Lesson 41: Market Microstructure and HFT

Module 4: Traditional Digital Finance

Digital Finance Course

2025

Learning Objectives

- Understand market microstructure theory and empirical regularities
- Analyze bid-ask spreads and their components
- Examine market maker economics and inventory management
- Evaluate high-frequency trading strategies and market impact
- Assess flash crashes and systemic stability concerns

Source: Financial industry data and regulatory publications

Microstructure Theory Overview

Core Questions:

- How are prices formed in continuous trading?
- What determines bid-ask spreads?
- How does information get incorporated into prices?
- What is the role of market makers and liquidity providers?
- How do trading protocols affect efficiency?

Key Concepts:

- **Price Discovery:** Aggregating dispersed information
- **Liquidity:** Ability to trade without price impact
- **Market Depth:** Volume available at various prices
- **Resilience:** Speed of price recovery after shocks

Trading Costs Framework:

- **Explicit Costs:** Commissions, fees, taxes
- **Implicit Costs:** Spread, impact, opportunity
- **Total Cost:** $TC = \text{Spread} + \text{Impact} + \text{Delay}$

Market Quality Metrics:

- **Efficiency:** Prices reflect available information
- **Liquidity:** Low cost, high volume capacity
- **Transparency:** Order flow and trade visibility
- **Stability:** Resistance to manipulation and crashes
- **Fairness:** Equal access and opportunity

Key insight: microstructure theory continues to evolve with technology advances.

Bid-Ask Spread Fundamentals

Spread Definitions:

- **Quoted Spread:** $S_Q = P_{ask} - P_{bid}$
- **Percent Spread:** $S_{\%} = \frac{P_{ask} - P_{bid}}{P_{mid}} \times 100$
- **Effective Spread:** $S_E = 2|P_{trade} - P_{mid}|$
- **Realized Spread:** $S_R = 2(P_{trade} - P_{mid+5min}) \times D$

where $D = +1$ for buyer-initiated, -1 for seller-initiated

Example:

- Bid: \$99.98, Ask: \$100.02
- Quoted spread: \$0.04 (4 cents)
- Percent spread: 0.04%
- Trade at \$100.01 (price improvement)
- Effective spread: 2 × (\$100.01 - \$100.00) = \$0.02

Spread Components (Stoll 1989):

- **Order Processing:** Fixed costs (40-50%)
- **Inventory Holding:** Risk aversion costs (10-20%)
- **Adverse Selection:** Informed trading (30-50%)

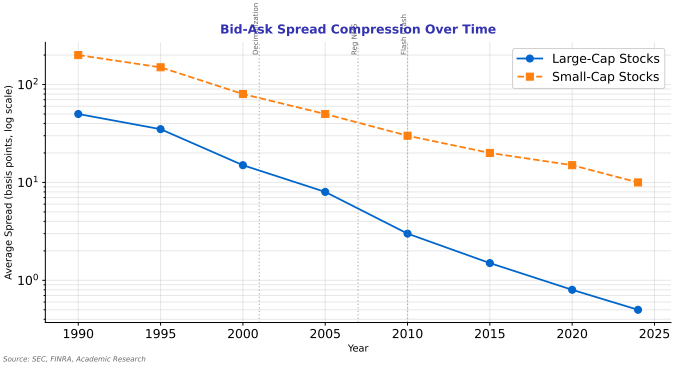
Determinants of Spreads:

- **Volume:** Higher volume → tighter spreads
- **Volatility:** Higher volatility → wider spreads
- **Competition:** More market makers → tighter
- **Tick Size:** Minimum increment constraint
- **Information Asymmetry:** More informed trading → wider

US large-cap spreads: 1-3 bps; small-cap: 10-50 bps (2024)

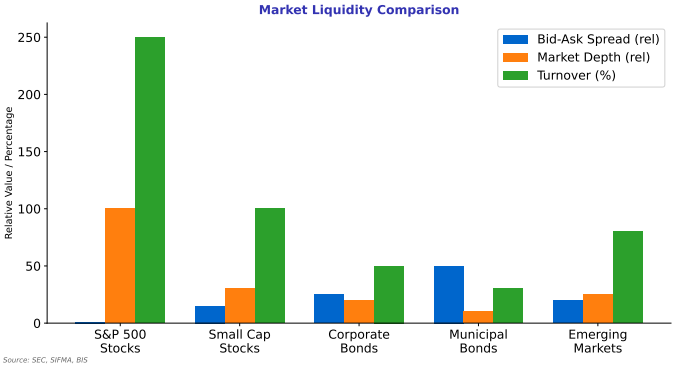
Source: Financial industry data and regulatory publications

