A Market Design Approach to the HFT Debate: The Case for Frequent Batch Auctions

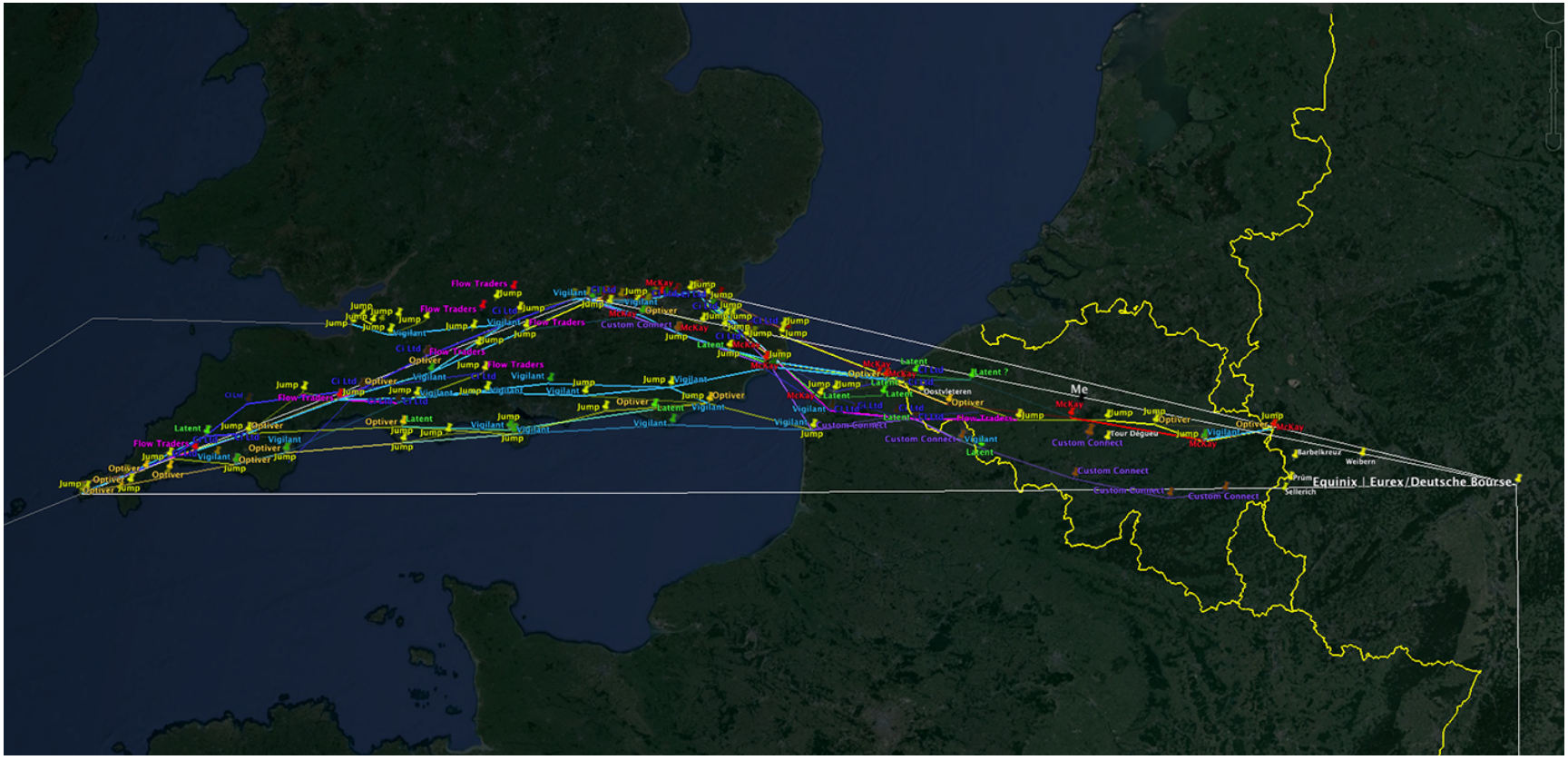
Eric Budish, University of Chicago

ECMI Annual Conference, National Bank of Belgium, Brussels
Oct 2014

The HFT Arms Race

- 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
- Last week alone
  - "Speed-of-Light Trading Expands in Europe with McKay Network" (Bloomberg)
  - "Run EDGAR Run: SEC Dissemination in a High Frequency World" (Chicago Booth Working Paper)

The HFT Arms Race



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: discrete-time trading.
  - 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: discrete-time trading.

1. Direct-feed millisecond level data: Continuous limit-order books don't actually "work" in continuous time
   - Market correlations completely break down
   - Frequent mechanical arbitrage opportunities
2. Mechanical arbs –> arms race. Arms race looks like a "constant"
3. Theory model: critique of the CLOB market design
   - Mechanical 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
   - Stops the arms race
   - Competition on speed –> competition on price. No sniping.
   - Enhances liquidity and social welfare

