L06: Financial Markets & Trading Infrastructure

Extended Slides – BSc Digital Finance Course

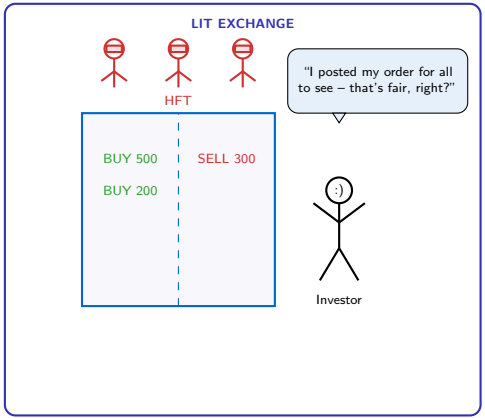
Digital Finance

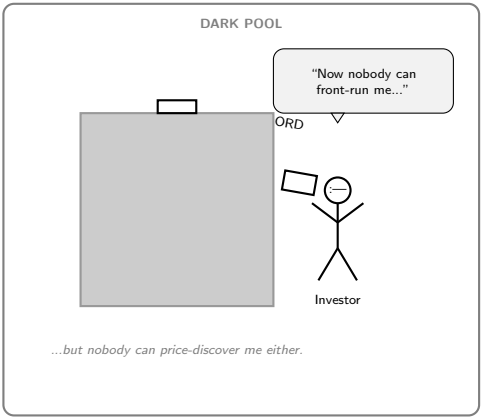
What Will You Be Able to Do After This Lecture?

By the end of this extended lecture, you will be able to:

1. Formally model the order book as a queuing system and derive the spread-depth-impact relationship
2. Test the Efficient Market Hypothesis using return autocorrelation and variance ratio tests in Python
3. Analyze HFT latency advantages using arrival-rate models
4. Model CCP netting efficiency as a graph problem
5. Implement a VWAP execution algorithm and measure TCA performance
6. Evaluate tokenized securities settlement economics

Six objectives: formal models (1, 4), Python implementation (2, 5), and applied evaluation (3, 6). This lecture combines theory with working code.





Transparency helps everyone – except when it helps the wrong people more.

How Does the Order Book Turn Individual Intentions into a Market Price?

The order book is a **queuing system** where limit orders arrive and depart stochastically.

Order arrival process. Let $\lambda_b(p)$ and $\lambda_a(p)$ denote bid and ask arrival rates at price p:

$$P_b^{(1)} = \max\{p : \lambda_b(p) > 0\}, \qquad P_a^{(1)} = \min\{p : \lambda_a(p) > 0\} \tag{1}$$

Bid-ask spread:

$$s = P_a^{(1)} - P_b^{(1)} \tag{2}$$

Kyle's lambda (permanent price impact per unit of net order flow):

$$\Delta P = \lambda \cdot \text{sign}(V) \cdot \sqrt{|V|} \tag{3}$$

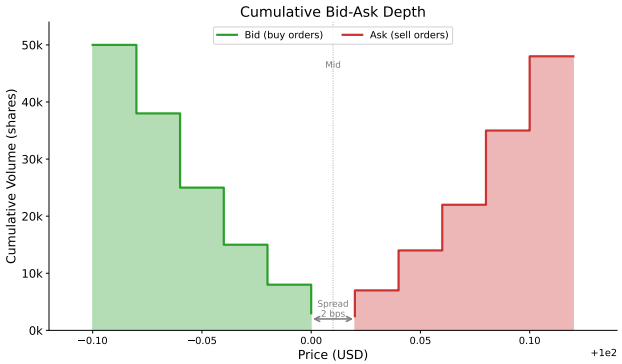
where V is signed volume and λ is the Kyle impact coefficient.

Glosten-Milgrom spread decomposition:

$$s = \underbrace{c_{\text{order}}}_{\text{processing}} + \underbrace{c_{\text{inventory}}}_{\text{risk}} + \underbrace{c_{\text{adverse}}}_{\text{informed traders}} + \underbrace{\pi}_{\text{profit}} \tag{4}$$

The spread is not a single cost – it is four costs stacked. Adverse selection dominates in liquid markets.

What Does the Full Depth of Supply and Demand Look Like?



Reading the depth chart

- The **bid curve** (green, left) shows cumulative buy volume at each price level; the **ask curve** (red, right) shows cumulative sell volume
- A **steep curve** near the midpoint signals high liquidity – many orders are clustered around the best price
- A **flat curve** signals thin liquidity – a large order would sweep through many price levels, causing significant impact
- The **gap** between the two curves at the midpoint is the bid-ask spread

Swiss context: SMI blue-chip names (Nestle, Novartis) typically show steep depth curves; Swiss mid-caps often have thinner books with wider spreads.

A depth chart is the order book made visual – steep curves mean cheap execution, flat curves mean expensive execution.