Bid-Ask Spread Analysis



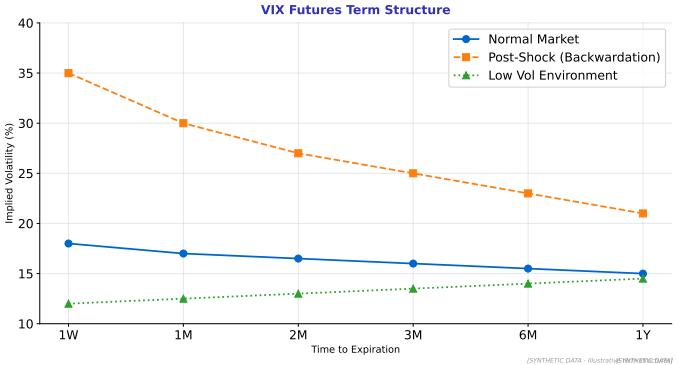
Bid-ask spread measures liquidity cost and information asymmetry.

Liquidity Metrics Across Markets



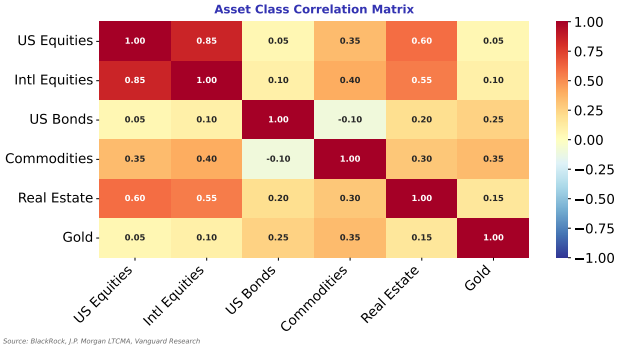
Source: Financial industry data and regulatory publications

Volatility Term Structure



Volatility term structure reveals market expectations across time horizons.

Asset Correlation Matrix



Asset correlations drive portfolio diversification and risk management.

Adverse Selection and Information Models

Glosten-Milgrom Model (1985):

- Sequential trade model with information asymmetry
- Informed traders know true value V
- Uninformed traders are noise/liquidity traders
- Market maker sets bid-ask to break even
- Spread compensates for adverse selection

Key Implications:

- Spread widens with more informed trading
- Price update after each trade (Bayesian learning)
- Bid-ask bounce causes negative autocorrelation
- Larger trades signal more information

Kyle Model (1985):

- Batch auction with strategic informed trader
- Informed trader optimizes profit vs price impact
- Market depth (lambda): $\lambda = \frac{dP}{dQ}$
- Price impact linear in order flow

Kyle's Lambda:

$$\lambda = \frac{\sigma_v}{2\sigma_u}$$

where σ_v = value volatility, σ_u = noise trading volatility

Empirical Evidence:

- Larger trades move prices more (concave impact)
- Block trades have 5-10x impact vs VWAP execution
- Price impact persists 30-60 minutes post-trade

Source: Financial industry data and regulatory publications

Market Maker Role and Economics

Market Maker Functions:

- Continuous two-sided quotes (bid and ask)
- Absorb temporary order imbalances
- Facilitate price discovery
- Reduce search costs for traders
- Profit from bid-ask spread capture

Designated Market Makers (DMM):

- **NYSE:** DMM obligations for assigned stocks
- **Nasdaq:** Competitive market makers (no exclusivity)
- **Obligations:** Maintain fair and orderly markets
- **Benefits:** Informational advantages, rebates

Profitability Sources:

- **Spread Capture:** Buy bid, sell ask
- **Maker Rebates:** 0.15-0.30 cents/share (US equities)
- **Order Flow Internalization:** Payment for order flow (PFOF)
- **Statistical Arbitrage:** Short-term mean reversion

Risks:

- **Inventory Risk:** Directional exposure
- **Adverse Selection:** Losing to informed traders
- **Volatility Spikes:** Widening spreads, reduced depth
- **Technology Failures:** Latency, connectivity issues