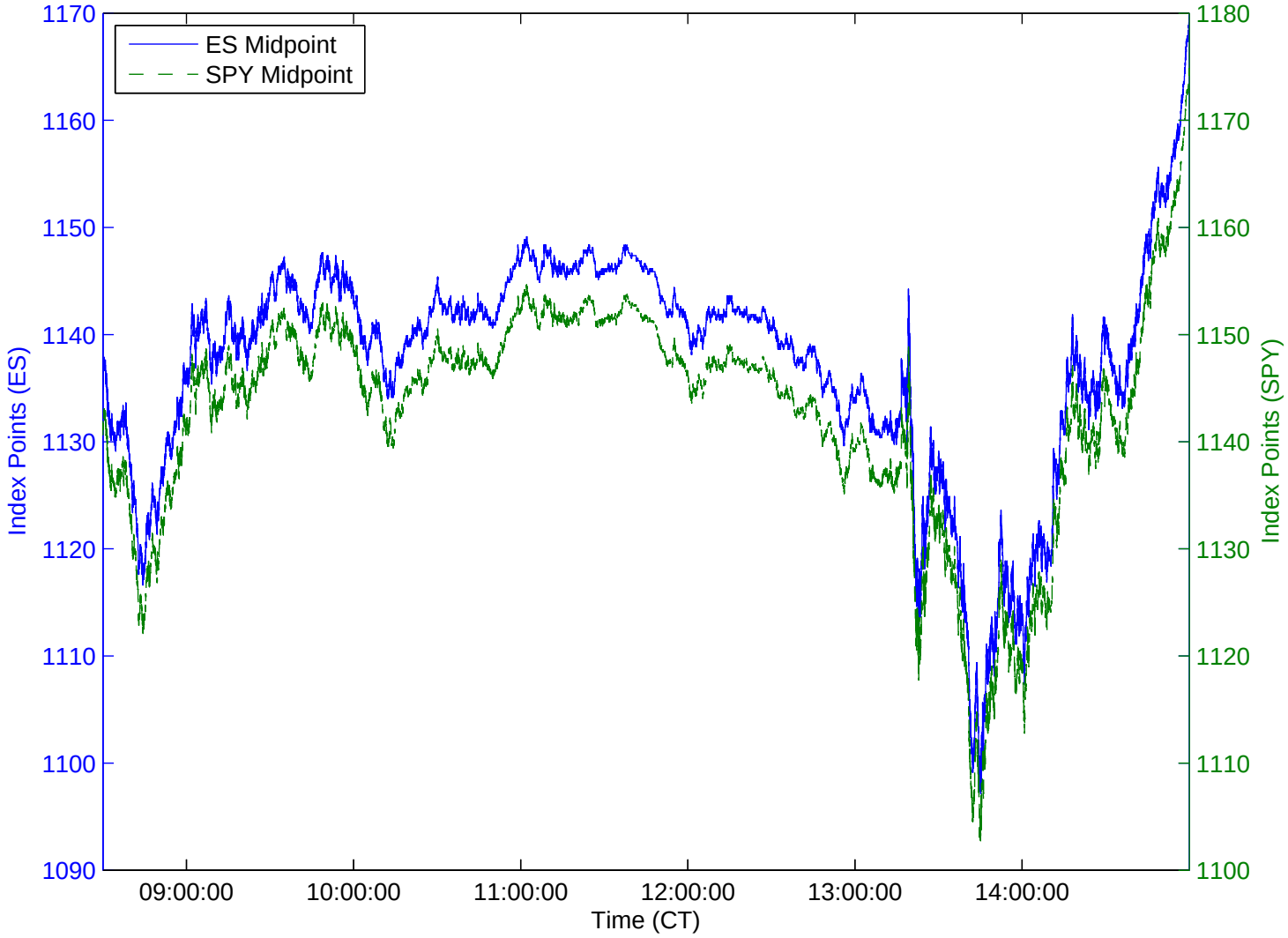
Frequent Batch Auctions

A simple idea: discrete-time trading.

1. **Direct-feed millisecond level data: Continuous limit-order books don't actually "work" in continuous time**
   - **Market correlations completely break down**
   - **Frequent mechanical arbitrage opportunities**
2. Mechanical arbs –> arms race. Arms race looks like a "constant"
3. Theory model: critique of the CLOB market design
   - Mechanical 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
   - Stops the arms race
   - Competition on speed –> competition on price. No sniping.
   - Enhances liquidity and social welfare

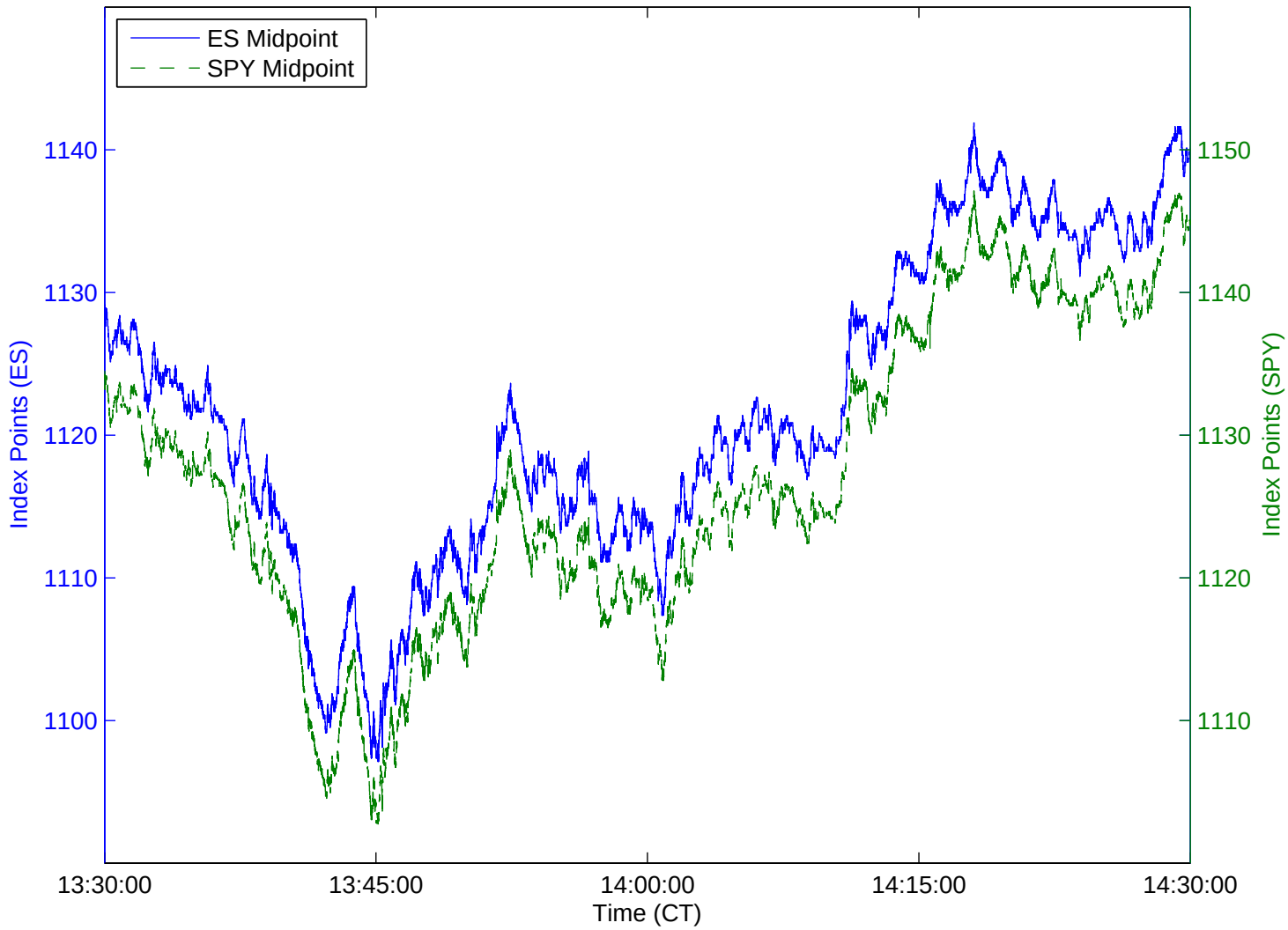
Market Correlations Break Down at High Frequency