Can You Simulate Price Discovery in 25 Lines of Python?

```python
import numpy as np
from collections import defaultdict

class OrderBook:
    def __init__(self):
        self.bids = defaultdict(float) # price -> vol
        self.asks = defaultdict(float)
    def limit_order(self, side, price, volume):
        book = self.bids if side=='buy' else self.asks
        book[price] += volume
    def market_order(self, side, volume):
        book = self.asks if side=='buy' else self.bids
        prices = sorted(book, reverse=(side=='sell'))
        filled, cost = 0, 0.0
        for p in prices:
            take = min(volume - filled, book[p])
            cost += take * p
            book[p] -= take
            filled += take
            if book[p] <= 0: del book[p]
            if filled >= volume: break
        return cost / filled if filled else 0
    @property
    def spread(self):
        if self.asks and self.bids:
            return min(self.asks) - max(self.bids)
        return float('inf')
```

What the code does

- Maintains two dictionaries: `bids` and `asks`, each mapping a price to cumulative volume
- **Limit orders** add volume at a specific price level – they provide liquidity
- **Market orders** walk through the book from best price outward, consuming liquidity at each level until filled
- The `spread` property computes the difference between the best ask and best bid

Key insight: Market orders execute at multiple price levels ("walking the book"). The deeper the book, the less the price moves.

Twenty-five lines: a working order book with limit orders, market orders, and a spread property.

(c) Joerg Osterrieder 2026

How Much Does Your Order Move the Market – and Can You Predict It?

The **Almgren-Chriss** framework decomposes execution cost into three components:

Square-root market impact (empirically validated across asset classes):

$$\text{Market Impact} = \sigma \cdot \sqrt{\frac{V}{V_{\text{ADV}}}} \tag{5}$$

where σ is daily volatility, V is order size, and V_{ADV} is average daily volume.

Total execution cost:

$$C_{\text{total}} = \underbrace{\frac{s}{2}}_{\text{half-spread}} + \underbrace{\sigma\sqrt{\frac{V}{V_{\text{ADV}}}}}_{\text{temporary impact}} + \underbrace{\gamma \cdot \frac{V}{V_{\text{ADV}}}}_{\text{permanent impact}} \tag{6}$$

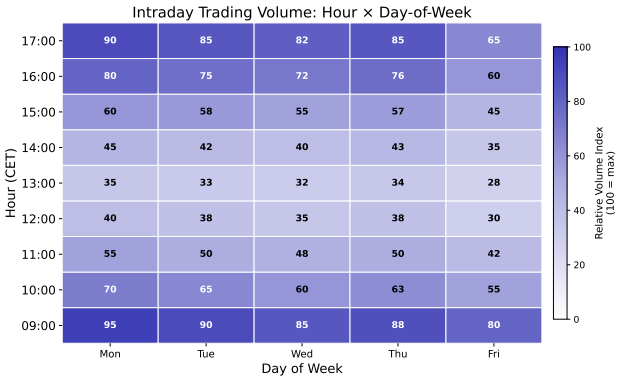
Optimal execution tradeoff: Execute faster $\Rightarrow$ higher market impact but lower timing risk. Execute slower $\Rightarrow$ lower impact but higher risk of price drift.

Numerical example: For $\sigma = 2\%$, $V/V_{\text{ADV}} = 5\%$, impact $\approx 2\% \times \sqrt{0.05} \approx 45$ bps. Doubling the order to 10% ADV yields $2\% \times \sqrt{0.10} \approx 63$ bps – a 41% increase, not 100%.

Market impact follows a square-root law: doubling order size increases impact by 41 percent, not 100 percent.

(c) Joerg Osterrieder 2026

When Is the Market Most Liquid – and When Should You Avoid Trading?



Reading the heatmap

- The **U-shape** within each day is the most robust pattern: volume is highest at the open and close, lowest at midday
- **Monday mornings** carry accumulated information from the weekend – spreads widen and impact costs rise
- **Friday afternoons** show lower volume as risk limits tighten before the weekend
- The **European-US overlap** (14:30–17:30 CET) is the deepest liquidity window for cross-listed securities

Implication for execution: VWAP algorithms exploit the U-shape by concentrating child orders at high-volume periods, reducing per-share impact cost.

The intraday U-shape is not noise – it reflects the information cycle: opening incorporates overnight news, closing locks in positions.

How Do You Statistically Test Whether a Market Is Efficient?

The **Efficient Market Hypothesis** implies that prices follow a random walk. Testable predictions:

Random walk model:

$$P_t = P_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2) \tag{7}$$

Return autocorrelation (should be zero under EMH):

$$\rho_k = \frac{\text{Cov}(r_t, r_{t-k})}{\text{Var}(r_t)} \approx 0 \quad \forall\, k \geq 1 \tag{8}$$

Lo-MacKinlay variance ratio test:

$$VR(q) = \frac{\text{Var}(r_t + r_{t+1} + \cdots + r_{t+q-1})}{q \cdot \text{Var}(r_t)} \tag{9}$$

Under the null (random walk), $VR(q) = 1$. If $VR(q) > 1$, returns are positively autocorrelated (momentum); if $VR(q) < 1$, returns mean-revert.

Test statistic:

$$z = \frac{VR(q) - 1}{\sqrt{\frac{2(2q-1)(q-1)}{3qn}}} \xrightarrow{d} N(0,1) \tag{10}$$

The variance ratio test converts EMH from a philosophical claim into a falsifiable statistical hypothesis.