Top HFT market makers: Citadel Securities, Virtu, Jane Street, Jump Trading

Source: Financial industry data and regulatory publications

Inventory Management Models

Stoll (1978) Inventory Model:

- Market maker adjusts quotes based on inventory
- High long inventory → lower ask, lower bid
- High short inventory → higher bid, higher ask
- Inventory mean-reversion via asymmetric quotes

Quote Adjustment:

$$P_{bid} = P_{mid} - \frac{S}{2} - \alpha \cdot I$$

$$P_{ask} = P_{mid} + \frac{S}{2} - \alpha \cdot I$$

where I = inventory position, α = inventory aversion

Example:

- Neutral: Bid \$99.98, Ask \$100.02
- Long 10k shares: Bid \$99.96, Ask \$100.00 (skewed to sell)

Avellaneda-Stoikov Model (2008):

- Optimal market making with risk aversion
- Maximizes expected utility of terminal wealth
- Incorporates arrival rates and fill probabilities
- Dynamic spread and mid-price adjustments

Optimal Quotes:

$$\delta^*_{bid} = \frac{1}{\gamma} \ln\left(1 + \frac{\gamma}{\kappa}\right) + \frac{\gamma\sigma^2(T-t)}{2} q$$

$$\delta^*_{ask} = \frac{1}{\gamma} \ln\left(1 + \frac{\gamma}{\kappa}\right) - \frac{\gamma\sigma^2(T-t)}{2} q$$

where γ = risk aversion, κ = order arrival intensity, q = inventory

Avellaneda-Stoikov model optimizes market making with risk aversion and dynamic spread adjustments. [Source: Industry benchmarks]

Payment for Order Flow (PFOF)

PFOF Mechanics:

- Retail broker routes orders to wholesaler
- Wholesaler internalizes (does not send to exchange)
- Wholesaler pays broker 0.1-0.5 cents/share
- Retail order receives NBBO or better (price improvement)
- Wholesaler profits from spread capture + rebates

Major Wholesalers (2024):

- Citadel Securities: 40-45% retail market share
- Virtu Financial: 25-30%
- Jane Street: 10-15%
- Two Sigma Securities: 5-10%

Volume: 40-45% of US equity volume is off-exchange (largely PFOF)

Controversies:

- **Conflict of Interest:** Broker incentive vs best execution
- **Market Segmentation:** Uninformed (retail) vs informed (institutional)
- **Reduced Exchange Volume:** Less price discovery on lit markets
- **GameStop (2021):** PFOF scrutiny after meme stock frenzy

Regulatory Responses:

- SEC considering PFOF ban (2022-2024 discussions)
- MiFID II (EU): PFOF banned for equities (2018)
- Enhanced Rule 606 disclosures (quarterly routing reports)
- Best execution analysis requirements

Retail traders receive average 0.5-1 cent price improvement per share vs NBBO

Understanding the process flow is key to identifying optimization opportunities. [Source: Nilson Report, World Bank 2024]

HFT Definition and Characteristics

HFT Defining Features:

- **Ultra-Low Latency:** Microsecond execution speeds
- **High Order-to-Trade Ratio:** 100:1 to 1000:1
- **High Daily Volume:** Millions of shares/contracts
- **Flat Overnight Positions:** Minimal directional risk
- **Co-Location:** Servers at exchange data centers
- **Direct Market Access:** Sponsored access or memberships

Market Share:

- US equities: 50-55% of volume (down from 60-65% in 2010)
- Futures: 60-70% of volume
- FX: 20-30% of spot volume
- European equities: 30-40% of volume

Technology Infrastructure:

- **Co-Location Costs:** $10k-50k/month per exchange
- **Market Data Feeds:** Direct feeds vs consolidated (SIP)
- **FPGA Acceleration:** Hardware-based order processing
- **Microwave Networks:** Chicago-NYC in 4.1 milliseconds
- **Kernel Bypass:** User-space networking (10-100x faster)

Latency Benchmarks:

- Matching engine: 10-50 microseconds
- Co-located order entry: 50-200 microseconds
- Cross-venue arbitrage window: 100-500 microseconds
- Human reaction time: 200,000 microseconds (200 ms)

Clear definitions are essential for understanding complex technical concepts. [Source: DeFi Llama, DeFi Pulse 2024]

HFT Strategies

Liquidity Strategies:

- **Market Making:** Spread capture + rebates (40-50% of HFT)
- **Stat Arb:** Mean-reversion, pairs trading at microsecond horizons