ES vs. SPY: 1 Day



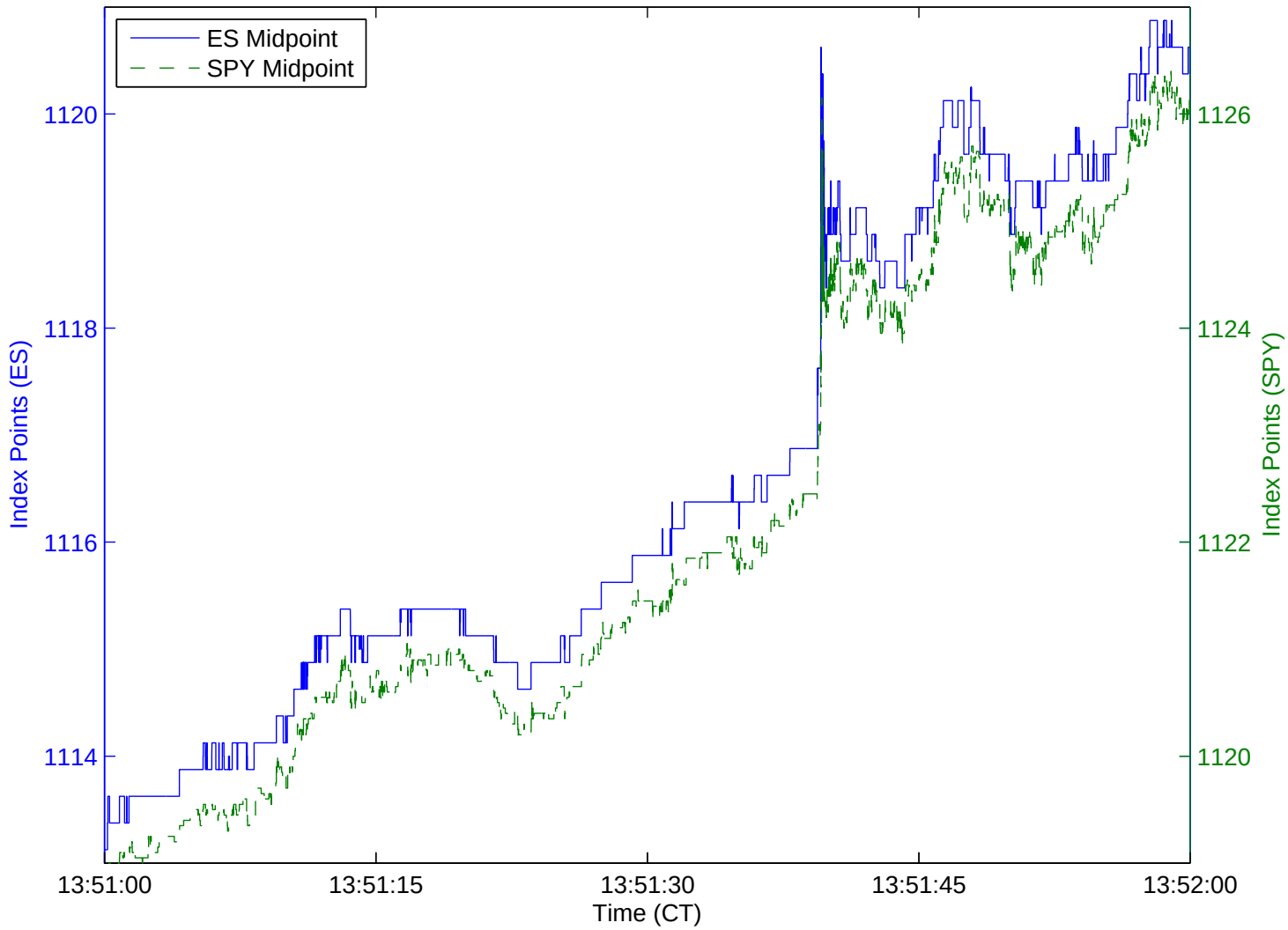
Market Correlations Break Down at High Frequency

