Latency arbitrage, market fragmentation, and efficiency: A two-market model

Elaine Wah and Michael P. Wellman

University of Michigan, Computer Science and Engineering

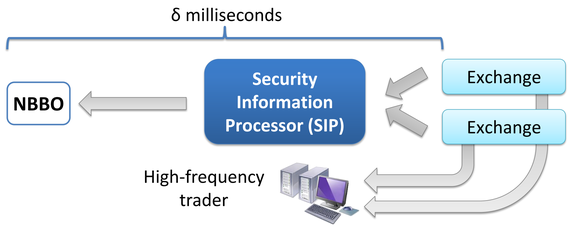
INTRODUCTION

Market fragmentation, where multiple trading venues compete with each other for orders, has become increasingly more prevalent; there are 40+ trading venues for stocks in the U.S. alone. This recent increase has come hand-in-hand with rise of **automated trading**, the use of quantitative algorithms to automate the process of buying and selling.

Fragmentation → *Potential for price disparities across markets*

Regulation NMS was created to mitigate the effects of fragmentation by (1) routing orders for best execution, and (2) creating the Security Information Processor (SIP) to compute and communicate the best price—the **National Best Bid and Offer (NBBO)**—across all exchanges, which it does with some latency (on the order of milliseconds).

Regulation NMS → *Creates exploitable latency advantages*



These advantages can be exploited by **high-frequency trading (HFT)**, characterized by large numbers of small orders with positions held for extremely short periods. Over 50% of total trading volume today is due to HFT, up from 0% in 1995. HF traders compute their own version of the NBBO in less than δ ms, which is before the SIP updates the NBBO. We focus on the HFT strategy of **latency arbitrage**, i.e., exploiting price disparities for nearly risk-free profits. These disparities arise due to market fragmentation and the delay in updating the public NBBO.

Latency arbitrage → *Latency arms race* where HFTs try to compute best prices as fast as possible (e.g., via *co-location*: placing computers as close to the exchange's servers as possible)

Speed advantages exist due to **clearing rules** in stock exchanges where orders are matched as they arrive, as in a **continuous double auction (CDA) market**. Matching orders at fixed intervals (as in a **discrete-time** or **call market**) introduces a delay & eliminates this advantage, as there is no benefit to receiving/responding to market information before others when all orders are processed at the same time. Also, switching to a centralized market would eliminate the effects of fragmentation.

Centralized discrete-time (call) market
prevents exploitation of latency advantages

Our IGERT program (Incentive-Centered Design for Information and Communication Systems) looks at how individual incentives align with system goals. In this context, we investigate how incentives of traders operating at different speeds affect overall efficiency in the market. We also look at the effects of fragmentation and clearing rules.

METHODOLOGY

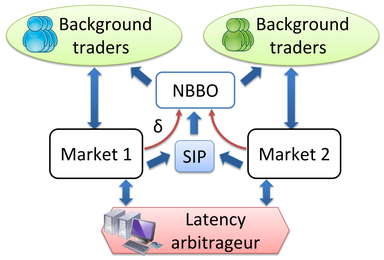
Agent-based modeling

- Allows us to specify agent behavior individually → overall market behavior can emerge over time
- Particularly conducive for modeling interactions between traders, exchanges, and the SIP

Discrete-event simulation

- Answer counterfactual questions
- Facilitate isolation of relationship between fragmentation, clearing rules, and latencies
- Allow precise specification of event occurrences and timing

TWO-MARKET MODEL



Orders are routed to the market offering the best execution, based on a comparison of the NBBO and prices in the alternate market.

Latency arbitrageur (LA)

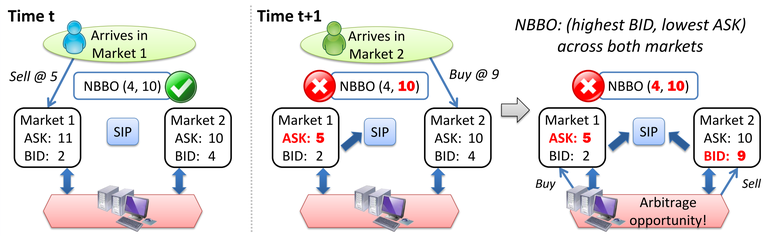
- Infinitely fast
- Arbitrage if market A's highest buy order (**BID**) > market B's lowest sell order (**ASK**)
  - Buy in market B & sell in market A