Speed-Based:

- **Latency Arb:** Exploit slow price updates, "arms race"
- **Event Arb:** News parsing (NFP, FOMC, earnings) in microseconds

Structural Strategies:

- **Order Anticipation:** Detect large orders via book patterns
- **Index Arb:** ETF vs basket, futures vs cash basis

Key Characteristics:

- Hold time: microseconds to minutes
- Requires fastest infrastructure
- 50%+ of US equity volume

HFT strategies exploit speed advantages and market structure inefficiencies. [Source: SEC Market Structure Reports 2024]

HFT Market Impact Debate

Positive Contributions:

- **Tighter Spreads:** Increased competition narrows bid-ask
- **Increased Liquidity:** Higher quoted depth
- **Faster Price Discovery:** Information incorporated quicker
- **Lower Trading Costs:** Spreads down 50-70% since 2000
- **Cross-Market Integration:** Arbitrage keeps markets aligned

Empirical Evidence (Brogaard et al., 2014):

- HFT trades align with permanent price changes
- Net provision of liquidity (market making)
- Faster incorporation of public information
- No evidence of systematic predatory behavior

Criticisms and Concerns:

- **Adverse Selection:** Institutional orders picked off
- **Phantom Liquidity:** Quotes vanish in stress periods
- **Flash Crashes:** Amplify volatility (e.g., 2010)
- **Unfair Advantage:** Speed and co-location benefits
- **Socially Wasteful:** Arms race with low social value

Market Stability Risks:

- Simultaneous withdrawal during stress
- Correlated algorithmic behavior
- Feedback loops and cascades
- Fragility due to speed dependencies

"HFT is like GPS: improves efficiency but introduces new failure modes" – SEC Commissioner (2014)

Source: Financial industry data and regulatory publications

May 6, 2010 Flash Crash

Event Timeline:

- **14:32 ET:** Large mutual fund executes $4.1B E-Mini sell program
- **14:41-14:45:** Dow drops 600 points in 5 minutes
- **14:45:** Markets recover 70% of losses in minutes
- **End of Day:** Dow closes down 348 points

Trigger Mechanism:

- Aggressive VWAP algorithm floods market
- HFT firms initially absorb selling
- Hot potato: HFTs trade among themselves
- Liquidity withdrawal as inventory limits hit
- Prices collapse due to lack of buyers

Key Findings (CFTC-SEC Report):

- E-Mini futures led equities down
- Cross-market arbitrage transmitted stress
- Stub quotes executed (e.g., Accenture to $0.01)
- 20,000+ trades broken (clearly erroneous)
- HFT exacerbated but did not cause crash

Regulatory Responses:

- **Limit Up-Limit Down (2012):** Single-stock circuit breakers
- **Clearly Erroneous Trades:** Standardized break criteria
- **Reg SCI (2015):** Systems compliance for critical infrastructure
- **Market-Wide Breakers:** 7%, 13%, 20% thresholds

Source: Financial industry data and regulatory publications

Other Notable Flash Events

October 15, 2014 Treasury Flash Rally:

- 10-year yield drops 37 bps in 12 minutes
- Largest intraday move in decades
- No fundamental news trigger
- Joint Staff Report: HFT amplified volatility
- Highlighted fragility in Treasury market structure

August 24, 2015 ETF Flash Crash:

- 1100+ trading halts in first 36 minutes
- 20% of ETFs trade 10%+ away from NAV
- LULD breakers overwhelmed by volume
- Exposing ETF market making fragility

GBP Flash Crash (October 7, 2016):

- Sterling drops 9% vs USD in minutes (Asian hours)
- Thin liquidity + algorithmic selling spiral
- Recovered within 30 minutes
- Highlighted FX market vulnerabilities

Common Patterns:

- Initial shock (fundamental or algorithmic)
- Liquidity provider withdrawal
- Cascading sell orders (stop losses, algos)
- Feedback loop amplification
- Recovery once human oversight intervenes

VPIN (toxicity indicator) spiked to 0.98 before 2010 Flash Crash (normal ¡ 0.5)

Source: Financial industry data and regulatory publications

Market Stability Mechanisms

Circuit Breakers (US Equities):

- **LULD Bands:** 5-20% depending on tier and time
- **Tier 1:** S&P 500, Russell 1000 (5% bands)
- **Tier 2:** Other NMS stocks (10% bands)
- **Trading Halt:** 5-minute pause if limit breached
- **Market-Wide:** 7%, 13%, 20% S&P 500 declines