ES vs. SPY: 1 hour



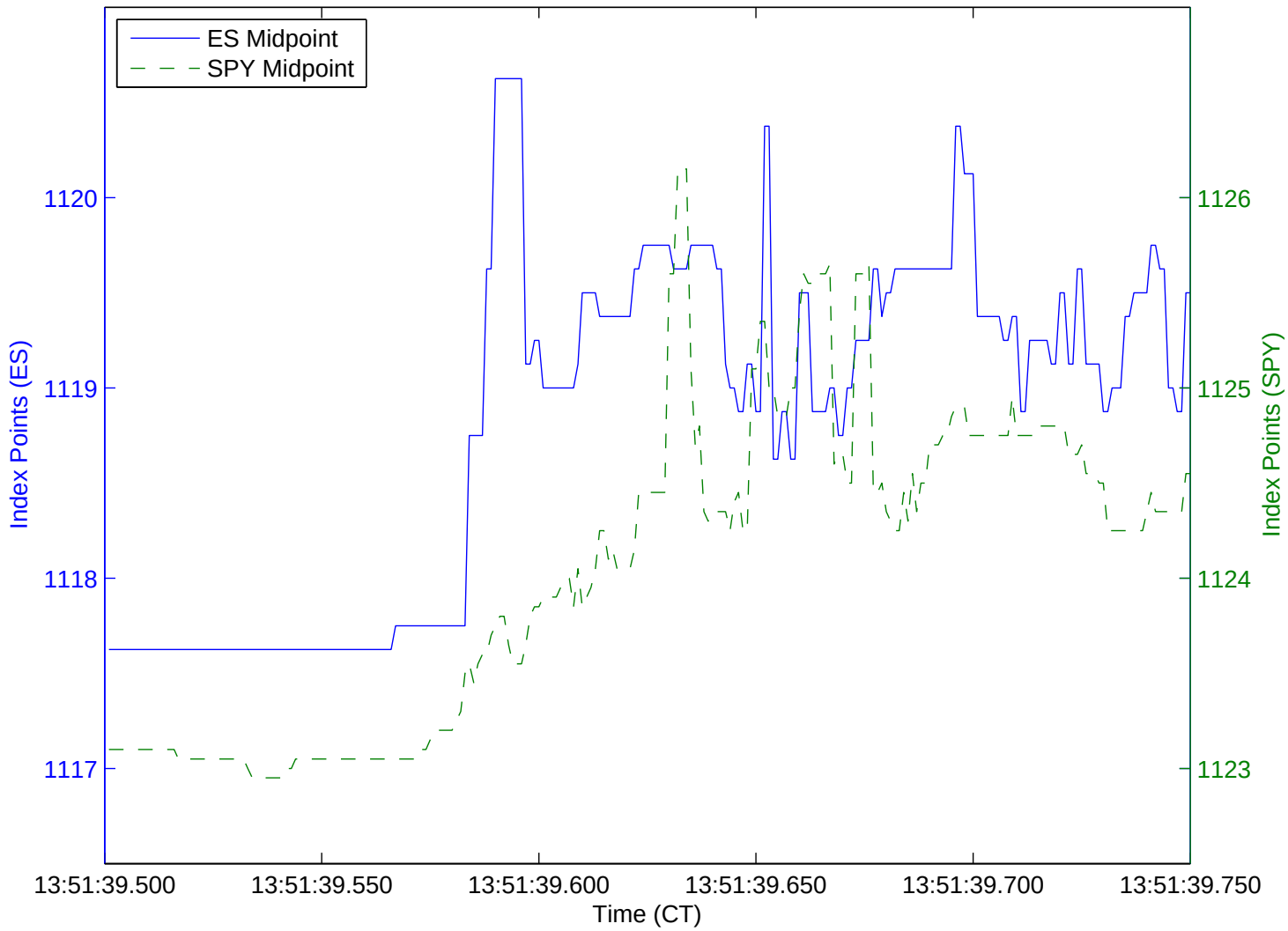
Market Correlations Break Down at High Frequency

ES vs. SPY: 1 minute



Market Correlations Break Down at High Frequency

ES vs. SPY: 250 milliseconds



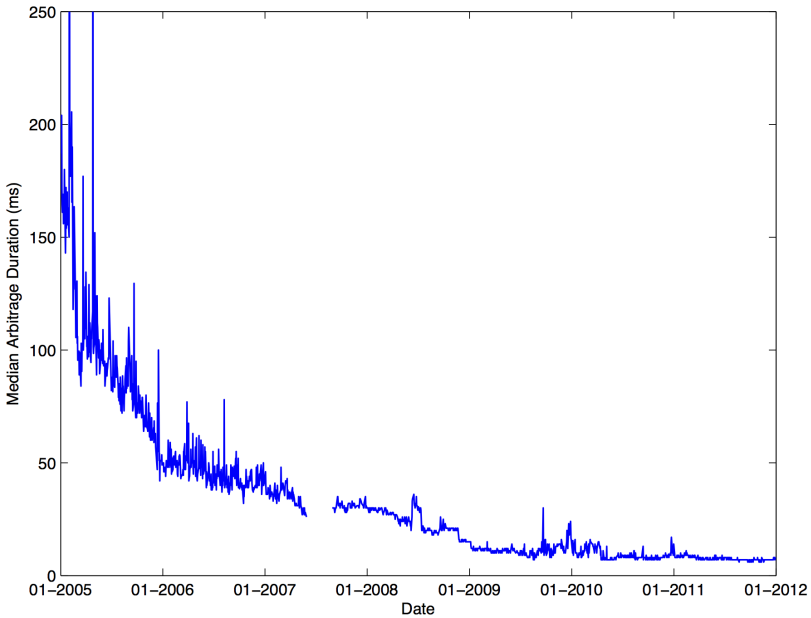
Frequent Batch Auctions

A simple idea: discrete-time trading.

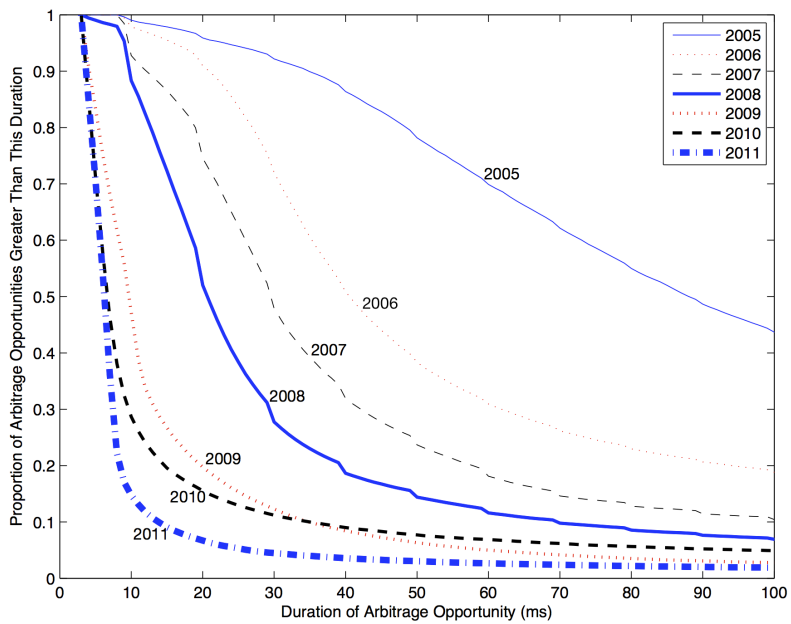
1. Direct-feed millisecond level data: Continuous limit-order books don't actually "work" in continuous time
   - Market correlations completely break down
   - Frequent mechanical arbitrage opportunities
2. **Mechanical arbs –> arms race. Arms race looks like a "constant"**
3. Theory model: critique of the CLOB market design
   - Mechanical 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
   - Stops the arms race
   - Competition on speed –> competition on price. No sniping.
   - Enhances liquidity and social welfare

