The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response

Eric Budish, Peter Cramton and John Shim

July 2014
In 2010, Spread Networks invests $300mm to dig a high-speed fiber optic cable from NYC to Chicago.
The HFT Arms Race: Example

- In 2010, Spread Networks invests $300mm to dig a high-speed fiber optic cable from NYC to Chicago.
- Shaves round-trip data transmission time ... from 16ms to 13ms.
The HFT Arms Race: Example

- In 2010, Spread Networks invests $300mm to dig a high-speed fiber optic cable from NYC to Chicago.
- Shaves round-trip data transmission time ... from 16ms to 13ms.
- Industry observers: 3ms is an “eternity”.

▶

Industry observers: 3ms is an “eternity”.

▶

Not tunnels, but micro waves (rst 10ms, then 9ms, now 8ms).

▶

Analogous races occurring throughout the financial system, sometimes measured as finely as microseconds or nanoseconds.

▶

Last month alone: Lasers, BusinessWire.
The HFT Arms Race: Example

- In 2010, Spread Networks invests $300mm to dig a high-speed fiber optic cable from NYC to Chicago.
- Shaves round-trip data transmission time ... from 16ms to 13ms.
- Industry observers: 3ms is an “eternity”.
- Joke at the time: next innovation will be to dig a tunnel, “avoiding the planet’s pesky curvature”.
The HFT Arms Race: Example

In 2010, Spread Networks invests $300mm to dig a high-speed fiber optic cable from NYC to Chicago.

Shaves round-trip data transmission time . . . from 16ms to 13ms.

Industry observers: 3ms is an “eternity”.

Joke at the time: next innovation will be to dig a tunnel, “avoiding the planet’s pesky curvature”.

Joke isn’t that funny . . . Spread’s cable is already obsolete!
The HFT Arms Race: Example

- In 2010, Spread Networks invests $300mm to dig a high-speed fiber optic cable from NYC to Chicago.
- Shaves round-trip data transmission time ... from 16ms to 13ms.
- Industry observers: 3ms is an “eternity”.
- Joke at the time: next innovation will be to dig a tunnel, “avoiding the planet’s pesky curvature”.
- Joke isn’t that funny ... Spread’s cable is already obsolete!
- Not tunnels, but microwaves (first 10ms, then 9ms, now 8ms).
The HFT Arms Race: Example

▶ In 2010, Spread Networks invests $300mm to dig a high-speed fiber optic cable from NYC to Chicago.
▶ Shaves round-trip data transmission time ... from 16ms to 13ms.
▶ Industry observers: 3ms is an “eternity”.
▶ Joke at the time: next innovation will be to dig a tunnel, “avoiding the planet’s pesky curvature”.
▶ Joke isn’t that funny ... Spread’s cable is already obsolete!
▶ Not tunnels, but microwaves (first 10ms, then 9ms, now 8ms).
▶ Analogous races occurring throughout the financial system, sometimes measured as finely as microseconds or nanoseconds
  ▶ Last month alone: Lasers, BusinessWire
The HFT Arms Race: Market Design Perspective

- We examine the HFT arms race from the perspective of market design.
  - We assume that HFT’s are optimizing with respect to market rules as they’re presently given
  - But, ask whether these are the right rules
    - Avoids much of the “is HFT good or evil?” that seems to dominate the discussion of HFT
    - Instead, ask at a deeper level what is it about market design that incentivizes arms race behavior, and is this design optimal
- Central point: HFT arms race is a symptom of a basic flaw in modern financial market design: continuous-time trading.
- Proposal: make time discrete.
  - Replace continuous-time limit order books with discrete-time frequent batch auctions: uniform-price double auctions conducted at frequent but discrete time intervals, e.g., every 1 second or 100ms.
Frequent Batch Auctions

A simple idea: make time discrete.

1. Continuous limit-order books don’t actually “work” in continuous time: market correlations completely break down; frequent technical arbitrage opportunities

2. Technical arbitrage opportunities → arms race. Arms race looks like a “constant” of the market design.

3. Theory model: critique of the CLOB market design
   - Technical arbs are “built in” to the market design. Sniping.
   - Harms liquidity (spreads, depth)
   - Induces a never-ending arms race for speed

4. Frequent Batch Auctions as a market design response
   - Benefits: eliminates sniping, enhances liquidity, stops arms race, computational advantages
   - Cost: investors must wait a small amount of time to trade
Frequent Batch Auctions

A simple idea: make time discrete.

