UUID v1 incorporates the current time and network address.

Underpricing: Setting an IPO offer price below what the market will pay, visible as a first-day price jump. The gap multiplied by the number of shares sold is called money *left on the table*.

Uncross The single atomic batch execution that occurs at the end of an auction phase. All crossable orders execute at the **equilibrium price** simultaneously.

Unrealized P&L The theoretical profit or loss on an open position if it were closed at the current market price. Changes continuously as the market moves. Becomes realized P&L when the position is closed.

V

Variation Margin: Daily cash payments reflecting mark-to-market gains and losses on open positions. Prevents losses from accumulating to unmanageable levels before settlement.

Venture Capital (VC): Private investment firms that provide funding to early-stage companies in exchange for equity stakes, before those companies are large enough to go public.

VIX (Volatility Index): A real-time index calculated by Cboe from the prices of S&P 500 options, representing the market's expectation of 30-day price volatility in the S&P 500. Commonly called the "fear gauge", when markets are uncertain or falling sharply, participants buy put options for protection, driving up option prices and therefore the VIX. Values above 30 typically indicate high market stress; values below 20 indicate relative calm. Derivatives on the VIX itself trade on Cboe.

VWAP (Volume-Weighted Average Price): The average price of a position weighted by the quantity traded at each price. Used as cost basis for P&L calculation. VWAP $= \frac{\sum(\text{price}_i \times \text{qty}_i)}{\sum \text{qty}_i}$\$

VWAP Algorithm: An execution algorithm that slices a large order and executes it throughout the day in proportion to expected trading volume, more during high-volume periods (open, close) and less during low-volume periods (mid-day). The goal is to achieve an average execution price close to or better than the day's volume-weighted average price. The most widely used institutional execution benchmark. Distinct from *VWAP (Volume-Weighted Average Price)*, which is a measurement; the VWAP algorithm is the execution strategy aimed at matching that measurement.

W

Wash Trading / Wash Trade: A transaction where the same participant is on both sides, generating artificial trading volume with no real change of ownership. Illegal in all major jurisdictions under market manipulation regulations (the EU's Market Abuse Regulation, the US SEC's Rule 10b-5, and others). Also called a wash sale. Sometimes performed accidentally, see *Self-Match Prevention (SMP)*, but when deliberate, it is a form of market abuse used to inflate trading volume figures or create false impressions of liquidity.

Write-Ahead Log (WAL): A persistent, append-only log of all input messages to the matching engine. Written before processing; used for recovery and replay. The source of truth from which all derived state (the in-memory order book) can be reconstructed.

Z

ZeroMQ (ZMQ): A messaging library providing efficient pub/sub, push/pull, and other communication patterns. Commonly used as the message bus in exchange systems. Uses a broker-less PUSH/PULL and PUB/SUB topology. No external message queue server is required.

References

The following sources were drawn upon in writing this document.

[1] Harris, Larry. *Trading and Exchanges: Market Microstructure for Practitioners.* Oxford University Press, 2003. ISBN 978-0-19-514470-3.

The standard academic and practitioner reference for exchange mechanics and market microstructure. Covers order types, price-time priority, the order book, spreads, market makers, adverse selection, and much of the vocabulary in this document. Recommended as the primary companion text for anyone building exchange software.

[2] Hasbrouck, Joel. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading.* Oxford University Press, 2007. ISBN 978-0-19-530164-3.

A rigorous academic treatment of the order book and matching process. Covers the econometrics of market data, information in order flow, and the theory underlying price-time priority. More mathematical than [1] but an authoritative source on how order books actually behave in practice.

[3] Brealey, Richard A., Stewart C. Myers, and Franklin Allen. *Principles of Corporate Finance.* 13th edition. McGraw-Hill Education, 2020. ISBN 978-1-260-01390-0.

The leading corporate finance textbook, used in business schools worldwide. Covers equity, bonds, dividends, IPOs, cost of capital, and the relationship between primary and secondary markets. The source for concepts in Part I of this document (equity, bonds, how companies raise capital).

[4] De la Vega, Joseph. *Confusión de Confusiones* (written 1688). Translated by Hermann Kellenbenz. Harvard University Graduate School of Business Administration, Kress Library of Business and Economics, 1957.