Arb Durations over Time: 2005-2011

Median over time

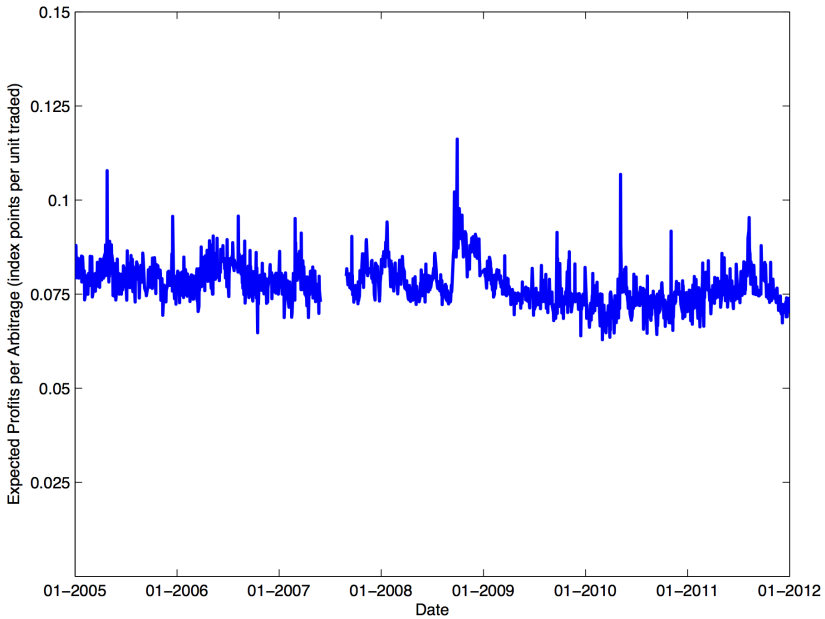


Distribution by year

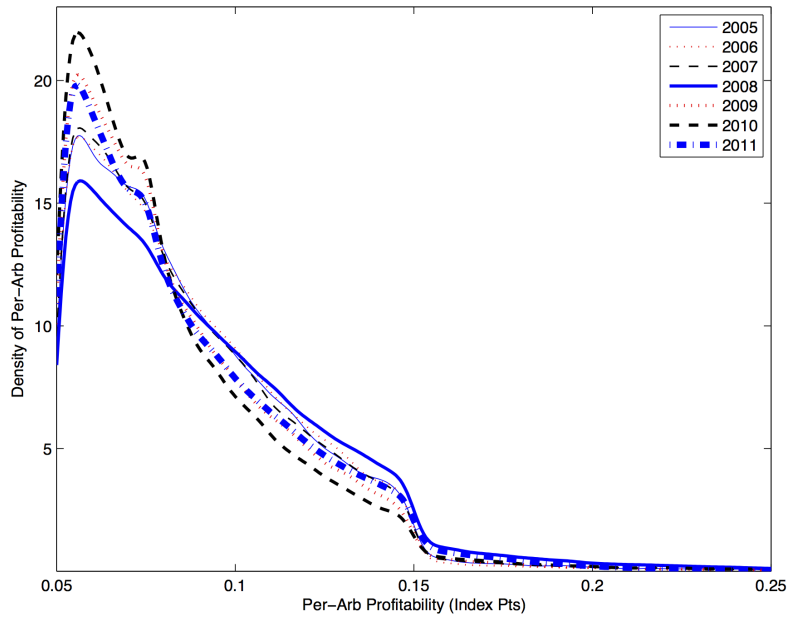


Arb Per-Unit Profits over Time: 2005-2011

Median over time

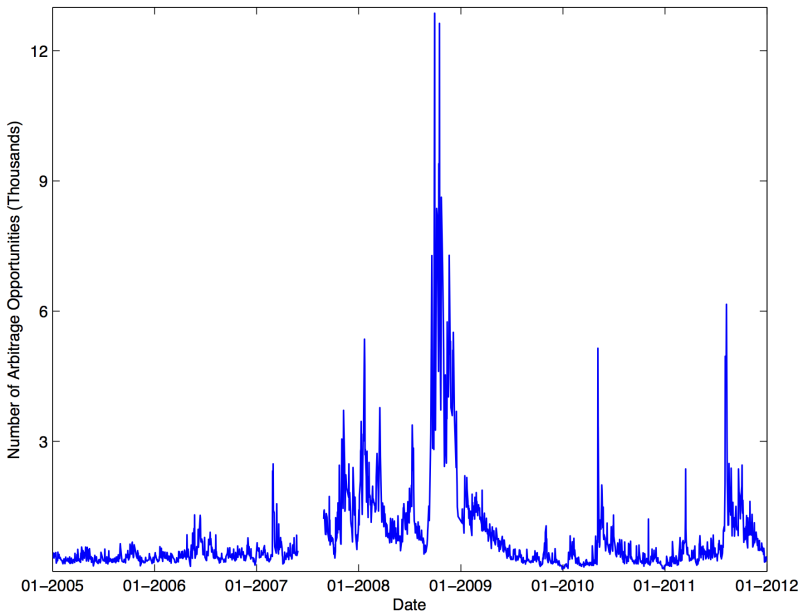


Distribution by year

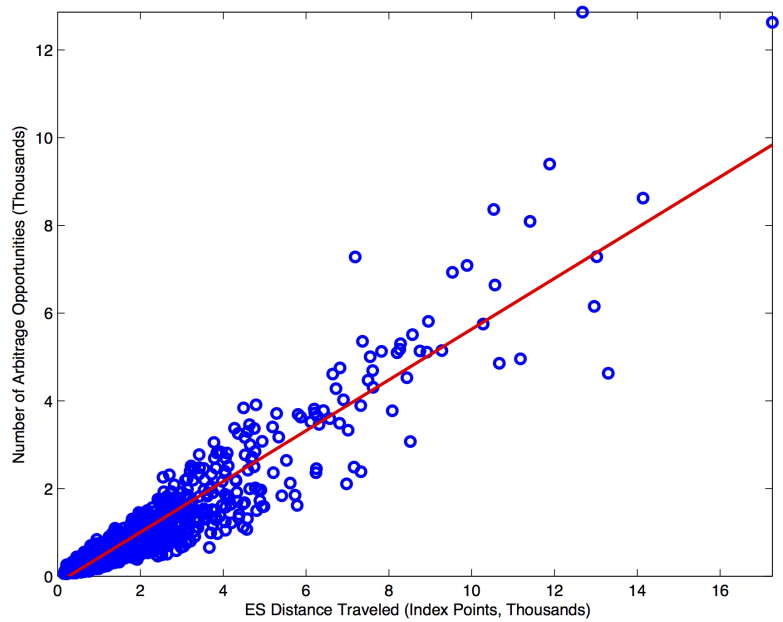


Arb Frequency over Time: 2005-2011

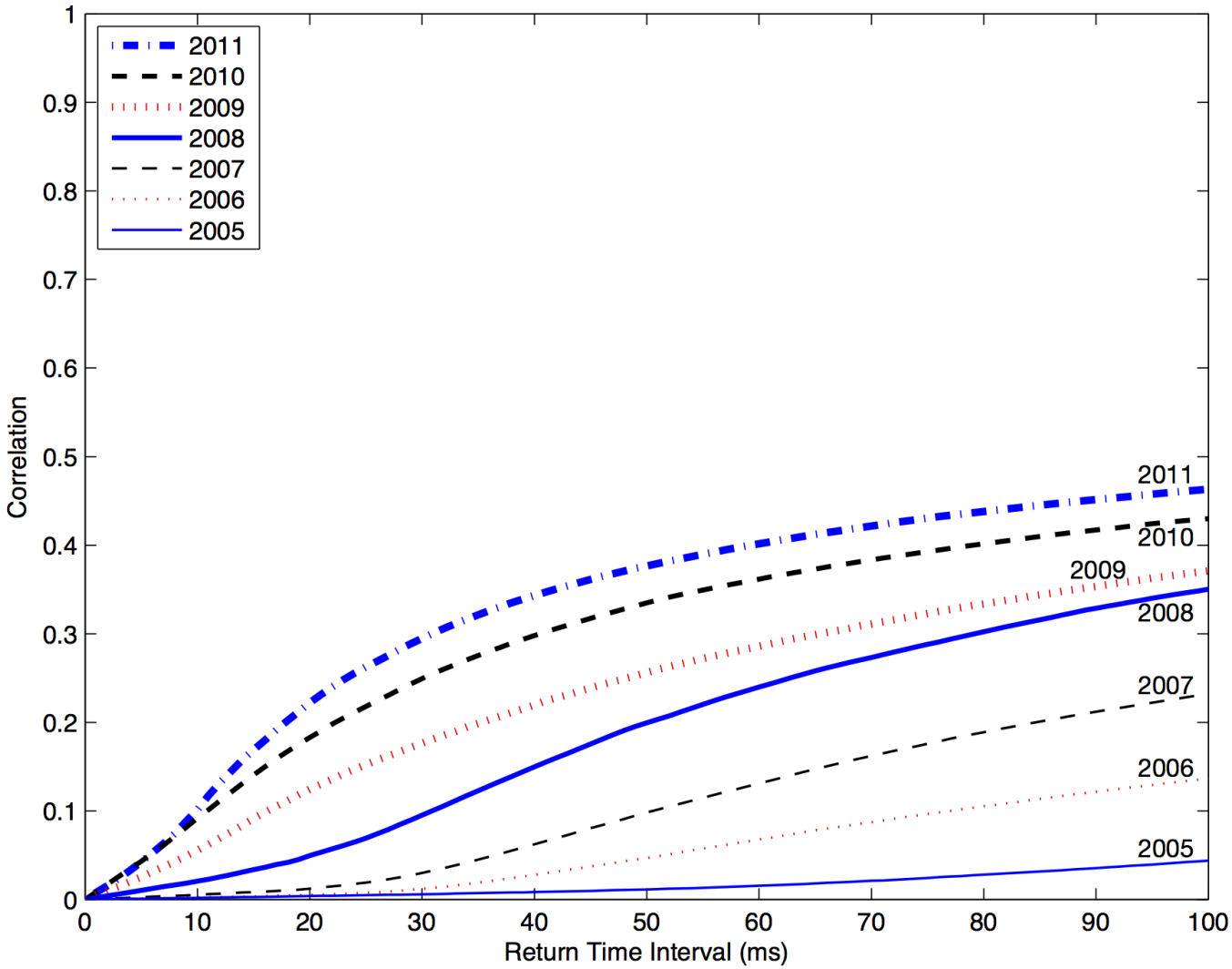
Frequency over time



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
- Mechanical 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: discrete-time trading.

1. Direct-feed millisecond level data: Continuous limit-order books don't actually "work" in continuous time
   - Market correlations completely break down
   - Frequent mechanical arbitrage opportunities
2. Mechanical arbs –> arms race. Arms race looks like a "constant"
3. **Theory model: critique of the CLOB market design**
   - **Mechanical 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
   - Stops the arms race
   - Competition on speed –> competition on price. No sniping.
   - Enhances liquidity and social welfare

Model: Key Idea

Key idea: mechanical arbitrage opportunities like ES-SPY are "built in" to the CLOB. The profits from these arbitrage opportunities harm liquidity provision.