1. Continuous limit-order books don’t actually “work” in continuous time: market correlations completely break down; frequent technical arbitrage opportunities

2. Technical arbitrage opportunities $\rightarrow$ arms race. Arms race looks like a “constant” of the market design.

3. Theory model: critique of the CLOB market design
   - Technical arbs are “built in” to the market design. Sniping.
   - Harms liquidity (spreads, depth)
   - Induces a never-ending arms race for speed

4. Frequent Batch Auctions as a market design response
   - Benefits: eliminates sniping, enhances liquidity, stops arms race, computational advantages
   - Cost: investors must wait a small amount of time to trade
Brief Description of the Continuous Limit Order Book

Basic building block: limit order

Species a price, quantity, and buy/sell (bid/ask)

Buy 100 shares of XYZ at $100.00

Traders may submit limit orders to the market at any time during the trading day.

Also may cancel or modify outstanding limit orders at any time.

Orders and cancelations are processed by the exchange one-at-a-time in order of receipt (serial process).

Set of outstanding orders is known as the limit order book.

Trade occurs whenever a new limit order is submitted that is either (i) bid ≥ lowest ask; (ii) ask ≤ highest bid.

New limit order is interpreted as accepting (fully or partially) one or more outstanding orders.
Brief Description of the Continuous Limit Order Book

- Basic building block: limit order
  - Specifies a price, quantity, and buy/sell (bid/ask)
  - “Buy 100 shares of XYZ at $100.00”

- Traders may submit limit orders to the market at any time during the trading day
  - Also may cancel or modify outstanding limit orders at any time
  - Orders and cancelations are processed by the exchange one-at-a-time in order of receipt (serial process)

- Set of outstanding orders is known as the limit order book

- Trade occurs whenever a new limit order is submitted that is either (i) bid $\geq$ lowest ask; (ii) ask $\leq$ highest bid
  - New limit order is interpreted as accepting (fully or partially) one or more outstanding orders
Data

- “Direct feed” data from Chicago Mercantile Exchange (CME) and New York Stock Exchange (NYSE)
  - Gives “play by play” of limit order book
  - Millisecond resolution time stamps
  - These are the data HFT firms subscribe to and parse in real time

- Focus primarily on a pair of securities that track the S&P 500 index
  - ES: E-Mini S&P 500 Future, traded on CME
  - SPY: SPDR S&P 500 Exchange Traded Fund, traded on NYSE (and other equities exchanges)

- Time period: 2005-2011
Market Correlations Break Down at High Frequency
ES vs. SPY: 1 Day

Diagram showing the comparison between ES and SPY index points over a 1-day period, with time in Central Time (CT) on the x-axis and index points on the y-axis.
Mark et Correlations Break Down at High Frequency

ES vs. SPY: 1 hour

Market Correlations Break Down at High Frequency
ES vs. SPY: 1 hour
Market Correlations Break Down at High Frequency

ES vs. SPY: 1 minute
Market Correlations Break Down at High Frequency
ES vs. SPY: 250 milliseconds
Frequent Batch Auctions

A simple idea: make time discrete.

1. Continuous limit-order books don’t actually “work” in continuous time: market correlations completely break down; frequent technical arbitrage opportunities

2. Technical arbitrage opportunities → arms race. Arms race looks like a “constant” of the market design.

3. Theory model: critique of the CLOB market design
   - Technical arbs are “built in” to the market design. Sniping.
   - Harms liquidity (spreads, depth)
   - Induces a never-ending arms race for speed

4. Frequent Batch Auctions as a market design response
   - Benefits: eliminates sniping, enhances liquidity, stops arms race, computational advantages
   - Cost: investors must wait a small amount of time to trade
Arb Durations over Time: 2005-2011

Median over time

Distribution by year
Arb Per-Unit Profits over Time: 2005-2011

Median over time

Distribution by year
Arb Frequency over Time: 2005-2011

Median over time

Frequency vs. Volatility
Correlation Breakdown Over Time: 2005-2011
Arms Race is a “Constant” of the Market Design

- Results suggest that the arms race is a mechanical “constant” of the continuous limit order book.
  - Rather than a profit opportunity that is competed away over time

- Correlation Breakdown
  - Competition **does** increase the speed with which information is incorporated from one security price into another security price
  - Competition **does not** eliminate correlation breakdown

- Technical Arbitrage
  - Competition **does** increase the speed requirements for capturing arbs (“raises the bar”)
  - Competition **does not** reduce the size or frequency of arb opportunities

- These facts both inform and are explained by our model
Total Size of the Arms Race Prize

- Estimate annual value of ES-SPY arbitrage is $75mm (we suspect underestimate, details in paper)
- And ES-SPY is just the tip of the iceberg in the race for speed:
  1. Hundreds of trades very similar to ES-SPY: highly correlated, highly liquid
  2. Fragmented equity markets: can arbitrage SPY on NYSE against SPY on NASDAQ! Even simpler than ES-SPY.
  3. Correlations that are high but far from one can also be exploited in a statistical sense. Example: GS-MS
  4. Race to top of book (artifact of minimum price tick)
  5. Race to respond to public news (eg Business Wire, Fed)

We don’t attempt to put a precise estimate on the total prize at stake in the arms race, but common sense extrapolation from our ES-SPY estimates suggest that the sums are substantial
Frequent Batch Auctions

A simple idea: make time discrete.