The oldest known book about stock market trading, written in Amsterdam in 1688 by a Spanish merchant named Joseph de la Vega. Describes, in dialogues, the practices of the Amsterdam VOC share market, including resting orders, market making, and speculative trading. A remarkable primary source demonstrating that many modern concepts are centuries old. The Kellenbenz translation includes a scholarly introduction providing historical context.

[5] Petram, Lodewijk. *The World's First Stock Exchange.* Columbia Business School Publishing, 2014. ISBN 978-0-231-16714-9. (Originally published in Dutch as *De wereld's eerste beurs.* Uitgeverij Atlas Contact, 2011.)

The definitive modern history of the Amsterdam Exchange and the VOC share market. Covers the founding of the VOC (1602), the development of the first secondary market in shares, Isaac Le Maire's 1609 short selling campaign against VOC stock, the 1610 Amsterdam edict attempting to ban short selling, and the broader financial innovations that originated in 17th-century Amsterdam. Directly cited for the Le Maire material in this document.

[6] Lewis, Michael. *Flash Boys: A Wall Street Revolt.* W.W. Norton & Company, 2014. ISBN 978-0-393-24413-3.

A narrative account of the founding of IEX by Brad Katsuyama and colleagues, describing the practice of high-frequency trading, its effects on institutional investors, and the design of the speed bump as a countermeasure.

The primary source for the IEX material in the *Speed Bumps* section of Part IV. Accessible to non-specialist readers; recommended as background reading for anyone working on exchange latency and fairness.

[**7**] U.S. Securities and Exchange Commission and Commodity Futures Trading Commission. *Findings Regarding the Market Events of May 6, 2010.* Joint Report. 30 September 2010.

The official regulatory investigation into the 2010 Flash Crash, which saw the Dow Jones Industrial Average fall approximately 1,000 points in minutes on May 6, 2010. Analyses the sequence of events, the role of automated trading, and the breakdown of circuit breakers.

- **Available at:** https://www.sec.gov/news/studies/2010/marketevents-report.pdf

[**8**] U.S. Securities and Exchange Commission. *SEC Adopts T+2 Settlement Cycle for Securities Transactions.* Press Release No. 2017-68. 22 March 2017.

The formal SEC announcement of the rule change shortening US equity settlement from T+3 to T+2, effective September 5, 2017. Contains the rationale for the change and the text of the amended rules.

- **Available at:** https://www.sec.gov/news/press-release/2017-68

[**9**] CME Group. *CME Group to Close Open Outcry Trading in Most Futures Pits.* Press Release. 4 February 2015. Available at: https://www.cmegroup.com

The formal announcement that CME Group would cease open outcry operations in most of its Chicago futures pits, transitioning those products to fully electronic trading. Marks the practical end of an era of physical floor trading that had defined futures markets for over a century.

[**10**] Michie, Ranald C. *The Global Securities Market: A History.* Oxford University Press, 2006. ISBN 978-0-19-928065-3.

A comprehensive scholarly history of securities markets worldwide from the 17th century to the present. Covers the origins of the NYSE and NASDAQ, the development of settlement practices, short selling regulation, and the transition from physical to electronic trading. Cited for historical facts about exchange foundations and the evolution of market structure.

[**11**] Goetzmann, William N., and K. Geert Rouwenhorst, eds. *The Origins of Value: The Financial Innovations that Created Modern Capital Markets.* Oxford University Press, 2005. ISBN 978-0-19-517571-4.

An edited volume of scholarly essays examining the historical development of financial instruments and markets. Includes detailed coverage of the Dutch East India Company, the Amsterdam Exchange, and the origin of tradeable shares, futures, and options. The chapter by Gelderblom and Jonker, "Amsterdam as the Cradle of Modern Futures and Options Trading, 1550–1650," is particularly relevant to the Amsterdam material in the *Language of the Market* section of Part I.

[**12**] Natenberg, S. *Option Volatility and Pricing.* McGraw-Hill, 2nd edition, 1994. ISBN 978-1-55738-486-7.

