Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model

Elaine Wah and Michael P. Wellman

University of Michigan
{ewah,wellman}@umich.edu

February 12, 2013

Abstract

We study the effect of latency arbitrage on allocative efficiency and liquidity in fragmented financial markets. We propose a simple model of latency arbitrage in which a single security is traded on two exchanges, with aggregate information available to regular traders only after some delay. An infinitely fast arbitrageur profits from market fragmentation by reaping the surplus when the two markets diverge due to this latency in cross-market communication. We develop a discrete-event simulation system to capture this processing and information transfer delay, and using an agent-based approach, we simulate the interactions between high-frequency and zero-intelligence trading agents at the millisecond level. We then evaluate allocative efficiency and market liquidity arising from the simulated order streams, and we find that market fragmentation and the presence of a latency arbitrageur reduces total surplus and negatively impacts liquidity. By replacing continuous-time markets with periodic call markets, we eliminate latency arbitrage opportunities and achieve further efficiency gains through the aggregation of orders over short time periods.

1 Introduction

Although program trading has been a reality for many years now, the pervasiveness, speed, and autonomy of trading algorithms are reaching new heights. *High-frequency trading* (HFT)—characterized by large numbers of small orders in compressed periods, with positions held for extremely short durations—is estimated to have accounted for as much as 78% of total trading volume in 2009, up from nearly zero in 1995 [Schneider, 2012].[1] The practice of HFT has generated several public controversies regarding its ramifications for transparency and fairness of market operations as well as its effects on market volatility and stability.

The debate has been spurred by recent high-profile events: for example, in August 2012, technology issues in the market-making unit at Knight Capital Group caused a flood of orders for approximately 150 stocks in the New York Stock Exchange. The repeated buying and selling of millions of shares caused dramatic price changes in these stocks, and as a result, all trades executed at 30% higher or lower than the opening price were later canceled [Popper, 2012, Valetkevitch and Mikolajczak, 2012]. Another incident of market turbulence was the so-called "Flash Crash" of May 6, 2010, during which the Dow Jones Industrial Average exhibited its largest intraday decline (approximately 1,000 points). During a five-minute period, some companies traded for as low as a penny and as high as nearly $100,000. The rout continued until an automatic stabilizer on the exchange paused trading for five seconds, after which the markets recovered [Bowley, 2010]. Some have argued that the fragmented nature of current equity markets is to blame for such abrupt and severe price changes [Madhavan, 2011, Golub et al., 2012]. These events and the controversy

[1] Definitive figures are elusive, but proportions exceeding two-thirds are widely reported, for instance 73% in "SEC runs eye over high-speed trading," *Financial Times*, 29 July 2009. This no doubt includes straightforward monitoring for arbitrage opportunities—for example between index securities and their defining constituents, which itself has long represented a large fraction of exchange trading volume.

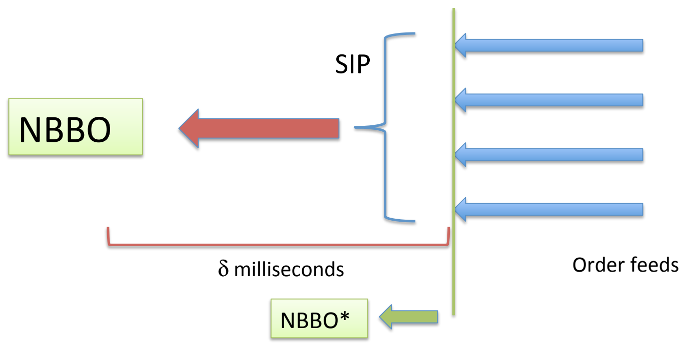


Figure 1: Front-running by exploiting latency differential. Rapid processing of the order stream enables private computation of the NBBO before it is reflected in the public quote from the SIP.

surrounding HFT underscore the necessity of gaining a greater understanding of how high-frequency trading and current market structure affects markets and their participants.

Many high-frequency trading strategies exploit advantages in *latency*—the time it takes to access and respond to market information. Trading on latency advantages has been estimated to account for \$21 billion in profit per year [Schneider, 2012].[2] HF traders achieve such advantages by investing in specialized computer hardware and software, co-locating servers on exchange floors, and in some cases even constructing dedicated communication lines.

The specific HFT strategy we examine here is *latency arbitrage*, where an advantage in access and response time enables the trader to book a certain profit. Arbitrage is the practice of exploiting disparities in the price at which equivalent goods can be traded in different markets. Such disparities can arise in financial markets in several ways, most directly by the fragmentation of securities markets across multiple exchanges. This fragmentation has been a major trend, particularly in the United States over the last decade [Arnuk and Saluzzi, 2012]. U.S. securities regulations have attempted to mitigate the effect of fragmentation through the formulation of Regulation NMS, which mandates cross-market communication and the routing of orders for best execution [Blume, 2007, Securities and Exchange Commission, 2005]. Orders stream into exchanges, which are required to feed summary information about their best buy and sell orders to an entity called the Security Information Processor (SIP). The SIP continually updates public price quotes called the "National Best Bid and Offer" (NBBO).