1. Continuous limit-order books don’t actually “work” in continuous time: market correlations completely break down; frequent technical arbitrage opportunities
2. Technical arbitrage opportunities \(\rightarrow\) arms race. Arms race looks like a “constant” of the market design.
3. Theory model: critique of the CLOB market design
   - Technical arbs are “built in” to the market design. Sniping.
   - Harms liquidity (spreads, depth)
   - Induces a never-ending arms race for speed
4. Frequent Batch Auctions as a market design response
   - Benefits: eliminates sniping, enhances liquidity, stops arms race, computational advantages
   - Cost: investors must wait a small amount of time to trade
Model: Goal

Simple new model which is motivated by, and helps to explain, these empirical facts. The model serves two related purposes

1. Critique of the continuous limit order book market design
2. Identifies the economic implications of the HFT arms race
Model: Preliminaries

- There is a security, $x$, that trades on a continuous limit-order book market.
- There is a publicly observable signal, $y$, of the value of security $x$.
- Purposefully strong assumption:
  - Fundamental value of $x$ is *perfectly* correlated to the public signal $y$.
  - $x$ can always be costlessly liquidated at this fundamental value.
- “Best case” scenario for price discovery and liquidity provision in a continuous limit order book.
  - No asymmetric info, inventory costs, etc.
- We think of $x$ and $y$ as a metaphor for pairs or sets of securities that are highly correlated.
  - Ex: $x$ is SPY, $y$ is ES.
  - Ex: $x$ is SPY on NYSE (NASDAQ, dark pools, etc.), $y$ is SPY on BATS.
Evolution of $y$

- The signal $y$ evolves as a compound Poisson jump process
- Arrival rate $\lambda_{jump}$
- Jump distribution $F_{jump}$
  - Finite support
  - Symmetric with mean zero
- Let $J$ denote the random variable formed by drawing randomly according to $F_{jump}$, and then taking the absolute value.
  - The “jump size” distribution
Players: Investors and Trading Firms

Investors

- Represent end users of financial markets: mutual funds, pension funds, hedge funds, etc.
- Since there is no asymmetric information about fundamentals, could be called “liquidity traders” or “noise traders”
- Arrive stochastically to the market with an inelastic need to either buy or sell 1 unit of $x$
- Poisson arrival rate is $\lambda_{invest}$. Equal probability of need to buy vs. need to sell
- Mechanical strategy: trade at market immediately upon arrival
  - This is microfounded in the paper (weak preference to transact sooner rather than later; assume that investors act only as “takers” of liquidity, not “makers”)

\[ \text{Poisson arrival rate is } \lambda_{\text{invest}}. \text{ Equal probability of need to buy vs. need to sell} \]

\[ \text{Mechanical strategy: trade at market immediately upon arrival} \]

\[ \text{This is microfounded in the paper (weak preference to transact sooner rather than later; assume that investors act only as “takers” of liquidity, not “makers”)} \]
Players: Investors and Trading Firms

Trading Firms

- Equivalently: HFTs, market makers, algorithmic traders
- No intrinsic demand to buy or sell $x$
- Their goal in trading is simply to buy $x$ at prices lower than $y$ and sell at prices higher than $y$. Payoffs:
  - Buy $x$ at price $p$ at time $t$: earn $y_t - p$
  - Sell $x$ at price $p$ at time $t$: earn $p - y_t$
- Goal is to maximize profits per unit time
- Entry
  - Initially: # of trading firms is exogenous, $N \geq 2$
  - Below, we will endogenize entry
Latency

**Exogenous entry case**
- No latency in observing $y$
  - All trading firms and investors observe innovations in the signal $y$ with *zero time delay, for free.*
- No latency in submitting orders to the exchange
  - If multiple orders reach the market at the same time, the order in which they are processed is random (serial processing)
  - Alternatively, orders are transmitted with small random latency, and processed in order of receipt (eg, colocation)
- Again, best case scenario for CLOB

**Endogenous entry case**
- Will add latency in observing $y$
“Sniping”

- Given the model setup – no asymmetric information, no inventory costs, everyone risk neutral – one might conjecture that (Bertrand) competition among trading firms leads to effectively infinite liquidity for investors
  - That is, trading firms should offer to buy or sell $x$ at price $y$ in unlimited quantity at zero bid-ask spread
- But that is not what happens in the CLOB market design, due to a phenomenon we call “sniping”
“Sniping”