Can You Test Market Efficiency in 20 Lines of Python?

```python
import numpy as np

def variance_ratio_test(returns, q=5):
    """Lo-MacKinlay variance ratio test."""
    n = len(returns)
    mu = np.mean(returns)
    var_1 = np.sum((returns - mu)**2) / (n - 1)
    r_q = np.array([
        np.sum(returns[i:i+q])
        for i in range(n - q + 1)
    ])
    var_q = np.sum((r_q - q*mu)**2) / (n - q)
    vr = var_q / (q * var_1)
    se = np.sqrt(2*(2*q-1)*(q-1) / (3*q*n))
    z = (vr - 1) / se
    return vr, z

np.random.seed(42)
r = np.random.normal(0.0004, 0.015, 1000)
vr, z = variance_ratio_test(r, q=5)
print(f"VR(5) = {vr:.4f}, z = {z:.2f}")
print("Reject EMH" if abs(z)>1.96 else "Fail to reject")
```

How the test works

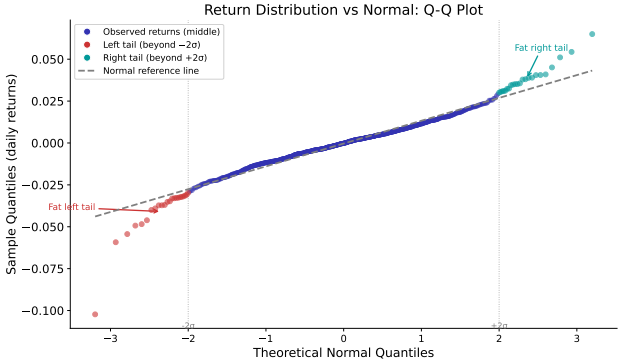
- Computes the variance of q-period returns and compares it to q times the variance of 1-period returns
- Under a random walk, these are equal, so $VR(q) = 1$
- The z-statistic tests whether the deviation from 1 is statistically significant
- With simulated i.i.d. returns, the test correctly fails to reject EMH

Real-world results:

- Most developed equity markets: $VR \approx 1$ at daily frequency
- Emerging markets: $VR > 1$ (momentum, serial correlation)
- Intraday data: $VR \neq 1$ due to microstructure effects

Twenty lines of Python convert a century-old debate about market efficiency into a testable, reproducible hypothesis.

Are Stock Returns Really Normally Distributed – or Is That a Dangerous Assumption?



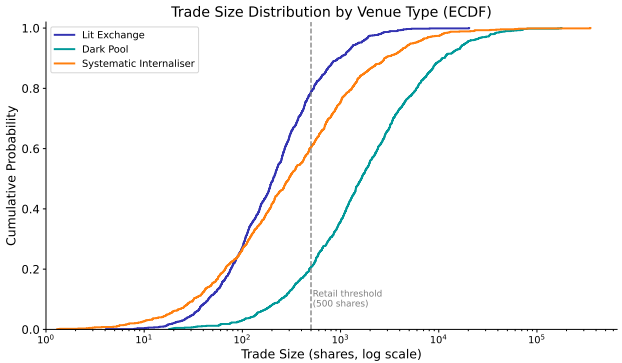
Reading the Q-Q plot

- If returns were normal, all points would lie on the diagonal line
- **Fat tails** appear as upward deviations at the right end and downward deviations at the left end
- Real equity returns exhibit excess kurtosis of $\sim$6–9, meaning extreme events are 10–100x more likely than a Gaussian model predicts
- The **October 1987** crash (-22% in one day) is a 25-σ event under normality – effectively impossible

Why it matters: VaR models, option pricing (Black-Scholes), and margin calculations all assume normality. Fat tails mean these models systematically underestimate tail risk.

The Q-Q plot is a diagnostic: if the tails deviate, your risk model is wrong – and it is wrong in the direction that matters most.

What Fraction of Trades Are Truly Large – and Why Does the Distribution Matter?



Reading the ECDF

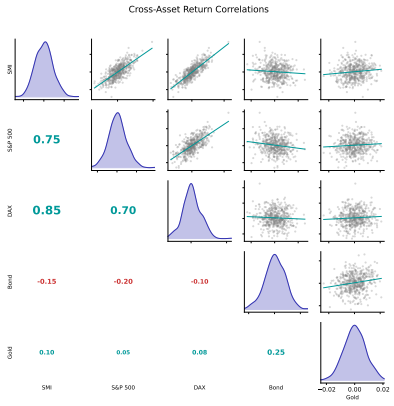
- The empirical CDF shows what fraction of trades fall below a given size
- In **lit markets**, roughly 80% of trades are under 500 shares – the market is dominated by small retail and algorithmic orders
- In **dark pools**, the curve shifts right: the 90th percentile trade size is 5–10x larger than in lit venues
- This confirms the economic rationale: dark pools exist to execute large institutional orders without information leakage

Regulatory implication: If dark pools attract only large orders, they complement lit markets. If they attract small orders too, they fragment price discovery.

The ECDF separates fact from narrative: most trades are small, but most volume comes from the large orders that dark pools are designed to protect.

(c) Joerg Osterrieder 2026

Do Markets Move Together – and When Does Diversification Fail?



Reading the scatter matrix

- Each panel shows the joint distribution of daily returns for two markets
- In **calm periods**, correlations between European indices (SMI, DAX, CAC) range from 0.5–0.7 – sufficient for diversification benefits
- In **crisis periods** (2008, 2020), correlations converge toward 0.90–0.95, destroying diversification precisely when it is needed most
- This is the "correlation breakdown" problem: correlations are not constant – they spike in stress

Swiss context: SMI-DAX correlation reaches $\sim$0.95 in stress, making geographic diversification within Europe ineffective during crises.