We illustrate this process and the potential for latency arbitrage in Figure 1. Given order information from exchanges, the SIP takes some finite time, say δ milliseconds, to compute and disseminate the NBBO. A computationally advantaged trader who can process the order stream in less than δ milliseconds can simply out-compute the SIP to derive NBBO*, a projection of the future NBBO that will be seen by the public. By anticipating future NBBO, an HFT algorithm can capitalize on cross-market disparities before they are reflected in the public price quote, in effect jumping ahead of incoming orders to pocket a small but sure profit. Naturally this precipitates an arms race, as an even faster trader can calculate an NBBO** to see the future of NBBO*, and so on.

The latency arms race as sketched above is fundamentally an outgrowth of *continuous trading*: a property of mechanisms that distinguish precedence according to arbitrarily small time differences. By moving to a discrete-time model—which introduces short but finite clearing intervals (as in a *call market*)—we can neutralize small disparities in information access and response time. A driving question of this work is how such a mechanism-design intervention would affect market performance.

More broadly, we seek to understand not only the effects of latency arbitrage on market efficiency and liquidity, but also the interplay between market fragmentation, clearing mechanisms, and latency arbitrage strategies in producing this performance. Such questions about HFT implications are inherently *computational*, as the very speed of operation renders details of internal market operations—especially the structure

[2] Profit figures are considerably more uncertain than volume estimates. Kearns et al. [2010] present an interesting approach to derive an upper bound on HFT profits. Presumably the billions HFT firms invest annually in technology and infrastructure [Adler, 2012] represent a lower bound on gross trading profit.

of communication channels—systematically relevant to market performance. In particular, the latencies between market events (transactions, price updates, order submissions) and when market participants observe these activities become pivotal, as even the smallest differential latency can significantly affect trading outcomes. Lacking suitable data to study these questions empirically,[3] we pursue a simulation approach. Simulation modeling enables us to incorporate causal premises, specifically presumptions of how trading behavior is shaped by environmental conditions.

We propose a simple model that captures the effect of latency across two markets with a single security. Our model is the first to capture the interplay of latency and fragmentation as well as the regulatory environment responsible for current equity market structure, and we have the first results quantifying the effect of latency arbitrage on surplus allocation as a function of latency and market rules. Using an agent-based approach, we simulate the interactions between high-frequency and background traders. Our simulation system allows us to compare the performance of fragmented and consolidated market models under the same underlying order streams. We evaluate efficiency (as measured in terms of total surplus) arising from the simulated orders, under a range of latency settings. Our main finding is that latency arbitrage not only reduces profits of the background traders, but also diminishes surplus overall. Perhaps surprisingly, market fragmentation per se does not harm efficiency; in fact some degree of fragmentation mitigates inefficient trades that are often executed by a continuous mechanism. The discrete-time call market eliminates latency arbitrage by construction and, by virtue of temporal aggregation, yet more effectively matches orders, producing significantly greater surplus.

The paper is structured as follows. In Section 2, we discuss related work on agent-based financial markets and models of HFT and market structure. We describe our two-market model in Section 3. In Sections 4 and 5, we discuss our simulation system and experiments. We present our results in Section 6 and conclude in Section 7.

2 Related work

2.1 Agent-based financial markets

There is a substantial literature on agent-based modeling (ABM) of financial markets [Buchanan, 2009, Farmer and Foley, 2009, LeBaron, 2006], much of it geared to reproduce and thereby explain stylized facts from empirical studies of market behavior. For example, simulated markets have been constructed to reproduce phenomena observed in real stock markets, such as bubbles, crashes, and other high-volatility episodes [LeBaron et al., 1999, Lee et al., 2011]. Because agent behavior is shaped by the market environment, which includes interactions with other agents over time, such models can support causal reasoning (as in the study by Thurner et al. [2012] establishing the effect of leverage on price volatility). ABM has also been used to model financial markets for applications such as portfolio selection [Jacobs et al., 2004] and determining the distributions of order and trading waiting times in a limit order book [Raberto and Cincotti, 2005].

One prominent example of an agent-based financial market is the Santa Fe artificial stock market [Palmer et al., 1994, LeBaron, 2002]. JLM Sim, a discrete-event stock market simulator, is another example; however, it is designed chiefly as a tool for investors rather than for modeling market rules [Jacobs et al., 2004].

2.2 High-frequency trading models

Much of the current literature on the effects of HFT relies on the evaluation of historical order data. Hasbrouck and Saar [2012] use NASDAQ order data to construct sequences of linked messages describing trading strategies. They find that this low-latency activity improves short-term volatility, spreads, and market depth. Angel et al. [2011] conclude that the emergence of automated trading and HFT has improved various market measures such as execution speed and spreads. Additional work suggests a link between HFT and increased

[3]Order activity at the temporal granularity of interest here is generally unavailable for public research, and it is unclear whether data on communication latencies and the end-to-end routing of orders among brokers and exchanges is available from any source. What high-frequency trading data does exist commercially is prohibitively expensive. Moreover, even full details on conceivably observable trading activity could not directly resolve counterfactual questions, such as the response of financial markets to possible shocks or the effects of alternative market rules and regulations.

volatility [Arnuk and Saluzzi, 2012]. In a high-profile study released a few months ago, Baron et al. [2012] find that some kinds of HFT activities directly harm ordinary investors [Popper and Leonard, 2012].