- Suppose $y$ jumps, e.g., from $y$ to $\bar{y}$
  - This is the moment at which the correlation between $y$ and $x$ temporarily breaks down
- Trading firms providing liquidity in the market for $x$ send a message to the CLOB
  - Withdraw old quotes based on $y$
  - Replace with new quotes based on $\bar{y}$
“Sniping”

- However, at the exact same time, other trading firms send a message to the CLOB attempting to “snipe” the stale quotes before they are adjusted
  - Buy at the old quotes based on \( y \), before these quotes are withdrawn
- Since the CLOB processes messages in serial – that is, one at a time – it is possible that a message to snipe a stale quote will get processed before the message to adjust the stale quote
- In fact, not only possible but probable
  - For every 1 liquidity provider trying to get out of the way
  - \( N - 1 \) other trading firms trying to snipe him
  - Hence, when there is a big jump, liquidity provider gets sniped with probability \( \frac{N-1}{N} \)
“Sniping”

- Hence, in a CLOB, *symmetrically observed public information creates technical arbitrage*.
  - Technical arbitrage is not supposed to exist in an efficient market (Fama, 1970)
  - Closely associated with correlation breakdown phenomenon
  - Missed by extant literature
  - Mechanically very similar to Glosten-Milgrom (1985) adverse selection, but caused by the market design not asymmetric information
  - In equilibrium, gets passed on to investors
Equilibrium, Exogenous Entry

The unique static Nash equilibrium is described as follows:

- Investors: trade immediately when their demand arises, buying or selling at the best available ask or bid, respectively.
- Trading Firms: of the $N$ trading firms, 1 plays a role we call “liquidity provider” and $N - 1$ play a role we call “stale-quote sniper”.
- Liquidity provider
  - Maintain a bid and ask for 1 unit of $x$ at spread of $s > 0$, derived below (stationary)
  - If $y_t$ jumps, send a message to cancel old quotes and replace with new quotes
- Snipers
  - if $y_t$ jumps such that $y_t > y_{t-} + \frac{s}{2}$ or $y_t < y_{t-} - \frac{s}{2}$, attempt to trade at the stale quote (“immediate or cancel”)
- Trading firms are indifferent between these two roles in equilibrium. (In practice, sorting is stochastic)
Equilibrium Bid-Ask Spread

In equilibrium, the bid-ask spread is such that trading firms are indifferent between liquidity provision and sniping.

- Return to liquidity provision
  - Benefits: \( \lambda_{\text{invest}} \cdot \frac{s}{2} \)
  - Costs: \( \lambda_{\text{jump}} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N-1}{N} \)

- Return to sniping
  - Benefits: \( \lambda_{\text{jump}} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{1}{N} \)

- Indifference condition:
  \[
  \lambda_{\text{invest}} \cdot \frac{s}{2} = \lambda_{\text{jump}} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \quad (1)
  \]

- Uniquely pins down \( s \). Interpretation:
  - LHS: revenue from investors due to non-zero bid-ask spread
  - RHS: rents to trading firms from technical arbitrage
Remark: Thin Markets

- What happens if investors sometimes need to trade 1 unit but sometimes need to trade 2 units?
- If the liquidity provider provides a quote with depth 2 at the same bid-ask spread as above
  - Benefits scale less than linearly with quote size: sometimes investors only want 1
  - Costs scale linearly with quote size: if get sniped, get sniped for the whole amount!
- Hence, equilibrium bid-ask spread is wider for second unit than first
- Not only is there a positive bid-ask spread even without asymmetric information about fundamentals, but markets are thin too
Equilibrium, Endogenous Entry

- Now, endogenize entry.
  - Trading firms and investors observe the signal $y$ with a small time delay, $\delta_{\text{slow}} > 0$, for free.
  - Can pay a cost $c_{\text{speed}}$ to reduce latency from $\delta_{\text{slow}}$ to $\delta_{\text{fast}}$, with $0 \leq \delta_{\text{fast}} < \delta_{\text{slow}}$. Let $\delta = \delta_{\text{slow}} - \delta_{\text{fast}}$.