Background traders

- Private valuation for stock based on a global fundamental
- Zero-intelligence (ZI) strategy
  - Buy or sell 1 unit with probability ½
  - Price offset from private valuation
- Arrives in according to random process

Our two-market model (2M) of a single stock captures:

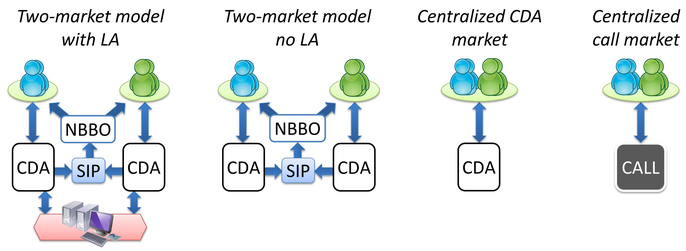
- Communication latencies (between exchanges, SIP, & NBBO)
- Current U.S. regulatory environment (order routing, Regulation NMS)
- Relationship between market fragmentation & latency arbitrage



Example: Latency arbitrage arising in 2M model; bold red price indicates mismatched NBBO

EXPERIMENTS

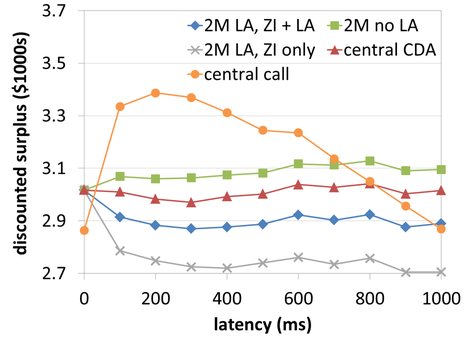
We simulate market models in parallel (background agent population & identical order streams) to isolate the effects of LA & fragmentation.



Our main market performance metric is **surplus** (the gain from trade), which measures **allocative efficiency** or how well resources are distributed to market participants. LA surplus is profit; background trader surplus is gain over the private valuation. We discount surplus by rate ρ to express traders' preference for trading sooner rather than later.

RESULTS

Total surplus results for 200 simulations with 250 background traders:



Comments

As latency ↑, NBBO more out-of-date, & orders more likely to be routed incorrectly

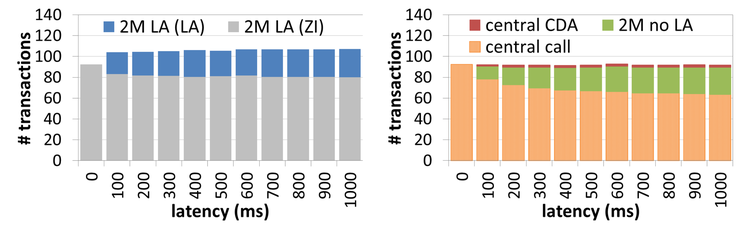
High ρ chosen to exert a strong bias in favor of LA and against periodic clearing

When latency = 0:

- All models generate identical trade sequences
- NBBO is always correct → no arbitrage opportunities & orders are routed correctly

Effect of LA and discrete-time market clearing on efficiency:

- LA takes surplus away from background traders; amount it deducts is greater than the total trading profit it makes → overall surplus ↓
- 2M no LA > central CDA: benefit to fragmentation as it makes inefficient trades less likely, since orders may be routed to the incorrect market
  - — Due to differences in the sequence of orders selected to trade
  - — LA removes this benefit: incorrectly routed orders are removed immed.
- Despite discounting, central call > 2M due to order aggregation over time



Relationship between total number of transactions and surplus:

- 2M LA has most transactions but lowest surplus → other models have higher avg surplus / transaction (since different orders are trading)
- Central call lets orders wait before matching → highest avg surplus/trans

CONCLUSIONS

We introduced a two-market model of latency arbitrage, which we implemented in a system combining **agent-based modeling** and **discrete-event simulation**. We found that:

Latency arbitrage → **degrades total surplus**
(due to differences in the orders selected to trade)

Fragmentation → **some surplus benefit** *(which LA eliminates)*
Centralized call market → **significantly improves efficiency**

ACKNOWLEDGMENTS

This work was supported in part by Grant CCF-0905139 and an NSF IGERT Fellowship through the STIET (Socio-Technical Infrastructure for Electronic Transactions) Program at the University of Michigan.