Others rely on theoretical analysis to determine the optimal behavior of high-frequency traders. Avellaneda and Stoikov [2008] derive an optimal limit order submission strategy for a single high-frequency trader acting as a liquidity provider, running numerical simulations to assess the agent's performance under varying strategies. Cohen and Szpruch [2012] propose a single-market model of latency arbitrage with one limit order book and two investors operating at different speeds. The fast trader employs a front-running strategy to determine in advance the quantity the slow investor intends to trade, using this information to generate a risk-free profit.

In a rare application of ABM to HFT, Hanson [2012] finds that market liquidity and total surplus vary directly with the number of HF traders.

2.3 Modeling market structure and clearing rules

Several prior works seek to identify the effects of market fragmentation and clearing rules, mainly via anecdotal evidence elicited from historical data. On the theoretical side, Mendelson [1987] investigates the effect of consolidation versus fragmentation of periodic call markets, without consideration of arbitrage between the submarkets. O'Hara and Ye [2011] use historical quote data and execution metrics to demonstrate that market fragmentation does not appear to harm measures such as spreads, execution speed, and efficiency. Bennett and Wei [2006] compare the execution costs of stocks that have switched from the NASDAQ to the more consolidated NYSE, finding evidence that execution costs decline with order flow consolidation. Amihud et al. [2003] examine the response of equities on the Tel Aviv Stock Exchange to the exercise of corporate warrants, concluding that consolidation improves liquidity. However, none of these prior studies attempt to directly model the communication latencies arising from market fragmentation and the resultant arbitrage opportunities.

Switching to a discrete-time clearing mechanism, as in a call market, has been proposed as a means to eliminate the exploitation of latency differentials across multiple exchanges [The Government Office for Science, London, 2012, Sparrow, 2012]. Empirical work on the effects of such a change is limited and again relies largely on the analysis of historical events. For example, Amihud et al. [1997] investigate the effect of switching from a daily call auction to a combination of a daily call and continuous trading in the Tel Aviv Stock Exchange, finding that this event is associated with improvements in market liquidity.

2.4 Our model in relation to prior work

To study latency arbitrage as made possible by market fragmentation, we develop an agent-based model populated by representative trading strategies interacting within carefully specified market mechanisms. Our model comprises a latency arbitrageur and multiple non-HF traders, with a single security whose trading is fragmented across two markets. Our proposed two-market model is unique in capturing the connections between market fragmentation, communication latencies, regulations, and latency arbitrage. As discussed above, previous analytical or agent-based HFT models employ a single market or order book—rendering them incapable of capturing the effect of fragmentation—and they do not incorporate the communication delays that enable cross-exchange arbitrage.

The focus on accurately modeling communication latencies motivates the study of latency arbitrage and its effect on efficiency and liquidity from a computational perspective. We implement our model in a discrete-event simulation system that captures the processing and information transfer delay in the dissemination of the public NBBO price quote by explicitly modeling the communication patterns between background investors, exchanges, and the SIP operating in current US equity markets.

3 Two-market model

We propose a simple model for latency arbitrage across two markets populated by a single high-frequency trader and multiple background traders. We describe the specifics of this model in Section 3.1. In Sections 3.2 and 3.3, we discuss the behaviors of the latency arbitrageur and background traders, respectively. We present an example of how a latency arbitrage opportunity may arise in this two-market model in Section 3.4.

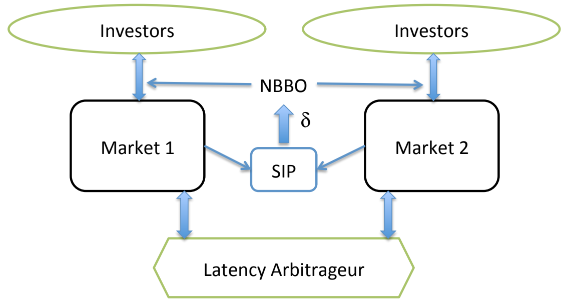


Figure 2: Two-market model with one infinitely fast latency arbitrageur and multiple background investors. A single security is traded on the two markets. Each background investor is associated primarily with one of the two markets, and its order is routed to its alternate market if and only if the NBBO quote indicates an immediate execution. The latency arbitrageur has undelayed access to both markets, so it can immediately detect arbitrage opportunities arising from the delay in NBBO calculation.