- Equilibrium
  - Very similar structure to above: 1 liquidity provider, $N - 1$ stale-quote snipers.
  - $N$ now endogenous: the number of fast traders (for simplicity, allow $N$ real not integer).
  - Fast traders are indifferent between two roles as above.
  - Fast traders earn zero profits (could generalize to give inframarginal fast traders positive profits).
  - No role for slow traders in equilibrium.
Equilibrium, Endogenous Entry

- Zero-profit condition for liquidity provider

\[
\lambda_{\text{invest}} \cdot \frac{s}{2} - \lambda_{\text{jump}} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N - 1}{N} = c_{\text{speed}}
\]  

(2)

- Zero-profit condition for stale-quote snipers

\[
\lambda_{\text{jump}} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{1}{N} = c_{\text{speed}}
\]  

(3)

- Together, equations (2) and (3) describe equilibrium, by uniquely pinning down the bid-ask spread \( s \), the total entry of trading firms \( N \), and the indifference of trading firms between the two roles they might play.
Equilibrium, Endogenous Entry

- Adding (2) and \( N - 1 \) times (3) yields

\[
\lambda_{\text{invest}} \cdot \frac{s}{2} = Nc_{\text{speed}}
\]  \hspace{1cm} (4)

- Economic interpretation: all of the expenditure by trading firms on speed technology ultimately is borne by investors, via the bid-ask spread.
  - Arms-race prize = expenditures on speed = cost to investors
  - Remember: arms-race profits have to come from somewhere
What’s the Market Failure?

Chicago question: isn’t the arms race just healthy competition? what’s the market failure?
What’s the Market Failure?

Market Failure 1: Sniping

- Technical arb opportunities are “built in” to CLOB market design
- These arb opportunities violate weak-form EMH (Fama, 1970)
- Market looks highly efficient in time space, but it isn’t efficient in volume space
  - Lots of volume gets transacted at the instant prices become stale
- HFTs earn rents from symmetrically observed public information

Market Failure 2: Arms Race

- The arb rents then induce an arms race for speed
- Mathematically, a prisoners’ dilemma
Remarks on the Equilibrium

**Arms Race is a “constant”**

- Comparative static: the negative effects of the arms race on liquidity and welfare do not depend on either
  - the cost of speed (if speed is cheap, there will be more entry)
  - the magnitude of speed improvements (seconds, milliseconds, microseconds, nanoseconds, ...)
- The problem we identify is an equilibrium feature of continuous limit order books
  - not competed away as HFTs get faster and faster
  - ties in nicely with empirical results
Remarks on the Equilibrium

Role of HFTs

- In our model HFTs endogenously perform two functions
  - Useful: liquidity provision / price discovery
  - Rent-seeking: sniping stale quotes
- HFTs are indifferent between these two roles in equilibrium of our model
- The rent-seeking seems like zero-sum activity among HFTs
  - but we show that it ultimately harms real investors
- Frequent batching preserves the useful function but eliminates the rent seeking function (or at least reduces)
**Remark: Empirical Evidence of Effect of HFT on Liquidity**

Consistent with “IT Good, Speed Race Bad”

Virtu IPO Filing (Spreads)

Angel, Harris and Spatt (Cost to Trade Large Blocks)

---

**Block trade transaction costs have also fallen.**

Average Transaction Cost Estimate for 1M Shares in a $30 Stock

Source: Authors’ analysis of Ancerno trade data.
Frequent Batch Auctions

A simple idea: make time discrete.

1. Continuous limit-order books don’t actually “work” in continuous time: market correlations completely break down; frequent technical arbitrage opportunities

2. Technical arbitrage opportunities → arms race. Arms race looks like a “constant” of the market design.

3. Theory model: critique of the CLOB market design
   ▶ Technical arbs are “built in” to the market design. Sniping.
   ▶ Harms liquidity (spreads, depth)
   ▶ Induces a never-ending arms race for speed

4. Frequent Batch Auctions as a market design response
   ▶ Benefits: eliminates sniping, enhances liquidity, stops arms race, computational advantages
   ▶ Cost: investors must wait a small amount of time to trade
Frequent Batch Auctions: Overview

- High level: analogous to a CLOB, except time is discrete
- Discrete time then necessitates batch processing, using an auction
Frequent Batch Auctions: Definition (1 of 3)

- The trading day is divided into equal-length discrete batch intervals, e.g., 100 milliseconds.
- During the batch interval, traders submit bids and asks:
  - Can be freely modified, withdrawn, etc.
  - If an order is not executed in the batch at time $t$, it automatically carries over for $t + 1, t + 2, \ldots$.
- At the end of each interval, the exchange “batches” all of the outstanding orders, and computes market-level supply and demand curves.
Case 1: supply and demand don’t cross

- No trade
- All orders remain outstanding for the next batch auction
- We expect this case to be quite common
- Analogous to liquidity provision in a CLOB

Case 2: supply and demand cross (at price $p^*$)