- Why? Consider the race from a liquidity provider's perspective
  - Suppose there is a publicly observable news event that causes his quotes to become "stale"
    - E.g., a change in the price of a highly correlated security (ES/SPY), central bank announcement, company announcement
  - 1 of him, trying to adjust his stale quotes
  - Many others, trying to "snipe" his stale quotes
  - In a continuous limit order book, messages are processed one-at-a-time in *serial* ...
  - so the 1 usually loses the race against the Many ...
  - Even if he, too, is at the cutting edge of speed
- Hence, in a CLOB, *even symmetrically observed public information creates arbitrage rents*
  - Such arbitrages are not supposed to exist in an efficient market (Fama, 1970). Matters because the arbs harm liquidity.

Model: Key Idea

- This technical cost of providing liquidity – "sniping" – is incremental to the usual fundamental costs of providing liquidity
  - Asymmetric information, inventory costs, search costs
- In a competitive market, sniping costs get passed on to investors
  - Thinner markets, wider bid-ask spreads
- Sniping creates a never-ending race for speed
  - Snipers: win race to pick off stale quotes
  - Liquidity providers: get out of the way of the snipers!
- Ultimately, in equilibrium of our model, all of the $ spent in the arms race come out of the pockets of investors
  - 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

- Mechanical 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
  - Even for public / technical info (e.g., a jump in ES): *somebody is always first to react*

Market Failure 2: Arms Race

- The arb rents then induce an arms race for speed
- Mathematically, a prisoners' dilemma

Model: Additional Remarks

The Arms-Race is a "Constant"

- Comparative static: the negative effects of the arms race 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
  - *Takeaway: the race for speed will never end as long as we have continuous-time trading*

Model: Additional Remarks

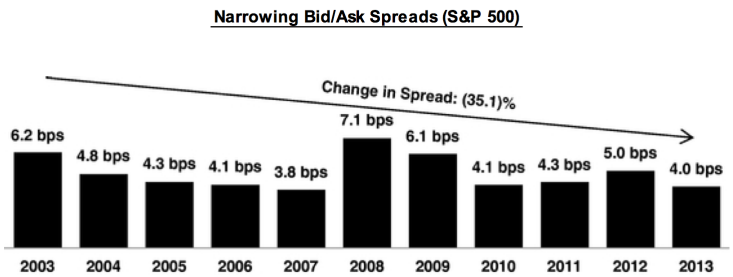
Role of HFTs

- In our model HFTs endogenously perform two functions
  - Useful: liquidity provision / price discovery
  - Rent-seeking: picking off stale quotes
- 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)
- Nuance
  - Our results *do not* imply that on net HFT has been negative for liquidity or social welfare.
  - Our results *do* say that sniping is negative for liquidity and that the speed race is socially wasteful.

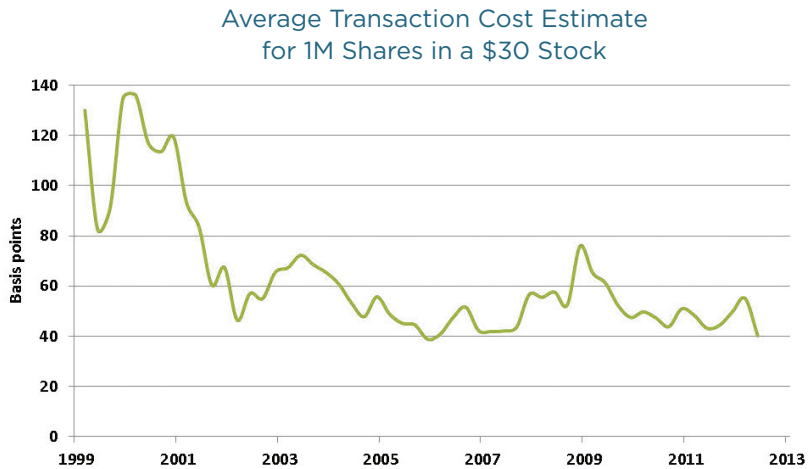
Remark: Empirical Evidence of Effect of HFT on Liquidity

Consistent with "IT Good, Speed Race Bad"

Virtu IPO Filing (Spreads)



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



Source: Authors' analysis of Ancerno trade data.

Frequent Batch Auctions

A simple idea: discrete-time trading.

1. Direct-feed millisecond level data: Continuous limit-order books don't actually "work" in continuous time
   - Market correlations completely break down
   - Frequent mechanical arbitrage opportunities
2. Mechanical arbs –> arms race. Arms race looks like a "constant
3. Theory model: critique of the CLOB market design
   - Mechanical 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**
   - **Stops the arms race**
   - **Competition on speed –> competition on price. No sniping.**
   - **Enhances liquidity and social welfare**

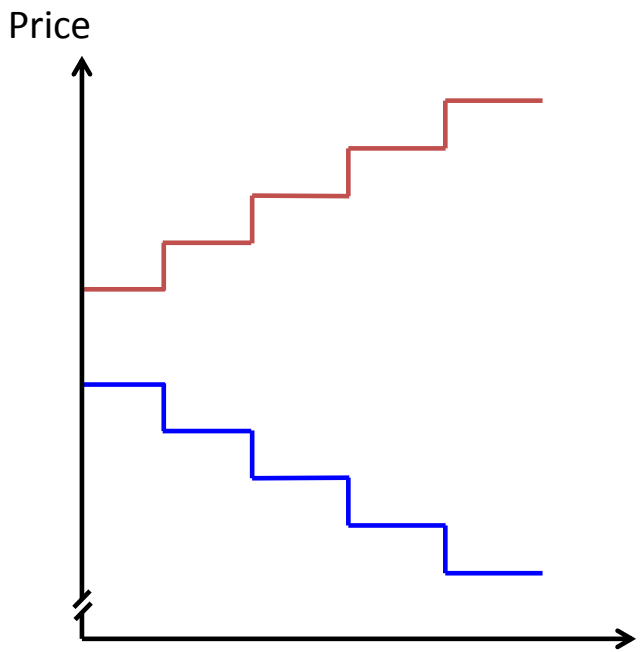
Frequent Batch Auctions: Overview

- High level: analogous to a CLOB, except time is discrete
- Discrete time then necessitates batch processing, using an auction