3.1 Model description

Our model of latency arbitrage consists of one security traded on two markets, each employing a *continuous double auction* (CDA) mechanism. The CDA is a simple and standard two-sided market that forms the basis for most financial and commodities markets Friedman [1993]. Agents submit bids, or *limit orders*, specifying the maximum price at which they would be willing to buy a unit of the security, or the minimum price at which they would be willing to sell.[4] CDAs are *continuous* in the sense that orders may be submitted at any time. When a new order matches an existing order in the *order book*, the market clears immediately and the trade is executed at the limit price of the incumbent order—which is then removed from the book. A buy order matches and transacts with a sell order when the limits of both parties can be mutually satisfied. CDA markets also continually publish a *price quote* consisting of two parts: The BID quote is the highest-price buy order in the order book, and the ASK quote is the lowest-price sell order. The difference between the two quote components is called the *BID-ASK spread*. A CDA invariant is that $BID < ASK$; otherwise, the orders would have matched and been removed from the order book.

The two markets are linked by a public NBBO signal (see Figure 2). Limit orders lodged in either market are forwarded to the SIP, which calculates and reports an NBBO—based on the quotes from the two markets—with some finite delay up to δ. This latency reflects the time required to receive information about activities in the two markets and compute an updated public price signal.

Retail and institutional investors generate limit orders according to an evolving fundamental (driven by news) and other private factors. Each non-HF investor is primarily associated with one of the markets, to which its orders are routed by default. The determination of the optimal order routing is based on the NBBO quote. The order is sent to the trader's primary market unless the NBBO indicates that the order could be executed in the alternate market at a better price than is currently available on the primary market.

More precisely, let BID_j and ASK_j, where $j \in \{1, 2\}$, denote the current BID and ASK quotes, respectively, in market j. Similarly, let BID_N and ASK_N represent the NBBO quote. Background traders have direct access to the quotes on their primary market and the NBBO, but not to those on the alternate market. Suppose a trader associated with market 1 generates a limit order to buy a unit at price p. This order goes to market 2 if and only if $p \geq ASK_N$ and $ASK_N < ASK_1$. Otherwise, the order is routed to market 1, the trader's primary market. Note that the conditions for submitting to the alternate market entail that the trader's order would execute there immediately, if in fact the NBBO reflects the current global state. If the order is routed to the primary market, it may execute right away (if $p \geq ASK_1$); otherwise, it is added to market 1's order book. The rule for routing sell orders is analogous.

The latency arbitrageur in this model can determine the best prices in each market before the NBBO updates, due to its ability to receive and process order streams faster than background investors. It can thus immediately detect an arbitrage situation, which occurs whenever $BID_1 > ASK_2$ or $BID_2 > ASK_1$. We assume the arbitrageur can respond infinitely fast, so it quickly takes the profit from such arbitrage situations

[4] We assume that there is a limit on the granularity of prices, and thus we represent prices here by integers.

by submitting executable orders to the two markets. Note that the arbitrage opportunity can arise only to the extent that the NBBO information is out of date. If the SIP were able to compute and disseminate the NBBO with zero latency, then a new order would always be routed correctly and would thereby execute if there were a matching order in either market. Any finite delay, however, opens the possibility that an order is routed to the investor's primary market when there is a matching order in the alternate market that had arrived too recently to be admitted in the available NBBO. An out-of-date NBBO can also cause an order to be improperly routed to the alternate market despite it no longer matching there, even if there is a matching order in the primary market.

3.2 Latency arbitrageur

The latency arbitrageur (LA) in the two-market model operates as follows. LA first obtains current price quotes in both markets, then checks whether an arbitrage situation exists. Denote the best price available to sell at by $BID^* \equiv \max\{BID_1, BID_2\}$, and let $ASK^* \equiv \min\{ASK_1, ASK_2\}$ be the best price available to buy. Given a threshold $\alpha \geq 0$, LA deems the current state a worthwhile arbitrage opportunity if and only if $BID^* > (1+\alpha)\,ASK^*$. To execute the arbitrage, LA submits orders exploiting the price differential to the two markets simultaneously. Under our assumption that LA is infinitely fast, bidding any price at or better than the current quote would lead to successful execution at the quoted prices. In our implementation, LA calculates the midpoint m between BID^* and ASK^*, then submits an order to buy at $\lfloor m \rfloor$ to the market with the better ASK price and an order to sell at price $\lceil m \rceil$ to the market with the better BID price. LA surplus (i.e., profit) for these trades is $BID^* - ASK^*$.

3.3 Background traders

Prices in our model are driven by the activity of background investors. We assume a large population of potential investors, who arrive to trade one unit in the market according to a Poisson process with rate λ.

The bid or offer price the agent submits is determined by two components: its underlying *valuation* for the security, which is a product of both fundamental and private factors, and its *trading strategy*, which specifies the price of its limit order.

3.3.1 Valuation model

Each agent possesses a private valuation for the security. This depends on a public (global) fundamental value r_t, which evolves according to a mean-reverting stochastic process (similar to the model of LeBaron [2002]):

$$r_t = \max\left\{0,\ \kappa\bar{r} + (1-\kappa)\,r_{t-1} + u_t\right\},$$

where $\kappa \in [0,1]$ specifies the degree to which the fundamental value reverts back to the mean price $\bar{r}$. The u_t term represents the system-wide shock at time t, which is normally distributed: $u_t \sim \mathcal{N}\left(0, \sigma_s^2\right)$.