- All bids strictly greater than $p^*$ and all asks strictly lower than $p^*$ transact their full quantity at $p^*$ ("uniform price")
- Bids and asks of exactly $p^*$ may get rationed
- Suggested rationing rule: pro-rata with time priority across batch intervals but not within batch intervals
- Historical aside: uniform-price auction originally proposed by Milton Friedman in the 1960s, for Treasury auctions.
Frequent Batch Auctions: Illustrated

(a) Case 1: No Trade

(b) Case 2: Trade
Frequent Batch Auctions: Definition (3 of 3)

- After the auction is computed, the following information is announced publicly
  - The market-clearing price (or the outcome “no trade”)
  - The quantity cleared (possibly zero)
  - The supply and demand curves
  - Optional: individual bids comprising the supply and demand curves

- Key point: new activity for the time $t$ batch auction is not visible during the time $t$ batch interval. Instead, everything is announced after the time $t$ auction is conducted. This is to prevent gaming.

- Analogous to current practice under the CLOB
  - (i) submit order; (ii) exchange processes order; (iii) public announcement
  - For auction at time $t$, traders see bids and asks at time $t - 1, t - 2, t - 3, \ldots$ (see market as it was a latency ago)
Why and How Batching Eliminates the Arms Race

Reason 1: frequent batch auctions reduce the value of a tiny speed advantage.

- Consider a slow trader who attempts to provide liquidity to investors
- There is 1 fast trader present in the market
- Continuous market: liquidity provider is vulnerable to being picked off by a fast trader for all jumps in $y$.
- Discrete market: liquidity provider is vulnerable to being picked off by a fast trader for only $\frac{\delta}{\tau}$ proportion of jumps in $y$:

![Diagram showing time intervals for slow and fast traders]
Reason 2: frequent batch auctions change the nature of competition when there are multiple fast traders: from competition on speed to competition on price

- As above, suppose a slow trader attempts to provide liquidity to investors
- There are $N \geq 2$ fast traders present in the market (exogenously)
- Suppose $y$ jumps in the interval $[\tau - \delta_{slow}, \tau - \delta_{fast}]$ where the liquidity provider is vulnerable.
- All of the fast traders wish to exploit the stale quote ... but this means that Bertrand competition drives the price of $x$ to the new, correct level.
- No longer can make money from symmetrically observed public information.
Ex: the Fed announces policy change at 2:00:00.000pm ...

- Continuous market: competition manifests in a race to exploit. Someone earns a rent for information that multiple market participants observe at basically the same time (logical extreme of Hirshleifer 1974)
- Batched market: competition simply drives the price to its new correct level for 2:00:01.000. Lots of orders reach the exchange by the end of the batch interval. Nobody earns a rent

N.B.: with batch intervals of e.g. 1 second, there is still plenty of scope for market participants to develop genuinely asymmetric information about security values, for which they will earn a rent
Equilibrium of Frequent Batch Auctions, Exogenous Entry

- $N \geq 2$ fast traders, exogenously in the market
- Description of equilibrium:
  - Bertrand competition drives bid-ask spread to zero, effectively infinite depth
  - No sniping
  - Fast traders earn zero gross profits (do not recover costs, treated as sunk)
- Highlights the difference between frequent batching and CLOB
  - No rents from symmetrically observed public information
  - No technical arbitrage opportunities
  - Bertrand competition competes spread to zero, as expected given model setup
- May also be useful for thinking about the transition to frequent batch auctions (some fast traders already present in the market with sunk costs in speed technology)
Equilibrium of Frequent Batch Auctions, Endogenous Entry

- Description of equilibrium
  - If $\tau$ sufficiently long relative to $\delta$, then in equilibrium trading firms do not pay $c_{speed}$ to be $\delta$ faster.
  - Slow trading firms provide $\bar{Q}$ units of liquidity at zero-bid ask spread.

- Key condition: not worth it for a fast trader to enter to pick off the slow traders:
  \[
  \frac{\delta \lambda_{jump}}{\tau} E(J) \cdot \bar{Q} < c_{speed}
  \]

- The fraction $\frac{\delta \lambda_{jump}}{\tau}$ is the proportion of time during the trading day during which the fast trader has a profitable sniping opportunity.