Clearly Erroneous Trades:

- Numerical thresholds for breaking trades
- $0-$25 stocks: 10% from reference price
- $25-$50: 5%, Over $50: 3%
- Must be reported within 30 minutes
- Exchange decision (not automatic)

Kill Switch Requirements (MiFID II):

- Ability to cancel all orders in under 2 seconds
- Mandatory for algorithmic traders
- Pre-trade risk controls on parameters
- Post-trade monitoring and alerts

Volatility Auctions:

- European markets use volatility interruptions
- 2-5 minute random auction when thresholds hit
- Allows human intervention and price discovery
- Less disruptive than hard halts

Speed Bumps:

- IEX: 350-microsecond delay on all orders
- Prevents latency arbitrage strategies
- Controversial: reduces efficiency vs stability

Source: Financial industry data and regulatory publications

Intraday Patterns and Anomalies

U-Shaped Volume Pattern:

- High volume at open (09:30-10:00 ET)
- Low volume midday (11:00-14:00)
- High volume at close (15:30-16:00)
- Opening 30 min: 15-20% of daily volume
- Closing 30 min: 20-25% of daily volume

U-Shaped Volatility:

- Spreads widest at open (information asymmetry)
- Tightest spreads midday (steady state)
- Widening into close (position squaring)

Bid-Ask Bounce:

- Trades alternate between bid and ask
- Induces negative autocorrelation in returns
- Roll (1984) model: Spread $= 2\sqrt{-\text{Cov}(r_t, r_{t-1})}$
- Empirically: first-order autocorr = -0.05 to -0.15

Price Impact Asymmetry:

- Buy orders have larger impact than sells
- More pronounced for small-cap stocks
- Attributed to short-sale constraints
- Temporary component decays exponentially

Opening cross (NYSE, Nasdaq): 5-10% of daily volume in single batch auction

Source: Financial industry data and regulatory publications

Tick Size and Market Quality

US Tick Size Evolution:

- Pre-2001: 1/16 dollar ($0.0625)
- 2001: Decimalization to $0.01
- Result: Spreads compressed 30-50%
- But: Depth per price level declined
- Trade-off: Tighter spreads vs lower depth

Tick Size Pilot (2016-2018):

- SEC mandated $0.05 tick for 1200 small-cap stocks
- Goal: Improve liquidity and market making economics
- Results: Wider spreads, lower volume, minimal IPO impact
- Pilot ended; no permanent changes adopted

Optimal Tick Size Theory:

- Too small: Excessive queue jumping, minimal profit
- Too large: Constrained price discovery, wide spreads
- Optimal tick proportional to stock price
- Harris (1994): Tick-to-price ratio $= 0.1\text{-}0.5\%$

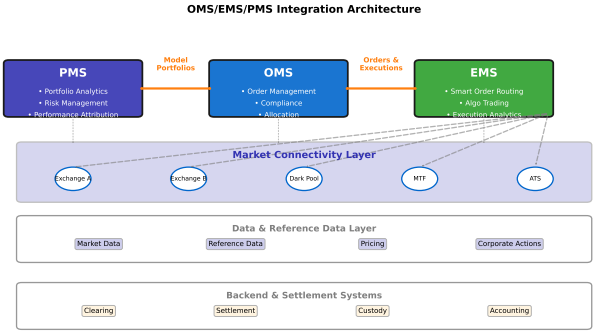
Empirical Regularities:

- Quotes cluster at round numbers
- Spreads often exactly 1 tick (minimum)
- 60-70% of stocks trade at 1-cent spread
- Sub-penny trading banned for most stocks (Reg NMS)

European markets: varying tick sizes by price range (MiFID II tick size regime)

Source: Financial industry data and regulatory publications

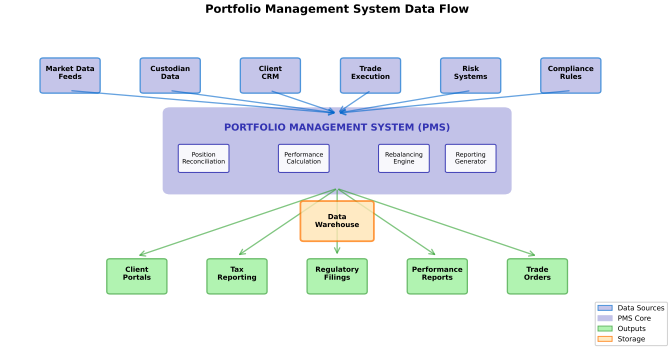
OMS, EMS, and PMS Integration



Source: FIX Trading Community, Celent, Charles River

Modern trading ecosystems integrate portfolio management, order management, and execution.