The private valuation PV_i for background trader i is simply a perturbed version of the public fundamental at the arrival time $t(i)$ of trader i:

$$PV_i = \max\left\{0,\ d_i\right\},$$

where the deviated value is $d_i \sim \mathcal{N}(r_{t(i)}, \sigma_{PV}^2)$.

3.3.2 Trading strategy

There is an extensive literature on heuristic strategies for trading in CDAs [Friedman and Rust, 1993, Wellman, 2011]. Our investigation employs what is perhaps the simplest strategy from this literature: the aptly named *zero intelligence* (ZI) strategy [Gode and Sunder, 1993]. ZI and related trading strategies have been widely employed in agent-based financial models [Farmer et al., 2005, Paddrik et al., 2012], including MAS studies [Das, 2008, Niu et al., 2010].

A background trader i in our model calculates its private valuation PV_i as described above. It then decides whether to buy or sell a unit of the good (each with probability 1/2). The ZI strategy dictates that the agent submits a bid or offer offset from its valuation by a random amount—essentially the surplus

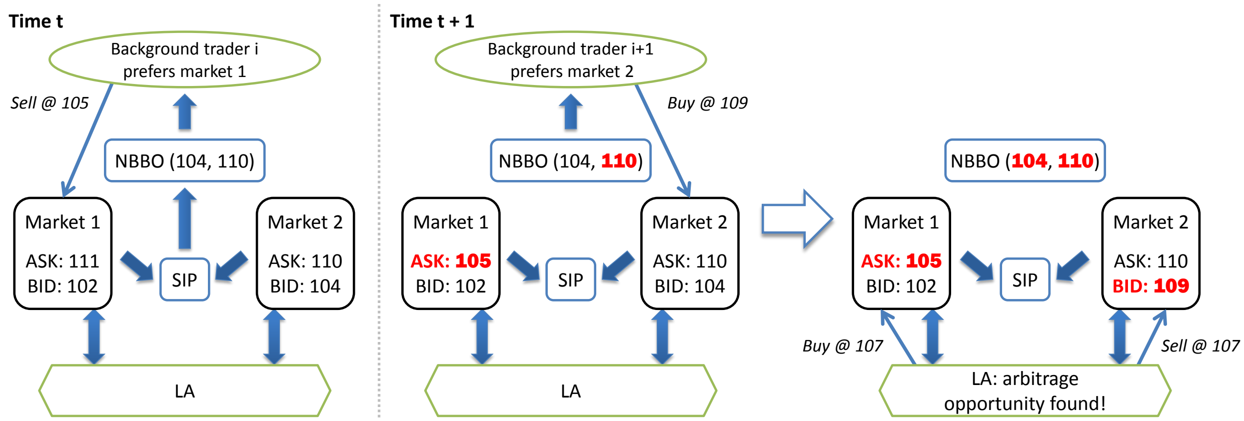


Figure 3: Emergence of a latency arbitrage opportunity over two time steps in our two-market model. All orders are for single-unit quantities. A red, bolded price highlights a discrepancy between the actual market state and the NBBO, represented in the diagram as (BID_N, ASK_N). At time t, the NBBO is up to date. Background trader i wishes to sell at price 105. Since $BID_N < 105$ (which indicates non-immediate execution), the investor's order is routed to market 1. At time $t+1$, the NBBO is out of date, as the SIP updates the public quote with some delay δ. Background trader $i+1$ wishes to buy at 109; based on the NBBO, its order is routed to market 2, its primary market. (Had its order been routed to market 1, its bid would have transacted immediately.) The submission of its order to the inferior market opens up an arbitrage opportunity between the two markets ($BID_2 > ASK_1$), which LA immediately exploits for a guaranteed profit.

the agent seeks from the trade. Let R specify the range of offset values. ZI agent i submits its bid at a nonnegative price $p_i \sim \mathcal{U}\left[PV_i - R, PV_i\right]$ for buy orders or $p_i \sim \mathcal{U}\left[PV_i, PV_i + R\right]$ for sell orders.

To measure market efficiency, we compute total surplus (the sum of buyer and seller surplus) for all background traders. If trader i's limit order transacts at price p_t, it achieves raw (undiscounted) surplus:

$$\begin{cases} PV_i - p_t & \text{for buy transactions, or} \\ p_t - PV_i & \text{for sell transactions.} \end{cases}$$

It follows that the total raw surplus when agent i buys from agent j is $PV_i - PV_j$.

We discount a background trader's raw surplus back to its arrival time at rate ρ, as in the model of Goettler et al. [2009]. The discount is intended to represent not the time value of money (which negligible at this time scale), but rather the traders' general preference for orders to trade earlier rather than later. Such time preference may be due to execution risk, for example, or other costs of delay for related transactions. The raw surplus is discounted by a factor $e^{-\rho T}$, where the execution time T is the difference between transaction time t and the trader's arrival time $t(i)$. For a transaction at time t, the total surplus with discounting is:

$$e^{-\rho(t-t(i))}\left(PV_i - p_t\right) + e^{-\rho(t-t(j))}\left(p_t - PV_j\right).$$

If its limit order never transacts, a trader's surplus is zero.