- For any finite $\bar{Q}$, the condition is satisfied for long enough $\tau$.
  - Hence, any desired market depth can be provided by slow traders at zero cost if the batch interval is sufficiently long.
How Long is Long Enough to Stop the Speed Race

Very rough calibration of $\tau$, using annual ES-SPY arbitrage estimates and other sources

$$\frac{\delta \lambda_{jump}}{\tau} E(J) \cdot \bar{Q} < c_{speed}$$

- $\delta$: speed difference between state of the art HFT in 2014 and state of the art in 2013.
  - ES/SPY: $\leq 500$ microseconds (.0005 seconds)
  - Equities: $\leq 100$ microseconds

- $\lambda_{jump}$: number of arbitrage opportunities, annualized
  - ES/SPY: 200,000 (800 per day)

- $E(J)$: size of each arb opportunity
  - ES/SPY: in data, 0.01 per share, but we expect narrower spreads so use 0.02 per share
How Long is Long Enough to Stop the Speed Race

Very rough calibration of $\tau$, using annual ES-SPY arbitrage estimates and other sources

$$\delta \lambda_{jump} \frac{E(J)}{\tau} \cdot \bar{Q} < c_{speed}$$

- $\bar{Q}$: depth of the order book in the batch auction
  - SPY depth averages 30,000 shares in our data. We expect greater depth, so use 100,000 shares

- $c_{speed}$, annualized
  - Virtu IPO filing: $100mm$ per year on capex and D&A. Assume 50% is speed related, 1% is ES/SPY related. $500k$ per year (Getco roughly double but includes Knight)
  - ES/SPY arb: $75mm$ per year, assume 20 firms competing for prize, suggests $3.5mm$ per year
\[
\tau > \delta \lambda_{\text{jump}} E(J) \cdot \bar{Q} \frac{1}{c_{\text{speed}}}
\]

\[
> .0005 \times 200,000 \times .02 \times 100,000 \times \frac{1}{500,000}
\]

\[
> .4 \text{seconds}
\]

- Very rough:
  - we got numbers as low as 12ms
    - (using \(\delta = .1 \text{ms}\) and \(c_{\text{speed}} = $3.5 mm\))
  - and as high as 6s
    - (using \(\delta = .5 \text{ms}, \lambda_{\text{jump}} = 1600, \bar{Q} = 500,000, E(J) = 0.03\))
- N.B. in paper we also analyze equilibrium with \(\tau\) very short and endogenous entry
  - Liquidity provider invests in speed but nobody else does (no sniping)
  - Spreads not zero but narrower than in CLOB, welfare higher than in CLOB
Summary: Equilibrium Costs and Benefits of Frequent Batching

- **Benefits**
  - Enhanced liquidity
    - Narrower spreads
    - Increased depth
  - Eliminate socially wasteful arms race
- **Costs**
  - Investors must wait until the end of the batch interval to transact
Computational Benefits of Frequent Batching

- **Overall**
  - Continuous-time markets implicitly assume that computers and communications technology are infinitely fast.
  - Discrete time respects the limits of computers and communications. Computers are fast but not infinitely so.

- **Algorithmic traders**
  - Continuous: Always uncertain about current state; temptation to trade off robustness for speed
  - Discrete: Everyone knows state at time $t$ before decision at time $t + 1$

- **Exchanges**
  - Continuous: Computational task is mathematically impossible; latencies and backlog unavoidable
  - Discrete: Computation is easy

- **Regulator**
  - Continuous: Audit trail is difficult to parse; who knew what when? in what order did events occur across markets?
  - Discrete: Simple audit trail; state at $t, t + 1, ...$
Alternative Responses to the HFT Arms Race

- Tobin Tax
  - Does partially mitigate sniping
  - But: cost of tax gets passed on to investors

- Random delay
  - Does mitigate incentive to invest in speed
  - Does \textit{not} mitigate sniping
  - Each message to snipe is like a lottery ticket
  - Explosion in message traffic

- Message-to-trade ratios
  - Hard to analyze
  - But: note that high message-to-trade ratios are \textit{equilibrium} feature of CLOB

- Minimum resting times
  - Exacerbates sniping

- IEX speed bump + price sliding to NBBO midpoint
  - Ingenious, eliminates sniping
  - But, only works while IEX is small relative to the rest of the continuous market (free-rides off price discovery elsewhere)
Open Questions

▶ Another Chicago question: if this is such a good idea, why hasn’t an exchange already tried it? Potential reasons:

▶ Relatively new problem
▶ Coordination challenge
▶ Regulatory ambiguities
▶ Vested interests in the current market structure

▶ Issues due to fragmentation of US equities markets

▶ What is equilibrium if there are both batch and continuous exchanges operating in parallel?
▶ Mechanics if multiple exchanges each run batch (how to ensure law of one price)
▶ Interaction with Reg NMS
Open Questions

- Market stability
  - Common claim among policy makers is that stopping the HFT arms race would enhance market stability (meaning vulnerability to flash crashes, exchange outages, programming glitches, etc.)
  - This is another potential welfare benefit of frequent batching, but not yet modeled

- Implementation details
  - Optimal batch interval, and how this varies by security
  - Tick sizes?
  - Circuit breakers?
  - Further details of information policy
  - Note: these questions will likely require a richer model (e.g. with asymmetric information)
“Insider Trading 2.0 – A New Initiative to Crack Down on Predatory Practices”

We have to review, and it’s something I want to raise, and I’m sure it will be discussed at the panel, and carefully consider a proposal that I like very much. It was put forward by economists at the University of Chicago School of Business – not an enemy of free markets, the University of Chicago School of Business, by any means.