Correlations are unstable: they are lowest when you need diversification least and highest when you need it most.

How Much Is One Microsecond Worth in the Latency Arms Race?

An HFT firm profits when it detects and acts on a price discrepancy **before** competitors.

Arrival-rate advantage model. Let α be the rate of arbitrage opportunities per second, $\bar{\pi}$ the average profit per opportunity, and $P_{\text{win}}(\Delta t)$ the probability of capturing it given latency advantage Δt:

$$\Pi_{\text{fast}} = \alpha \cdot \bar{\pi} \cdot P_{\text{win}}(\Delta t) \cdot T \tag{11}$$

where T is the number of trading seconds per year ($\sim 2.3 \times 10^7$).

Value of one microsecond:

$$\frac{\partial \Pi}{\partial(\Delta t)} = \alpha \cdot \bar{\pi} \cdot \frac{\partial P_{\text{win}}}{\partial(\Delta t)} \cdot T \tag{12}$$

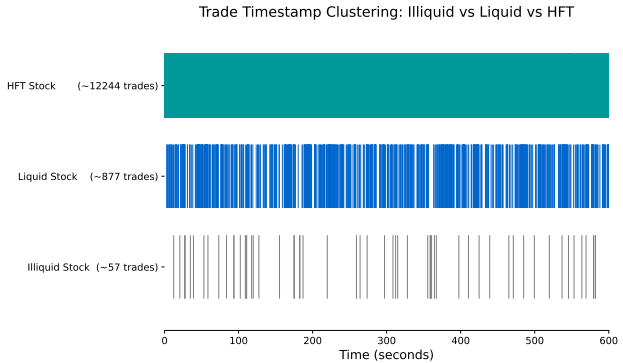
Numerical example: With $\alpha = 100$ opportunities/sec, $\bar{\pi} = \$0.01$ per event, and $\partial P_{\text{win}}/\partial(\Delta t) = 0.001$ per μs, one microsecond is worth:

$$100 \times 0.01 \times 0.001 \times 2.3 \times 10^7 \approx \$23{,}000/\text{year}$$

Across 200+ correlated instruments: \$5–50M per year per microsecond.

One microsecond of latency advantage is worth \$5–50 million per year – which explains billion-dollar investments in fiber and microwave links.

Can You See the Speed of Markets – and the Gaps Between Trades?



Reading the eventplot

- Each vertical tick marks one trade; the horizontal axis is time
- **Dense clusters** ("barcodes") indicate bursts of activity – often triggered by news, order book events, or algorithmic cascade
- **Gaps** between clusters reveal periods of low activity where market makers widen spreads
- The clustering pattern is not random: it follows a **Hawkes process** – each trade increases the probability of the next trade occurring soon

Implication: Trade arrival is self-exciting. Algorithms that detect burst onsets can position ahead of the cascade.

Trade timestamps reveal the rhythm of the market – bursts and silences that no summary statistic can capture.

Can You Build a VWAP Execution Algorithm That Beats the Benchmark?

```python
import numpy as np

def vwap_schedule(total_shares, hist_vol):
    """VWAP schedule from historical volume."""
    frac = hist_vol / hist_vol.sum()
    return np.round(total_shares * frac).astype(int)

def execute_vwap(schedule, prices, slip_bps=2):
    """Simulate VWAP execution with impact."""
    total_cost, total_qty = 0.0, 0
    for qty, price in zip(schedule, prices):
        if qty > 0:
            impact = slip_bps*1e-4 * np.sqrt(qty/250)
            fill = price * (1 + impact)
            total_cost += fill * qty
            total_qty += qty
    algo_vwap = total_cost / total_qty
    mkt_vwap = np.average(prices, weights=schedule)
    return algo_vwap, mkt_vwap, algo_vwap - mkt_vwap

vol = np.array([15,10,8,7,6,7,12,20])
px = np.array([100.,100.2,100.1,99.8,99.9,100.3,100.5,100.4])
s = vwap_schedule(50000, vol)
a, m, slip = execute_vwap(s, px)
print(f"Algo: {a:.4f} Mkt: {m:.4f} Slip: {slip:.4f}")
```

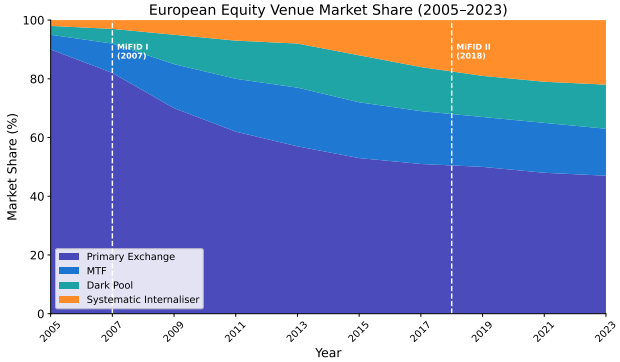
How VWAP execution works

- **Step 1:** Build a volume profile from historical data – how much of the day's volume trades in each interval
- **Step 2:** Allocate shares proportionally to each interval's historical volume share
- **Step 3:** Execute each child order with a square-root impact model
- **Benchmark:** Compare the algorithm's average fill price to the market VWAP

Why VWAP? It is the most common institutional benchmark because it measures whether you traded "with the market" – neither rushing nor lagging.