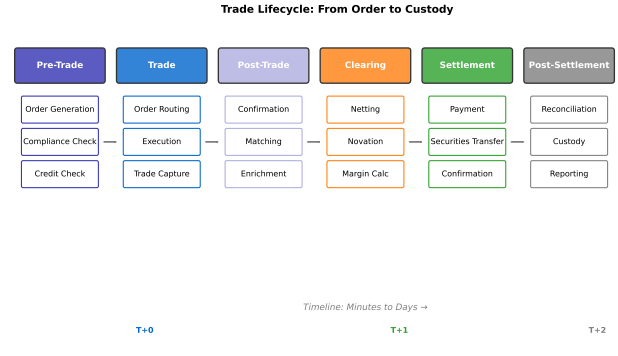
Portfolio Management System Data Flow



Source: Standard PMS architecture (CFA, CRD, Aladdin)

PMS systems aggregate positions, performance, and risk across multiple accounts.

Trade Lifecycle from Order to Settlement



The trade lifecycle involves multiple stages and systems from execution to settlement.

Summary and Key Takeaways

Microstructure Foundations:

- Bid-ask spreads compensate for order processing, inventory, and adverse selection
- Market makers provide liquidity and earn spread via inventory management
- Information asymmetry drives adverse selection costs
- PFOF internalizes retail flow (40-45% of US volume)

High-Frequency Trading:

- 50-55% of US equity volume (microsecond speeds)
- Strategies: market making, stat arb, latency arb
- Tighter spreads but phantom liquidity concerns
- Regulatory focus on stability and fairness

Flash Crashes and Stability:

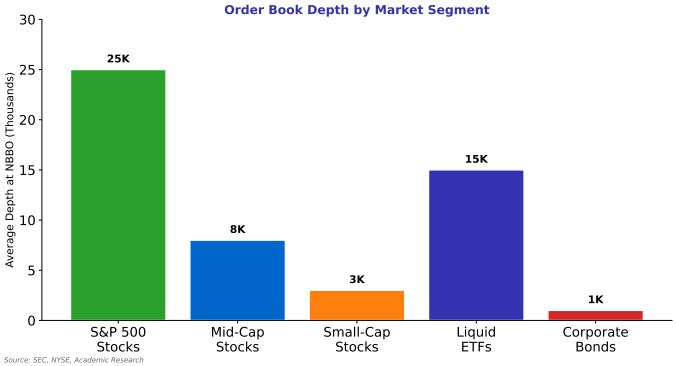
- May 6, 2010: 600-point drop in 5 minutes
- Liquidity withdrawal amplifies shocks
- LULD circuit breakers now standard
- Kill switches and risk controls mandatory (MiFID II, Reg SCI)

Market Quality:

- U-shaped volume and volatility patterns
- Decimalization tightened spreads (2001)
- Tick size pilot failed to improve small-cap liquidity
- Trade-off: efficiency vs stability in design

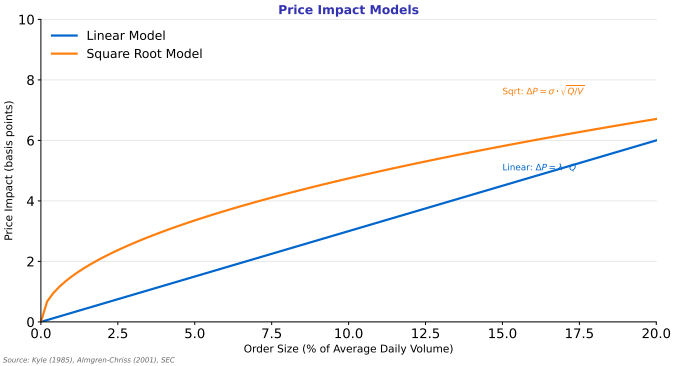
Data sources: McKinsey, Gartner 2024

Order Book Depth Analysis



Depth varies by asset class and market conditions.

Price Impact Function



Impact increases non-linearly with order size.