Market Fragmentation and the Latency Arms Race

Chris Rockwell, Erik Brinkman, Elaine Wah and Michael P. Wellman

University of Michigan, Computer Science and Engineering

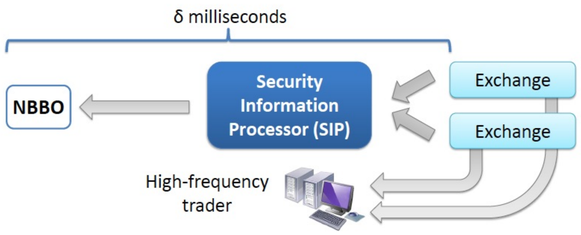
INTRODUCTION

High frequency trading (HFT) is the rapid trading of securities executed by computers. It is important today as it controls over 50% of total trading volume in stock markets[1]. Our research examines a specific type of HFT prevalent in markets today, latency arbitrage.

Latency arbitrage takes advantage of the latency (delay) across markets in attempt to generate income. In other words, Latency arbitrageurs get access to information from multiple markets faster than other traders, and can exploit this to find small price disparities between markets to make nearly risk-free profit. Latency arbitrage has been estimated to account for $21 billion in profit per year[2].

Currently there are over 40 trading venues in the U.S. used to trade stocks, each keeping track of its highest bid and lowest sell prices. Because of this **market fragmentation**, there is the possibility of price disparities across markets. **Regulation NMS** was created to mitigate this, as it routes orders for best execution. This means upon submitting orders, traders have access to all other markets through the **Security Information Processor (SIP).** The SIP has the aggregated lowest sell, or ask price among all markets, as well as the highest bid price, called the **National Best Bid and Offer (NBBO).**

Regulation NMS → *Creates exploitable latency advantages*



Traders face latency in their own market and in the NBBO. Because clearing rules in stock exchanges are in effect a **continuous double auction (CDA) market**, orders are matched as they arrive. Because this type of matching is based on time, latency arbitrage agents can take advantage of their fast speed.

Elaine Wah ran a previous simulation of this scenario using a simple model consisting of two markets, focusing on effects of latency arbitrageurs on **surplus and liquidity**[3]. We are extending her simulation to include delays for latency arbitrageurs, delay for agents in their local market, and multiple latency arbitrage agents. These additions will allow us to analyze competition of latency arbitrageurs over latency advantages.

We study the phenomenon of a **latency arms race**, where HFTs try to keep reducing their latencies, approaching none at all. In reality, HFT firms spend millions of dollars investing in technology to reduce latency, and we model this by looking at revenues in our model to predict how much arbitrageurs value decreasing latency.

METHODOLOGY

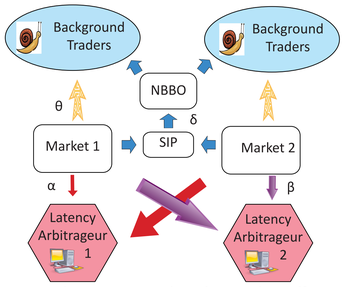
Agent-Based Modeling

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

Discrete-Event Simulation

- Facilitate isolation of relationship between fragmentation, clearing rules, and latencies
- Allow variable latency of information access for different agents to different markets

MODEL



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

Background Traders

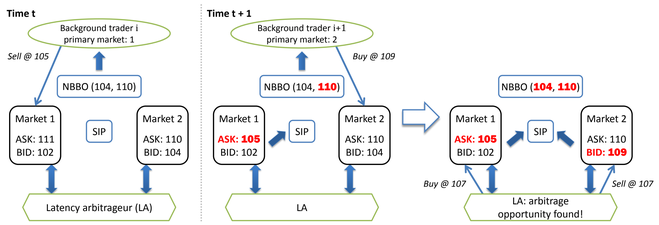
- Access to NBBO with latency δ
- Access to their market with latency θ < δ
- Trade based on private valuation of stocks, using simple strategies

Latency Arbitrageurs

- Access to all markets at a single latency α or β < θ
- Arbitrage if market 1's highest buy order (BID) > market 2's lowest sell order (ASK) or vice versa
  --Buy in market 1 & sell in market 2

Our model of a single equity captures:

- Communication latencies (between exchanges, information processors and traders)
- Current U.S. regulatory environment (order routing, Regulation NMS)
- Relationship between market fragmentation & latency arbitrage
- Competition between latency arbitrageurs



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

EXPERIMENTS

We simulate a two-market model with two latency arbitrage agents, measuring revenues and varying latencies between simulations. The two general scenarios are:

[a]. Both agents at an equal latency
[b]. One agent at a latency less than the other agent

We will compare LA revenue between different latencies and will use **Empirical Game Theoretic Analysis (EGTA)** to make sure fundamental agents respond appropriately to varying HFT strategies.

Our focus is on what reaction each HFT agent will take to different situations and how these decisions will change the shared revenue of the two agents.

HYPOTHESIS

We hypothesize the scenario HFT's face is a prisoner's dilemma in the form of a latency arms race because HFT's will invest in faster technology regardless of the other's decision [i, ii], and they would be better off not doing this [iii]:

i. If the other HFT **does not invest** in faster technology, this HFT has incentive to invest in faster technology to receive all of the revenue, rather than half. Moreover, we predict it to increase in speed only marginally as increasing more than this is unnecessary to capture nearly all arbitrage opportunities.
ii. If the other HFT **does invest** in faster technology, this HFT has incentive to invest so it can capture half of the revenue, rather than none. If possible, it will speed up even faster as it now faces [i].
iii. Assuming cost is nonzero and revenue for arbitrageurs is not significantly correlated with latency when greater than zero, arbitrageurs are best off mutually agreeing to not invest in technology.

	No Investment	Small Investment	Large Investment
No Investment	(½, ½)	(0,1)	(0,1)
Small Investment	(1,0)	(½, ½)	(0,1)
Large Investment	(1,0)	(1,0)	(½, ½)

A simplistic model estimating revenues for the two LAs. Note if we were looking at profit, having an investment would involve a higher cost and thus lower profit. Payoffs are a fraction of total revenue earned using latency arbitrage

FUTURE WORK

In the future, we could look at changing the minimum bid increment to see what effects it has on the system. We could also look at different market model structures, including noting differences between a two call market[4] system and a two CDA market system. An alternative project could analyze more complex latency arbitrage strategies.

CITATIONS

[1] According to Financial Times, 29 July 2009, 73% in "SEC runs eye over high-speed trading."
[2] D. Schneider. The microsecond market. IEEE Spectrum, pages 66 - 81, June 2012.
[3] Elaine Wah and Michael P. Wellman. Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model, February 2013
4 Eric Budish, Peter Cramton, and John Shim. The High-Frequency Trading Arms Race: Frequenct Batch Auctions as a Market Design Response, July '13

ACKNOWLEDGMENTS

This work was supported by NSF Award 1156469, for REU Site: Incentive-Centered Design. PI Erin Lea Krupka.