VWAP is the most widely used execution benchmark: if you match it, you traded at the market's average price – no better, no worse.

Where Have All the Trades Gone – and Why Are They Spread Across So Many Venues?



Reading the stacked area

- The **primary exchange** (bottom layer) shrank from ∼90% of volume in the early 2000s to roughly 50% today
- **Dark pools** (middle layers) grew steadily after MiFID I (2007) opened competition among execution venues
- **Systematic internalisers** capture retail order flow, especially in Europe post-MiFID II
- The total still sums to 100% – volume did not disappear, it fragmented

MiFID I / II effects: MiFID I broke the exchange monopoly. MiFID II tried to push dark volume back to lit venues via the double volume cap. The result: some dark volume migrated to periodic auctions instead.

Market fragmentation is not a bug – it is the intended consequence of competition policy. The question is whether fragmentation helps or harms price discovery.

(c) Joerg Osterrieder 2026

How Do Dark Pools Match Orders Without Revealing Prices?

Dark pools use the **midpoint** of the lit market's National Best Bid and Offer (NBBO) as the reference price.

Midpoint matching:

$$P_{\text{dark}} = \frac{P_b^{(1)} + P_a^{(1)}}{2} \tag{13}$$

Fill probability. Let Q_b and Q_a be the queue sizes on bid and ask sides in the dark pool:

$$P(\text{fill}) = \min\left(\frac{Q_a}{Q_b}, 1\right) \quad \text{for a buy order} \tag{14}$$

Information leakage cost. If a fraction ϕ of dark pool participants are informed:

$$C_{\text{leak}} = \phi \cdot \sigma \cdot \sqrt{\frac{\tau}{\Delta t}} \tag{15}$$

where σ is volatility, τ is the information half-life, and Δt is the matching interval.

Decision rule: Route to the dark pool when:

$$\underbrace{P(\text{fill}) \cdot \frac{s}{2}}_{\text{expected savings}} > \underbrace{C_{\text{leak}} + C_{\text{delay}}}_{\text{expected costs}} \tag{16}$$

Dark pools save the half-spread but risk information leakage – the routing decision is a cost-benefit calculation, not a philosophical choice.

(c) Joerg Osterrieder 2026

How Does a CCP Turn 1,000 Bilateral Exposures into 50 Net Positions?

Bilateral gross exposure among n participants:

$$E_{\text{bilateral}} = \sum_{i=1}^{n} \sum_{j \neq i} |G_{ij}| \tag{17}$$

where G_{ij} is the gross obligation from participant i to participant j.

Multilateral netting via CCP. The CCP computes each participant's net position:

$$N_i = \sum_{j \neq i} G_{ij} - \sum_{j \neq i} G_{ji} \tag{18}$$

$$E_{\text{CCP}} = \sum_{i=1}^{n} |N_i| \tag{19}$$

Netting efficiency:

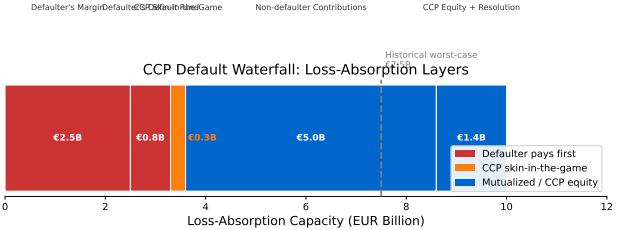
$$\eta = 1 - \frac{E_{\text{CCP}}}{E_{\text{bilateral}}} \qquad \text{(typically 95\%–99\%)} \tag{20}$$

Collateral savings: With $\eta = 98\%$, a market with \$1 trillion in gross obligations reduces to \$20 billion in net exposures, freeing \$980 billion in collateral for other uses.

CCP netting is the single most powerful risk-reduction mechanism in modern finance – 95–99 percent of bilateral exposures are eliminated.

(c) Joerg Osterrieder 2026

What Happens to Your Collateral If a CCP Member Defaults?



Reading the risk waterfall

- **Layer 1 – Defaulter's margin:** The first loss is absorbed by the defaulting member's own initial margin and variation margin
- **Layer 2 – Defaulter's default fund contribution:** The defaulter's share of the pooled default fund is consumed next
- **Layer 3 – CCP skin in the game:** The CCP's own capital – required by EMIR to align CCP incentives with members
- **Layer 4 – Non-defaulters' default fund:** Surviving members' contributions are mutualized
- **Layer 5 – Resolution tools:** Variation margin haircutting, forced allocation, or central bank intervention

EMIR: European Market Infrastructure Regulation requires CCPs to hold sufficient resources to survive the default of their two largest members simultaneously.

The five-layer waterfall ensures losses are absorbed in order: defaulter first, CCP capital third, mutualized funds fourth – but the last layer is everyone's problem.