3.4 Example

Figure 3 illustrates how a latency arbitrage opportunity may arise in our two-market model. At time t, the NBBO quote is $BID_N = 104$ and $ASK_N = 110$. Consider background trader i, who is primarily associated with market 1 and who wishes to submit a sell order at 105. To determine optimal order routing, BID_1 is compared with the NBBO to identify the better market. As $BID_N > BID_1$, the alternate market appears to be superior. However, as an offer to sell at 105 would not transact immediately (since $BID_N = 104$), agent i submits its order to market 1. At the beginning of time step $t+1$, for latency $\delta > 1$, the SIP has not yet updated the NBBO quote to include the order submitted at time t. Therefore, the NBBO available to background investors is out of date: the correct quote would be (104, 105), but the NBBO at time $t+1$ is still (104, 110). This means that the NBBO matches ASK_2 in market 2, the primary market for incoming agent $i+1$. Consequently, agent $i+1$'s buy limit order at price 109 is routed to its primary market. At this point, BID_2 (at price 109, submitted by agent $i+1$) exceeds ASK_1 (at price 105, submitted by agent i), which defines an arbitrage opportunity. Since LA is infinitely fast, it capitalizes on this disparity by submitting bids to buy at 107 in market 1 and sell at 107 in market 2, realizing a profit of 4.

4 Simulation System

The financial markets we study are stochastic dynamic systems with discrete states that change in response to communication events. These events occur at high frequency, and distinctions on the order of milliseconds can be significant. To faithfully model such systems in simulation, ensuring the unambiguous timing of agent and market interactions is paramount. We therefore design our system based on principles of discrete-event simulation (DES), which affords the precise specification of temporal changes in system state. In the DES framework, a simulation run is modeled as a sequence of events. Each event is an instantaneous occurrence that marks a change to the system state at a given time, and events are maintained in a queue ordered by time of occurrence [Banks et al., 2005].

Our DES system simulates the interactions among traders in a set of markets. An *event* in our system consists of a sequence of *activities* that are to be executed by various *entities* (traders, markets, and the SIP). The events are ordered in a priority queue by event time and executed sequentially until the event queue is empty. Multiple events may be scheduled for the same time step, in which case they are executed deterministically in the order in which they are enqueued.

The list of activities within each event is sequenced by priority, and activities with matching priorities are inserted in the order they arrive. Priorities are assigned based on activity type (e.g., bid submission, market clearing). This guarantees determinism in the sequential execution of activities and the correct operation of markets in our simulations. Using this framework, we ensure the latency arbitrageur is infinitely fast by inserting its trading strategy activity at the end of every relevant event (such as a market admitting a new order).

To control the latency of the SIP, we specify three activities: `SendToSIP`, `ProcessQuote`, and `UpdateNBBO`. The `SendToSIP` activity is inserted when a market publishes a quote at time t; upon execution of this activity, the market sends its updated quote to the SIP entity and inserts a `ProcessQuote` and an `UpdateNBBO` activity, both to execute at time $t+\delta$. When `ProcessQuote` is executed, the SIP updates its stored knowledge of the best quotes in the markets. It computes and publishes an updated NBBO based on this information during the execution of the `UpdateNBBO` activity.

Figure 4 illustrates how the activities in our simulation system are sequenced to reflect the communication latencies arising as a consequence of market fragmentation. Market 1 clears and publishes an updated quote at time t_1. Market 2 publishes its new quote at time t_2. For $\delta > t_2 - t_1$, a `ProcessQuote` followed by an `UpdateNBBO` activity are executed sequentially at $t_1+\delta$, as well as at time $t_2+\delta$. The `UpdateNBBO` executing at $t_1+\delta$ does not incorporate market 2's updated quote, as the `ProcessQuote` activity to add market 2's best quote (BID_2, ASK_2) is not executed until $t_2+\delta$. This process serves to model the behavior of the SIP with a delay of δ.

In our simulation system, a *market model* specifies the number of markets, their associated clearing rules, and the population of agents present within the model. To maximize the statistical power of our experimental comparisons, we simulate multiple market models in parallel. Our model-centric system enables the juxtaposition of fragmented and consolidated markets and facilitates the comparison of agent behavior under varying market configurations.

We specify an agent population by describing an arrival process, a process for assigning valuations, and the correspondent trading strategies. In our implementation, a separate pool of background investors are created for each market model under study. We ensure that identical sequences of arrival times and pseudorandom number generator seeds are used to initialize these agents. Since the global fundamental remains consistent across the market models, each ZI agent bid is essentially duplicated within each model. This conveniently allows us to compare the performance of multiple market configurations using the same underlying order stream.