The practitioner standard. Used as the internal training text at options market making firms for decades. Covers every significant multi-leg strategy, spreads, straddles, strangles, ratio spreads, butterflies, condors, with worked examples from first principles. It explains synthetic positions in depth: synthetic long stock (long call + short put at the same strike), synthetic short stock (short call + long put), synthetic calls and puts, and how put-call parity makes them equivalent. It explains why these constructions exist, delta hedging, volatility trading, arbitrage, not just what they are. The section on volatility spreads directly connects to why exchanges need to support combo orders natively.

[**13**] Hull, John C. *Options, Futures, and Other Derivatives.* Pearson/Prentice Hall, (13th edition, 2021). ISBN 978-0-13-813423-3.

The authoritative academic and practitioner reference for all derivatives. Chapter 12 covers trading strategies involving options in detail, spreads (bull, bear, box, butterfly), combinations (straddles, strangles, strips, straps), with numerical examples and payoff diagrams. The chapters on futures also cover calendar spreads and the basis. Hull grounds everything in put-call parity and arbitrage relationships, which is what makes synthetics work. More mathematical than Natenberg but more complete.

[**14**] *Oliver Gingold and "blue chip" (1923-1924):* The attribution of the first financial use of "blue chip stocks" to Dow Jones employee *Oliver Gingold* circa 1923-1924 is widely repeated in financial dictionaries (including Merriam-Webster). The reference used here comes from **Philadelphia University**, Blue Chip term. Note

the $200/share figure attached to Gingold's original usage is a nominal 1923 dollar amount; it has no bearing on "blue chip" classification today and should not be read as a price threshold (see the discussion in *Before the Exchange*, Part I).

[15] *Edward Calahan and the stock ticker (1867):* Widely documented in the history of telegraphy and American industry. The invention is discussed in Calahan's own patents of the period U.S. Patent No 76,157, Edward A. Calahan,

[16] *Mizuho Securities fat-finger incident (2005):* The December 2005 incident in which a Mizuho trader submitted an order to sell 610,000 shares at ¥1 rather than 1 share at ¥610,000 was extensively reported in media, e.g. FinExtra - Mizuho sues TSE over 'fat finger' trade botch-up, Guardian. It has become a standard illustrative example in financial risk management literature.

- Available at Guardian Newspaper: https://www.theguardian.com/business/2005/dec/09/japan.internationalnews1
- Available at FinExtra: https://www.finextra.com/newsarticle/16081/mizuho-sues-tse-over-fat-finger-trade-botch-up

[17] Lovelock, Mendel, Wright. *An Introduction to the Mathematics of Money.* Springer Verlag, 2007. ISBN 978-0387-34432-4.

A personal favourite, because it quietly dismantles concepts you were certain you understood. Take interest: you think you know what it is? Probably not. At the very least you will come away knowing how the *effective interest rate* on a loan ought to be calculated, and why an effective rate of a few million percent is rather less alarming than it sounds. In the EU this figure must, by law, be quoted on every instalment-based payment, which has inspired businesses to develop a remarkable repertoire of entirely legal "tricks" for arriving at the lowest quotable number.

[18] U.S. Securities and Exchange Commission. *Decimalization of Equity Markets.* Phased implementation order, effective through April 2001. See also Bacidore, Battalio, and Jennings, "Order Submission Strategies, Liquidity Supply, and Trading in Pennies on the New York Stock Exchange," and related Rule 612 (sub-penny) rulemaking under Regulation NMS, 2005.

The formal record of the switch from fractional to decimal quoting in US equities. The source for the decimalization history and Rule 612 material in the *Tick Sizes and Fractional Ticks* section of Part II.

[19] U.S. Securities and Exchange Commission. *Tick Size Pilot Program*, adopted under Rule 612, effective October 2016 to October 2018. Final assessment reports prepared for the SEC by academic and industry researchers.

The primary source for the Tick Size Pilot Program discussion in the *Tick Sizes and Fractional Ticks* section of Part II, including the pilot's design (control group plus three five-cent test groups across roughly 1,200 small-cap securities) and its mixed empirical findings.