(c) Joerg Osterrieder 2026

Can You Calculate How Much Collateral a CCP Saves the Market?

```python
import numpy as np

def simulate_netting(n_banks, n_trades, max_not=100):
    """Bilateral trades -> CCP netting efficiency."""
    np.random.seed(42)
    gross = np.zeros((n_banks, n_banks))
    for _ in range(n_trades):
        i, j = np.random.choice(n_banks, 2, replace=False)
        gross[i, j] += np.random.uniform(1, max_not)
    bilateral = np.sum(np.abs(gross))
    net_pos = gross.sum(axis=1) - gross.sum(axis=0)
    multilateral = np.sum(np.abs(net_pos))
    efficiency = 1 - multilateral / bilateral
    return bilateral, multilateral, efficiency

g, n, eff = simulate_netting(n_banks=20, n_trades=1000)
print(f"Gross: {g:,.0f} Net: {n:,.0f}")
print(f"Efficiency: {eff:.1%}")
print(f"Saved: {g-n:,.0f} ({eff:.1%} reduction)")
```

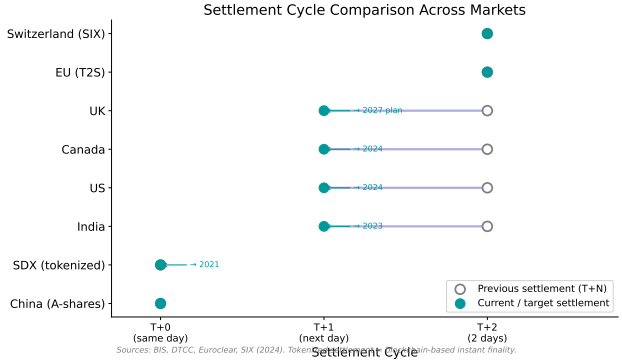
What the simulation shows

- Creates a random bilateral exposure matrix among 20 banks with 1,000 trades
- Computes gross bilateral exposure (sum of all obligations)
- Nets each bank's position: what it owes minus what it is owed
- **Typical result:** 95–98% netting efficiency, meaning the CCP reduces total exposure by a factor of 20–50x

Key insight: Netting efficiency increases with the number of participants and the number of trades – this is why CCP clearing exhibits strong network effects.

Fifteen lines of Python demonstrate why CCPs dominate post-trade: multilateral netting eliminates 95–98 percent of bilateral exposure.

Which Countries Settled Fastest – and Who Is Still Waiting?



Reading the Cleveland dot plot

- Each dot marks a country's settlement cycle; the horizontal axis measures days after trade execution
- **India** led the move to T+1 (January 2023), reducing counterparty exposure by 50% overnight
- **US and Canada** followed (May 2024), with Europe evaluating T+1 adoption
- **China** operates T+0 for equities on the Shanghai and Shenzhen exchanges
- **SDX (Switzerland):** tokenized securities settle in near-real-time via DLT, demonstrating T+0 feasibility for regulated instruments

The trend: Settlement cycles are compressing globally. DLT makes T+0 technically feasible; the remaining barriers are legal and operational.

India proved T+1 is operationally feasible at scale. The question is no longer whether shorter settlement is possible, but when others will follow.

Is Commission-Free Trading Really Free – or Does Someone Else Pay the Price?

Payment for Order Flow (PFOF) is the fee a market maker pays a broker to route retail orders to them.

PFOF revenue model:

$$R_{\text{PFOF}} = V_{\text{retail}} \cdot \bar{q} \cdot \bar{p} \cdot f_{\text{PFOF}} \tag{21}$$

where V_{retail} is retail order count, $\bar{q}$ is average quantity, $\bar{p}$ is average price, and f_{PFOF} is the per-share fee ($\sim$\$0.002–0.004/share in the US).

Value of retail flow to the market maker:

$$\Pi_{\text{MM}} = V_{\text{retail}} \cdot \bar{q} \cdot \left(\frac{s_{\text{NBBO}}}{2} - \text{PI} - f_{\text{PFOF}} \right) \tag{22}$$

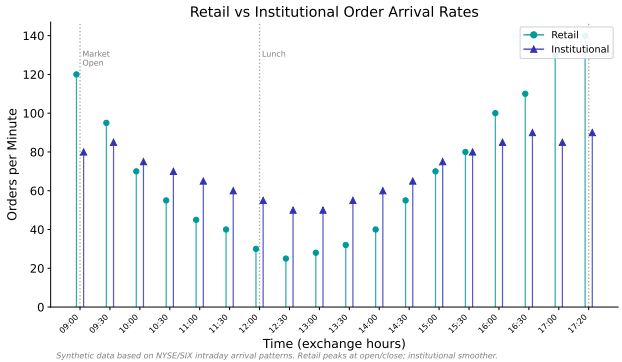
where PI is the price improvement given to the retail order. The market maker profits from the difference between the half-spread and the cost of price improvement plus PFOF.

MiFID II response: The EU banned PFOF in most jurisdictions (effective 2026), arguing it creates conflicts of interest incompatible with best execution obligations.

The paradox: US studies show retail orders routed via PFOF often receive *better* prices than on-exchange execution – but the comparison is against a spread that market makers themselves set.

PFOF makes brokers free for retail investors but raises the question: are you the customer or the product?

When Do Retail Traders Trade – and How Does It Differ from Institutions?



Reading the stem plot

- **Retail traders** show a pronounced U-shape: high activity at market open (reacting to overnight news) and close (end-of-day positioning)
- **Institutional traders** distribute orders more evenly across the day, using VWAP/TWAP algorithms to minimize impact
- **The mismatch:** Market makers profit from the predictability of retail timing patterns – they widen spreads at open when retail demand peaks
- **Lunchtime dip:** Both groups reduce activity during the midday low-volume window

Swiss pattern: Retail activity on Swissquote peaks at 09:00–10:00 CET (European open) and again at 15:30 CET (US open for cross-listed stocks).