In December, they issued a detailed and thoughtful proposal for reforms that would fundamentally reorient the markets in a very simple way that would help restore confidence in them. Their proposals would reaffirm the basic concept that the best price – not the highest speed – should win.

Currently, on our exchanges, securities are traded continuously, which means that orders are constantly accepted and matched with ties broken based on which orders arrived first. This system rewards high-frequency traders who continuously flood the market with orders – emphasizing speed over price.
The University of Chicago proposal – which I endorse – would, in effect, put a speed bump in place. Orders would be processed in batches after short intervals – potentially a second or less than a second in length – but that would ensure that the price would be the deciding factor in who obtains a trade, not who has the fastest supercomputer and early access to market-moving information.

This structural reform – sometimes called “frequent batch auctions” – would help catch and cap the supercomputer arms race now underway. This is tremendously important, because even advocates of high-frequency trading have always recognized that the potential for destabilization of the markets from volatility is a problem.

If you had frequent batch auctions, there’s no point in trying to get faster than whatever the interval is. It would discourage the risk taking that can cause flash crashes because, in the quest for greater and greater speed, there is, in and of itself, a threat to market stability. It rewards those who are taking chances. It rewards those who try risky new ways to gain a few milliseconds of speed. And that’s something we could put an end to if this proposal were successfully carried out.
“Enhancing Our Equity Market Structure”

... We must consider, for example, whether the increasingly expensive search for speed has passed the point of diminishing returns. I am personally wary of prescriptive regulation that attempts to identify an optimal trading speed, but I am receptive to more flexible, competitive solutions that could be adopted by trading venues. These could include frequent batch auctions or other mechanisms designed to minimize speed advantages. They could also include affirmative or negative trading obligations for high-frequency trading firms that employ the fastest, most sophisticated trading tools.

... A key question is whether trading venues have sufficient opportunity and flexibility to innovate successfully with initiatives that seek to deemphasize speed as a key to trading success in order to further serve the interests of investors. If not, we must reconsider the SEC rules and market practices that stand in the way.
“Slowing Down the Stock Market”

Today’s stock market is falling short. A wasteful arms race among high-frequency traders, the growth of dark pools (private trading venues) and assorted conflicts of interest have undermined its performance. If investors don’t trust the market, that hurts capital formation, not to mention retirement and college savings.

... Fixing the problems will require more than a tweak here and there. One idea that’s winning converts would replace the 24-hour, continuous trading of stocks with frequent auctions at regular intervals.

Why would that help? Because it would lessen the emphasis on speed and direct more attention to the price that investors are willing to pay for stocks, given the prospects of the companies concerned, their industries and the broader economy. The high-speed arms race would subside, because shaving another millisecond off the time it takes to trade would confer no benefit.

The idea has a good pedigree. It has been around at least since 1960, when Milton Friedman proposed a version for the sale of U.S. Treasury bonds. Researchers led by the University of Chicago’s Eric Budish refined the concept in a paper last year.
As well as prioritizing price over speed, this approach would make another flash crash less likely.

Auctions would also reduce the need for dark pools, because the orders of institutional investors wouldn’t be visible to other participants.

The conflicts of interest that brokers now face when they send orders to the trading venue that pays them the highest rebate or fee, rather than the one that offers the best execution, would recede as well.

Goldman Sachs Group Inc., among others, is interested enough in frequent batch auctions that it’s working with Budish to find an exchange that will conduct a pilot program and a regulatory agency that will monitor the results. Mary Jo White, the Securities and Exchange Commission chair, indicated in a June 5 speech her interest in batch auctions. She should make it a priority to conduct a test program. It’s a promising idea.
Summary

- We take a market design perspective to the HFT arms race.
- Root problem isn’t “evil HFTs”, it’s continuous-time / serial-process trading.
- Alternative: discrete-time / batch-process trading

1. Continuous-time markets are a fiction: correlations break down; frequent technical arbs
2. Technical Arbs $\rightarrow$ Arms Race. Arms Race is a “constant” of the market design.
3. Theory: root cause is the CLOB market design
   - Arms race is a never-ending, equilibrium feature of the CLOB
   - Arms race harms liquidity and is socially wasteful
4. Frequent Batch Auctions as a market design response
   - Benefits: eliminates sniping, stops arms race, enhances liquidity, computational advantages
   - Costs: investors must wait a small amount of time to trade, unintended consequences