- Available at: https://www.finra.org/rules-guidance/key-topics/tick-size-pilot-program
- Assessment report available at: https://www.sec.gov/files/TICK%20PILOT%20ASSESSMENT%20FINAL%20Aug
- Submission of Tick Size Pilot Plan: https://www.govinfo.gov/content/pkg/FR-2014-06-30/pdf/2014-15205.pdf
- SEC Landing Page: https://www.sec.gov/data-research/tick-size-pilot-program/tick-size-pilot-data-resources
- SEC Approval: https://www.sec.gov/files/rules/sro/nms/2015/34-74892-exa.pdf

[20] U.S. House of Representatives, Committee on Financial Services. *Game Stopped: Who Wins and Loses When Short Sellers, Social Media, and Retail Investors Collide.* Majority Staff Report, June 2021.

The primary congressional record of the January 2021 GameStop episode, including the NSCC margin call to Robinhood and other brokers and the resulting trading restrictions. Cited in the *Clearing and Settlement* section of Part III.

- Available at: https://democrats-financialservices.house.gov/uploadedfiles/6.22_hfsc_gs.report_hmsmeetbp.irm.nlr

[21] U.S. Securities and Exchange Commission and U.S. Commodity Futures Trading Commission, civil complaints against Samuel Bankman-Fried and FTX Trading Ltd., December 2022; U.S. Department of Justice, *United States v. Samuel Bankman-Fried*, S.D.N.Y., 2023 conviction.

The primary legal record of the FTX collapse, cited in the *Cryptocurrency and Digital Asset Venues* section of Part IV as an illustration of counterparty risk when matching, custody, and clearing are not institutionally

separated.

- **Wikipedia:** https://en.wikipedia.org/wiki/Trial_of_Sam_Bankman-Fried
- **SEC:** https://www.sec.gov/news/press-release/2022-219
- **CFTC:** https://www.cftc.gov/PressRoom/PressReleases/8638-22

- **DOJ:** https://www.justice.gov/archives/opa/pr/samuel-bankman-fried-sentenced-25-years-his-orchestration-multiple-fraudulent-schemes

Exchanges own technical references

Publicly available technical specifications and documentation provide exact structural logic and algorithmic details for combo and implied/synthetic orders across major electronic matching engines.

CME Globex Matching Engine & Implied Logic

CME Group provides open-access engineering wikis that map out their continuous multi-leg matching models.

- **Implied Orders Functionality:** To read the explicit state constraints, calculation boundaries, and engine generation rules for "Implied-In" and "Implied-Out" routing structures, visit the CME Globex Implied Orders Documentation.
- **Options-Specific Implied Generation:** For data regarding RFQ-triggered timers and second-generation restriction models, see the CME Globex Implied Options Documentation.
- **Algorithmic Match Priorities:** To inspect how implied quantities interact with Pro Rata or FIFO allocation passes when competing with outright resting orders, review the CME Globex Matching Algorithms Reference.
- **Implied Orders:** CME Group, *Client Systems Wiki, Implied Orders* (overview of implied IN / implied OUT functionality, per-product enablement, and implied dissemination on MDP3): https://cmegroupclientsite.atlass (search: "Implied Orders")

Cboe US Options Complex Book Architecture

Cboe details its execution logic for multi-leg derivative allocations, net premium sorting, and atomicity verification.

- **Technical Specifications Library:** Access the complete engineering documentation library, covering binary order entry and auction layout parameters, via the Cboe U.S. Options Technical Specifications.
- **Complex Book Fundamentals:** For an structural layout of multi-leg order handling constraints and leg ratio normalization, review the Cboe Titanium Complex Order Basics.
- **Complex Book Process Guide:** To dive into the complete structural manual—detailing Signed Values on the Complex Book (debit vs. credit), order sorting mechanisms, and Complex Order Auctions (COA)—view the comprehensive Cboe Titanium Complex Book Process Document.

Intercontinental Exchange (ICE) Order Books

ICE manages highly nested energy and financial spreads that utilize synthetic order generation.