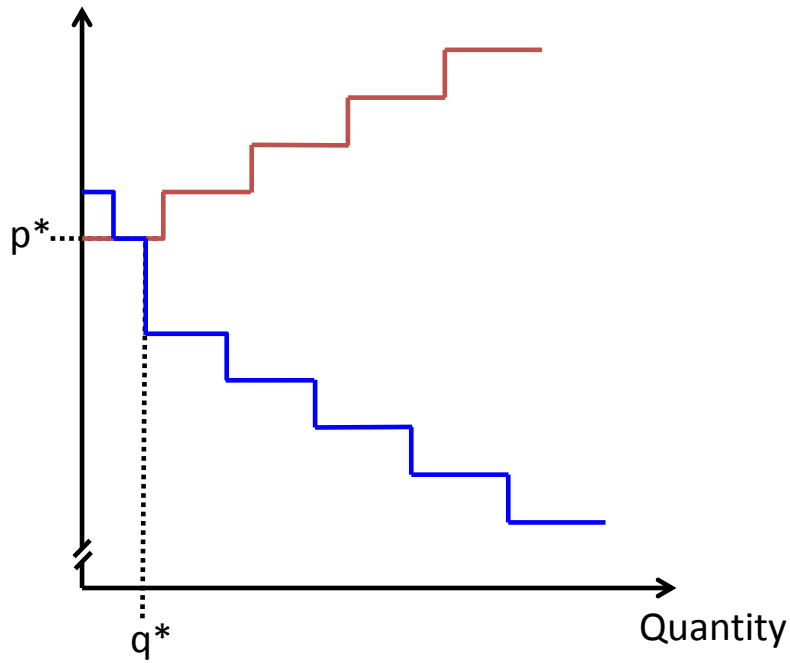
Frequent Batch Auctions: Definition

- The trading day is divided into equal-length discrete batch intervals, of length e.g. 1 second or 100ms.
- During the batch interval traders submits 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
- If supply and demand intersect, then all transactions occur at the same market-clearing price ("uniform price")
  - Priority: still price-time, but treat time as discrete. Orders submitted in the same batch interval have the same priority. Rationing is pro-rata.
- Information policy: orders are not visible during the batch interval. Aggregate demand and supply are announced at the end.
  - Analogous to current practice under the continuous limit-order book

Frequent Batch Auctions: Illustrated



(a) Case 1: No Trade



(b) Case 2: Trade

Frequent Batching Directly Solves the Problems with Continuous Trading

Reason 1: Discrete-time reduces value of tiny speed advantages

- Suppose there are two traders: one is faster than the other by 1ms
  - Continuous market: whenever anything happens, faster trader gets to act first.
  - Discrete market: most news events, either both slow and fast have plenty of time to react, or neither can react by end of interval
    - News has to occur at very precise moment in batch interval to give fast trader an advantage
    - If batch interval is 1 second, a 1 millisecond speed advantage is only $\frac{1}{1000}th$ as likely to matter
  - -> No more arms race

Frequent Batching Directly Solves the Problems with Continuous Trading

Reason 2: Batch auctions eliminate sniping

- Ex: ES jumps at 10:00:00.000am
  - Continuous market: competition manifests in a race to react in SPY market. *Someone is always first*.
  - Batched market: competition in the auction simply drives the price of SPY to its new correct level for 10:00:01.000.
- Notice: both fast and slow liquidity providers protected from sniping
  - Fast liquidity providers: plenty of time to cancel their quotes if there is news
  - Slow liquidity providers: even if something happens in the $\frac{1}{1000}th$ of the batch interval where they don't see it and fast traders do, they are protected by price competition in the auction.
- No more sniping -> improved liquidity

Equilibrium Costs and Benefits of Frequent Batch Auctions

- 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 (MacKenzie article)
  - 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, $t+1$,... (e.g., recent debates re dark pools, PFOF, SIP vs. Direct Feed)

Policy Debates Cleaned Up By Discrete Time

- Clock Synchorinization across exchanges
  - Continuous-time: challenging.
  - Discrete-time: trivial.
- Exchange Message Priority Rules
  - Continuous-time: details of message priority matter. Book updates vs. trade confirmation messages. CME controversy.
  - Discrete-time: issue goes away. plenty of time to disseminate all of the relevant info.
- "Level Playing Field" in access to info
  - Continuous-time: even if in principle info is released to all simultaneously, someone receives / acts on it first. arbitrage rents even from symmetrically observed public information.
  - Discrete-time: restores possibility of meaningfully symmetric information.
- Payment for order flow, Dark Pool debates
  - Continuous time: paper trail makes it hard for investors to know whether they got a fair price, versus a stale price
  - Discrete time: paper trail clean. Easier to discover if exploited.

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 *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 *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)

So, What Next?

- How do we get from continuous-time –> discrete-time?
- Approach 1: private sector innovation.
  - Another Chicago question: if this is such a good idea, why hasn't an exchange already tried it?
  - Potential frictions:
    - Coordination challenge
    - Regulatory ambiguities
    - Vested interests in the current market structure
- Approach 2: regulatory intervention
  - Potential friction: chicken-and-egg problem
    - Regulatory authorities want a high level of proof (rightly so).
    - But, to fully prove the case, someone has to try it first.
- Two things we can hopefully all agree on
  1. Value of a Pilot Test of Frequent Batch Auctions
  2. HFT Data Should Be More Easily Available to Academic Researchers

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. Direct-feed data: continuous-time markets don't actually work in continuous time: correlations completely break down; frequent mechanical arbs; never-ending arms race
2. 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
3. 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

Concluding Thought

There is enormous inertia—a tyranny of the status quo—in private and especially governmental arrangements. Only a crisis—actual or perceived—produces real change. When that crisis occurs, the actions that are taken depend on the ideas that are lying around. That, I believe, is our basic function [as economists]: to develop alternatives to existing policies, to keep them alive and available until the politically impossible becomes politically inevitable.

- Milton Friedman, Capitalism and Freedom