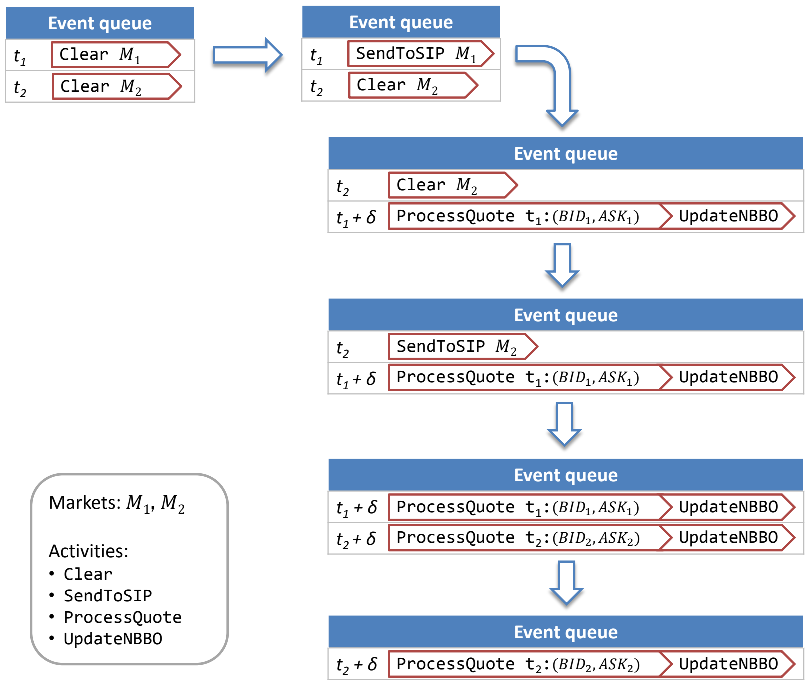


Figure 4: Event queue during the dissemination and processing of updated market quotes for NBBO computation, given latency $\delta > t_2 - t_1$. There are two markets, M_1 and M_2. When the NBBO update activity executes at time $t_1 + \delta$, the SIP has just processed market 1's best quote (BID_1, ASK_1) at time t_1; this is therefore the most up-to-date information that could be reflected in the NBBO at time $t_1 + \delta$.

To isolate the ramifications of market fragmentation, we consider two forms of centralized market models in our simulations: a CDA and a call market. In contrast to a continuous-time market, clearing in a discrete-time or *call market* takes place at designated intervals. A call market eliminates latency arbitrage opportunities, as the periodic clearing mechanism makes it impossible to gain or exploit informational advantages over other market participants within the clearing interval.

5 Experiments

Our experiments evaluate a variety of market configurations with respect to several performance measures. The configurations address the following central issues:

- **Presence of latency arbitrage**: We include configurations of the two-market model with and without LA.
- **Market fragmentation**: Along with the two-market model, we evaluate a centralized configuration where the two markets are consolidated as one.
- **Market clearing rules**: Along with continuous markets, we include a discrete-time call market setting. To facilitate direct comparison, in each run we set the clearing interval of the call market to equal the NBBO update latency.

We are interested in the following performance characteristics:

- **Allocative efficiency**: Total surplus (welfare) is our key measure of market performance.
- **Liquidity**: Markets are liquid to the extent they maintain availability of opportunities to trade at prevailing prices. Two indicators of liquidity are fast execution and tight *BID-ASK* spreads. We

measure *execution time* by the difference in time between order submission and transaction for orders that eventually trade. Execution time is potentially important to investors for many reasons, including the risk of changes in valuation while an order is pending, the effect of transaction delay on other contingent decisions, and general time preference. These factors are reflected in our surplus measures through the discount rate, but a direct evaluation of execution time may also be of interest. We also measure spread, which is the distance between prices quoted to buyers and sellers. Spreads are measured over the first 3000 milliseconds in each simulation, as the majority of background traders arrive within this time.

- **Volatility**: We measure volatility as the log of the standard deviation of midquote prices (as sampled every 250 time steps) over the same interval as spreads.

For each latency setting, we perform 200 simulation runs. The duration of each simulation is 15000 time steps (each step can be interpreted as one millisecond), and there are 250 ZI agents within each market model. An equal proportion of background traders is assigned primary affiliation with each market in a model. In the centralized call market, orders transact at a uniform price each time the market clears. This transaction price is the midpoint between the BID and ASK quotes in the discrete-time market at the time of the clear.

We select environment parameters that generate sufficient arbitrage opportunities. The threshold α for LA is fixed at 0.001. We set the mean fundamental value $\bar{r} = 100,000$, mean-reversion parameter $\kappa = 0.05$, and the variance parameters $\sigma_{PV}^2 = 100,000,000$ and $\sigma_s^2 = 150,000,000$. All bids have single-unit quantities, and we assume zero transaction costs. The range for bid shading by background traders is $R = 2000$. The arrival rate parameter is $\lambda = 0.075$; a ZI agent arrives, on average, every 13 to 14 time steps. All ZI agents submit their limit orders before the end of the simulation. The continuous discount rate ρ is 0.0006 for all background traders. We select this high value of ρ to exert a strong bias in favor of LA and against call markets—the profit of the latency arbitrageur is unaffected by discounting as it is infinitely fast, and call markets impose an inherent delay in trading.