Retail timing is predictable; institutional timing is algorithmic. Market makers profit from the gap between the two.

Can You Measure Whether Your Broker Gave You Best Execution?

```python
import numpy as np

def tca_report(orders, fills, market_vwap):
    """Transaction cost analysis: fills vs benchmarks."""
    results = []
    for order, fill in zip(orders, fills):
        sign = 1 if order['side']=='buy' else -1
        is_bps = sign * (fill['price'] - order['decision_price']) \
                 / order['decision_price'] * 1e4
        vw_bps = sign * (fill['price'] - market_vwap) \
                 / market_vwap * 1e4
        results.append({
            'is_bps': is_bps, 'vwap_bps': vw_bps,
            'qty': fill['qty']
        })
    wts = [r['qty'] for r in results]
    avg_is = np.average([r['is_bps'] for r in results],
                        weights=wts)
    avg_vw = np.average([r['vwap_bps'] for r in results],
                        weights=wts)
    return avg_is, avg_vw

orders = [{'side': 'buy', 'decision_price': 100.0}]
fills  = [{'price': 100.05, 'qty': 10000}]
is_c, vw_c = tca_report(orders, fills, market_vwap=100.03)
print(f"Impl. shortfall: {is_c:.1f} bps")
print(f"VWAP slippage: {vw_c:.1f} bps")
```

Two benchmarks, two stories

- **Implementation shortfall** (IS): compares fill price to the decision price – the price when the PM decided to trade. Captures the full cost of delay + impact
- **VWAP slippage**: compares fill price to the market VWAP. Measures whether you traded "with the market"
- A broker can beat VWAP but still have high IS if they delayed execution while the price moved

MiFID II requirement: Brokers must publish quarterly execution quality reports including IS and VWAP metrics, enabling clients to compare execution across brokers.

TCA is the accountability mechanism: without it, you cannot tell whether your broker optimized execution for you or for themselves.

How Does Tokenization Change the Economics of Owning and Trading Securities?

Traditional settlement cost (per transaction):

$$C_{\text{trad}} = c_{\text{custody}} + c_{\text{clearing}} + c_{\text{settlement}} + c_{\text{reconciliation}} + c_{\text{messaging}} \tag{23}$$

Estimated at 5–15 bps per trade for institutional settlement.

Tokenized settlement cost:

$$C_{\text{token}} = c_{\text{gas}} + c_{\text{smart contract audit}} + c_{\text{oracle}} \tag{24}$$

Estimated at 0.5–2 bps per trade, eliminating reconciliation and messaging entirely.

Break-even volume:

$$V_{\text{break}} = \frac{F_{\text{infra}}}{C_{\text{trad}} - C_{\text{token}}} \tag{25}$$

where F_{infra} is the fixed cost of building DLT infrastructure.

T+0 risk reduction. Counterparty exposure is proportional to settlement time:

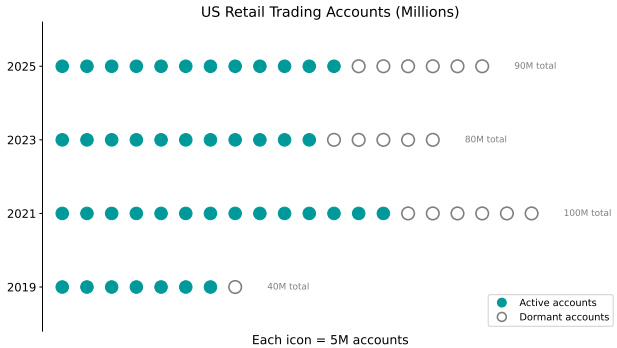
$$\text{Exposure}_{\text{T+0}} = \frac{0}{2} \cdot \text{Exposure}_{\text{T+2}} = 0 \tag{26}$$

Atomic delivery-versus-payment eliminates counterparty risk entirely.

Tokenization does not just speed up settlement – it eliminates entire cost layers (reconciliation, messaging, custodian fees) from the value chain.

(c) Joerg Osterrieder 2026

How Many Retail Investors Entered the Market – and What Changed?



Sources: FINRA, Schwab/Robinhood disclosures (2024). 2025 = estimate. Dormant = no trade in 12 months.

Reading the pictogram

- The **COVID surge** (2020–2021) saw retail trading volumes double or triple in most markets, driven by lockdowns, stimulus checks, and zero-commission apps
- **Post-surge dormancy:** Many new accounts became inactive within 12 months – the "tourist trader" phenomenon
- **Structural shift:** Even after the surge faded, retail's share of US equity volume settled at ∼25%, up from ∼10% pre-COVID
- **Meme stocks** (GameStop, AMC) demonstrated that coordinated retail flow can overpower institutional positioning

Swiss context: Swissquote reported record new account openings in 2020–2021, with the median client age dropping from 45 to 38.

The retail surge was not a blip – it permanently raised retail's structural share of market volume and forced regulators to rethink investor protection.

(c) Joerg Osterrieder 2026

What Have We Learned – and What Remains Unsolved?