- **Implied Prices and Generations:** ICE's FAQ on implied prices, including the Sep/Oct/Nov multi-generation example used in Part II: ICE Futures Implied Prices.
- **Market Data & Connectivity Controls:** For foundational parameters on data fields, systemic electronic metrics, and electronic volume calculations within the ICE infrastructure, refer to the Intercontinental Exchange Connect User Guide.

NYSE Pillar (Equity Trading Platform)

The core documentation for understanding order behavior, priority categories, and amendment rules across the NYSE (New York Stock Exchange) platform

- Pillar is NYSE technology platform which enables members to access NYSE equities and options markets using standard protocols, NYSE Pillar
- The structural and architectural details are broken down in the official NYSE Pillar Functional Differences Guide. This resource outlines how market collars, auction imbalances, and continuous trading interactions operate inside the core matching engine.
- Order types Matrix (XLSX) supported by NYSE

Further Sources Cited in the Text

Implied orders and spreads

- CME Group Client Systems Wiki, Futures Implied Order Matching Priority (real orders before implied orders; tie-breaks between implied sources).
- CME Group Client Systems Wiki, MDP 3.0 – Implied Book (two-deep implied book dissemination).
- CME Group, Treasury Futures Calendar Spreads, 2016 (spread buy/sell convention; leg-price assignment).
- Isaac Carruthers, Inter-Commodity Spreads and Implied Pricing, 2018 (outward rounding of implied prices; finer implied grids for ratio spreads).
- Eurex, Matching principles (synthetic matching) and Passive Liquidity Protection.
- CME press releases on implied functionality for Eurodollar butterflies, 8 March 2004 and 18 October 2004.

Crude oil and negative prices

- CME Group, NYMEX Rulebook Chapter 200: Light Sweet Crude Oil Futures (1,000 barrels; $0.01 tick).
- CFTC, Interim Staff Report on NYMEX WTI Crude Oil Futures Trading around April 20, 2020, November 2020.
- Congressional Research Service, Crude Oil Futures Prices Turn Negative (IN11354), 2020.
- CFTC press release 8432-21, 28 September 2021 (Interactive Brokers: negative-price handling failures).

Market events, enforcement and statistics

- SEC and CFTC, Findings Regarding the Market Events of May 6, 2010, September 2010.
- SEC, In the Matter of Knight Capital Americas LLC, Release No. 34-70694, 16 October 2013.
- SEC Division of Economic and Risk Analysis, Assessment of the Plan to Implement a Tick Size Pilot Program, 2018.
- FINRA, Rule 11892: Clearly Erroneous Transactions in Exchange-Listed Securities.
- NYSE, Closing Auction Fact Sheet; BMLL, Into the Close: Unpacking U.S. Closing Auction Dynamics, 2024.
- Federal Reserve Bank of St. Louis, FRED series SP500 (S&P 500 closing levels, March 2020).
- CME Group, Record annual ADV of 28.1 million contracts in 2025, 5 January 2026.
- Databento, CME matching algorithms explained.
- Eurex, T7 Functional and Interface Overview (Market Maker Protection limits).
- NYSE, Enhancement to Reopening Process after LULD Trading Pauses, 2017 (auction collar widening in 5% steps).
- SEC, Order approving Nasdaq Halt Cross price protections, Release No. 34-102336, 10 February 2025.

- Deutsche Börse, Protective Mechanisms in Auctions (Xetra volatility interruptions); CME Group,
- Deutsche Börse, Price Limits.
- Deutsche Börse, Market Model for the Trading Venue Deutsche Börse Xetra
- Wikipedia, Initial public offering of SpaceX; CoinDesk, SpaceX prices shares at $135 in largest IPO ever, 11 June 2026.
- CME Group, CME Group to Permanently Close Most Open Outcry Trading Pits, 4 May 2021.
- SEC, In the Matter of The NASDAQ Stock Market, LLC (Facebook IPO), Release No. 34-69655, 29 May 2013.
- CCP Global, The Lehman Case (LCH SwapClear close-out of Lehman's portfolio).