Five sections, five key insights:

1. **Market Microstructure:** The order book is a queuing system where spread, depth, and impact are mathematically linked. Kyle's lambda and Glosten-Milgrom decomposition explain *why* spreads exist.
2. **Market Efficiency:** EMH is testable via variance ratios and autocorrelation. Real returns exhibit fat tails and time-varying correlations that standard models underestimate.
3. **Trading Infrastructure:** HFT latency advantages are worth millions per microsecond. Venue fragmentation is the regulatory consequence of competition policy – and dark pools are a cost-benefit routing decision.
4. **Post-Trade:** CCP multilateral netting eliminates 95–99% of bilateral exposure. The default waterfall protects the system – until it does not.
5. **Digital Markets:** PFOF subsidizes retail trading but creates conflicts. Tokenization eliminates entire cost layers. Retail's structural share has permanently increased.

Unsolved: How do you regulate algorithms that operate faster than human oversight? How do you preserve price discovery when half the volume is dark? How do you protect retail investors who are both empowered and exposed?

The transparency paradox was never solved – it was distributed across lit exchanges, dark pools, internalisers, and DLT settlement layers.

(c) Joerg Osterrieder 2026

Key Takeaways

1. The **bid-ask spread** decomposes into four costs (order processing, inventory, adverse selection, profit). Adverse selection dominates in liquid markets; understanding the decomposition is essential for evaluating execution quality.
2. The **Efficient Market Hypothesis** is testable, not philosophical. Variance ratio tests and autocorrelation analysis convert EMH into falsifiable hypotheses – and the evidence is mixed, varying by market, frequency, and time period.
3. **Market impact follows a square-root law**: doubling order size increases impact by 41%, not 100%. This single fact drives the design of every execution algorithm in institutional finance.
4. **CCP netting** eliminates 95–99% of bilateral exposure, making it the most powerful risk-reduction mechanism in post-trade infrastructure. The tradeoff is risk concentration in the CCP itself.
5. **Dark pools** are a routing optimization, not a conspiracy. They save the half-spread but risk information leakage. Smart order routers solve a continuous cost-benefit problem.
6. **Tokenized securities** compress settlement from T+2 to T+0, eliminating counterparty risk and multiple intermediary cost layers. Switzerland's SDX demonstrates feasibility under full regulation.

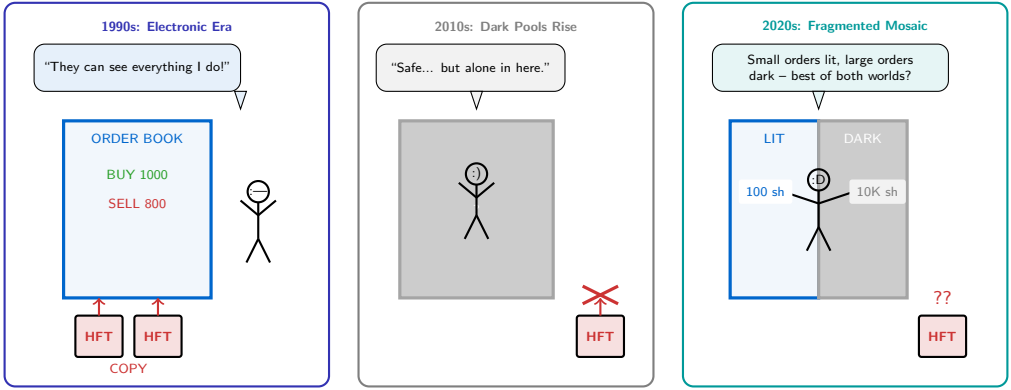
Six takeaways spanning microstructure, efficiency, execution, clearing, venue choice, and digital transformation – the full lifecycle of a trade.

(c) Joerg Osterrieder 2026

References and Further Reading

1. Kyle, A.S. (1985). "Continuous auctions and insider trading." *Econometrica*, 53(6), 1315–1335. *Foundational model of informed trading and price impact (Kyle's lambda).*
2. Glosten, L.R. & Milgrom, P.R. (1985). "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders." *Journal of Financial Economics*, 14(1), 71–100. *Derives the spread decomposition from adverse selection.*
3. Almgren, R. & Chriss, N. (2001). "Optimal execution of portfolio transactions." *Journal of Risk*, 3(2), 5–39. *Square-root impact model and optimal execution framework.*
4. Lo, A.W. & MacKinlay, A.C. (1988). "Stock market prices do not follow random walks: Evidence from a simple specification test." *Review of Financial Studies*, 1(1), 41–66. *Variance ratio test for market efficiency.*
5. Comerton-Forde, C. & Putnins, T.J. (2015). "Dark trading and price discovery." *Journal of Financial Economics*, 118(1), 70–92. *Empirical analysis of dark pool impact on price formation.*
6. Zhu, H. (2014). "Do dark pools harm price discovery?" *Review of Financial Studies*, 27(3), 747–789. *Theoretical model showing dark pools can improve welfare by sorting informed and uninformed traders.*

Core references: Kyle and Glosten-Milgrom for microstructure, Almgren-Chriss for execution, Lo-MacKinlay for efficiency, Comerton-Forde/Zhu for dark pools.



Markets evolved from transparent pits to opaque pools to fragmented mosaics. The transparency paradox was never solved – it was distributed.