IPO pricing and valuation

- Jay R. Ritter, IPO Data, University of Florida (first-day returns, money left on the table, gross spreads).
- Aswath Damodaran, interview with The Acquirer's Multiple, 2024 (pricing versus valuing).
- Investor.gov (SEC), Lockup Agreements.
- SEC, Regulation S-K Item 504: Use of Proceeds.
- European Union, Prospectus Regulation (EU) 2017/1129 and Listing Act Regulation (EU) 2024/2809; Bird & Bird, 2025 (Listing Act overview).
- S&P Dow Jones Indices, S&P U.S. Indices Methodology (IPO seasoning, profitability rule); S&P DJI announcement, 4 June 2026.
- WeWork Companies Inc., Form S-1 registration statement, 14 August 2019.
- News coverage used for deal details: *Forbes* (Facebook, 17 May 2012; WeWork, 22 April 2019 and 7 November 2023), *The Register* (Google auction, 19 August 2004), CNBC (Snowflake, 15 September 2020; Robinhood, 2 July 2021; Rivian, 9 November 2021), CoreWeave press release and Reuters (27 March 2025), *Fortune* (SpaceX S-1, 20 May 2026).

History and etymology

- Christopher Steiner, "How the Nasdaq Got Hacked", *Forbes*, 30 August 2012 (Thomas Peterffy's robot typist); see also his book *Automate This* (2012).
- Online Etymology Dictionary, short ("sell short" attested by 1852).
- Hendrik Bessembinder, "Trade Execution Costs and Market Quality after Decimalization", *Journal of Financial and Quantitative Analysis* 38(4), 2003.
- Richard Smitten, *Jesse Livermore: World's Greatest Stock Trader* (Wiley, 2001).

Notes on Sources Not Cited Inline

Some specific facts in this document are well established in the historical and financial literature but could not be attributed to a single verifiable primary academic document within this text, but rather second hand sources such as NYSE own publications and posts. For example:

Closing auction volume: The share of daily volume executed in US closing auctions varies by year, instrument, and calendar (it spikes on index rebalancing days); the figures in the text are from BMLL's 2024 analysis cited above. See The shifting dynamics of the NYSE Closing Auction

Back-Cover Text

You can write production-grade software in many domains without ever asking what a "spread" really is, why a matching engine must be deterministic, or what happens when a market suddenly halts.

Financial exchanges are not one of those domains.

How a Financial Exchange Works is a practical conceptual guide for developers, architects, product owners, QA engineers, and operations teams who need to understand market structure fast, without wading through rulebooks before they can ship. It explains the language, mechanics, and failure modes of modern markets in plain English: from bids, asks, and order books to auctions, implied matching, risk controls, clearing, settlement, surveillance, and market data economics.

This second edition is significantly expanded and updated, with deeper treatment of exchange microstructure, implementation-relevant architecture, and the real operational realities teams face in live systems.

Inside you will learn:

- How exchanges create price discovery, liquidity, and fairness
- How orders, priority rules, and matching logic produce trades
- Why tick sizes, auctions, and market fragmentation matter in practice
- How risk controls, conformance testing, and post-trade processes keep markets stable
- Where real-world edge cases appear, and what they imply for system design

No hype. No hand-waving. No prerequisite finance degree.

Just the map you need to stop translating jargon and start reasoning clearly about the system you are building.

About the Author

Johan Persson is a software engineer and musician with over 30 years of experience in high-performance, safety-critical systems. He has worked in multiple fields in software design and among them modern financial trading systems, as well as with thruster design for satellite control systems. Johan is passionate about making complex systems understandable and accessible and actually enjoys writing documentation. Johan lives in Sweden with his family and enjoys playing banjo (5-string, of-course!), cycling, wood-working, and when time allows, doing electronics design for analog computers.

As a companion to this book Johan has also written an educational exchange platform `EduMatcher` which implements many of the techniques discussed in this book. While out of necessity it is simplified compared with real exchanges it is still a surprisingly complete multi-process, message-bus based trading platform that can run on a laptop. It uses its own protocol suit with names that is a wink to the fintech industry; ALF, BALF, CALF, RALF, and LALF, with ALF=**AL**most **F**ix, BALF=**B**inary **ALF**, CALF=**C**hannel **ALF**, RALF=**R**econcilliation **ALF**, LALF=**L**ogging **ALF**. It makes an excellent companion to run while reading this book!