6 Results

We find that the presence of a latency arbitrageur reduces total surplus (Section 6.1) and has a mixed effect on market liquidity, reflected in slightly improved execution times but widened bid-ask spreads (Section 6.2). Eliminating fragmentation with a central CDA lowers spreads while producing surplus and execution metrics between the with and without LA cases. Replacing continuous markets with periodic call markets eliminates latency arbitrage opportunities and achieves substantial efficiency gains (Section 6.3).

6.1 Effect of LA on market efficiency

Figure 5 displays the total discounted surplus, over multiple latency settings, for the centralized CDA and the two-market model with and without a latency arbitrageur. The total surplus of the two-market model without LA, as well as that of the centralized CDA market (an unfragmented continuous-time market), exceeds that of the two-market model with LA, whether or not the profits of LA are counted. In other words, the latency arbitrageur takes surplus away from the background investors, and the amount it deducts exceeds the gross trading profit it accrues.

Note that when latency is zero, the various market models generate identical trade sequences for any given order stream. The NBBO is always correct if there is no delay, so background trader orders are always routed to the right market and no arbitrage opportunities emerge. It follows that the various market models at zero latency produce the same total *undiscounted* surplus. There is a subtle disparity, however, in *discounted* surplus between the CDA and call markets—even at zero latency. CDA trades are executed at the price of the incumbent order, whereas call markets set uniform prices. The pricing rule of the market effectively dictates how surplus is distributed. For a zero-length clearing interval, the call market's uniform price occurs at the midpoint between the incumbent and new matching orders. Since the new matching order clears immediately, only the incumbent order's surplus is discounted; therefore, different ways of distributing the surplus yield different discounted totals. Among the CDA models, the surplus division is the same, so

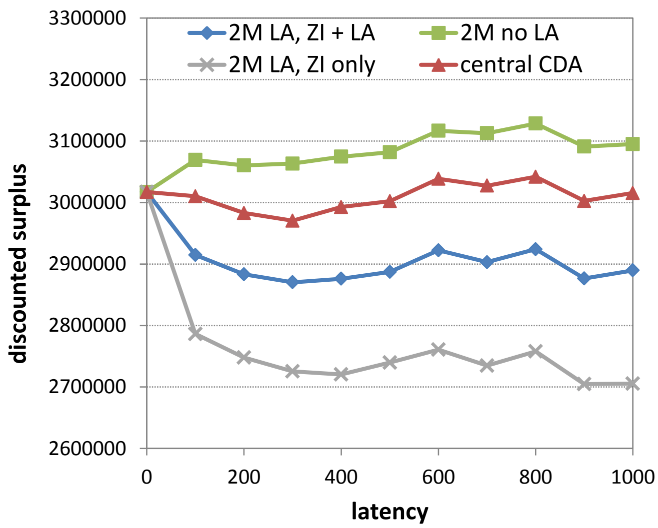


Figure 5: Total discounted surplus in the two-market (2M) model, both with and without a latency arbitrageur, and in the centralized CDA market. In the two-market model with LA, both the total surplus (ZI + LA) and discounted background trader surplus (ZI only) are plotted. The discount rate ρ is 0.0006. Each point reflects the average over 200 runs for each latency setting.

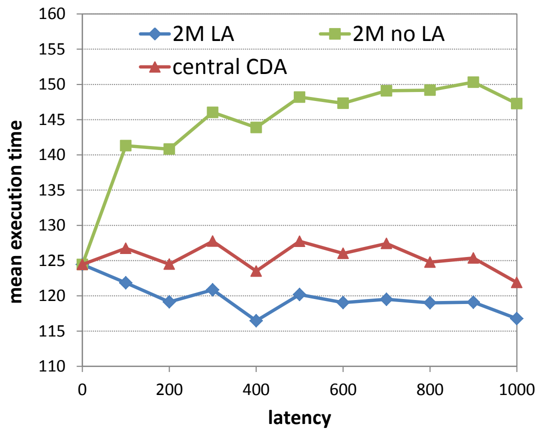
Table 1: P-values for the comparison of total surplus between the centralized CDA/call markets and the two-market model (2M) with and without LA. For instance, the row "CDA vs 2M (LA)" gives the p-values for the superiority in surplus of the centralized CDA market over the two-market model with LA. The p-values are computed by resampling 10,000 times.

Latency	0	100	200	300	400	500	600	700	800	900	1000
CDA vs 2M (LA)	0.4938	0.0015	0.0004	0.0002	0	0	0	0	0	0	0
Call vs 2M (LA)	1.0000	0	0	0	0	0	0	0	0	0.0038	0.7548
2M (no LA) vs 2M (LA)	0.4952	0	0	0	0	0	0	0	0	0	0
2M (no LA) vs CDA	0.5046	0.0350	0.0047	0.0004	0.0035	0.0020	0.0031	0.0027	0.0032	0.0022	0.0046
Call vs 2M (no LA)	1.0000	0	0	0	0	0	0	0.2153	0.9905	1.0000	1.0000

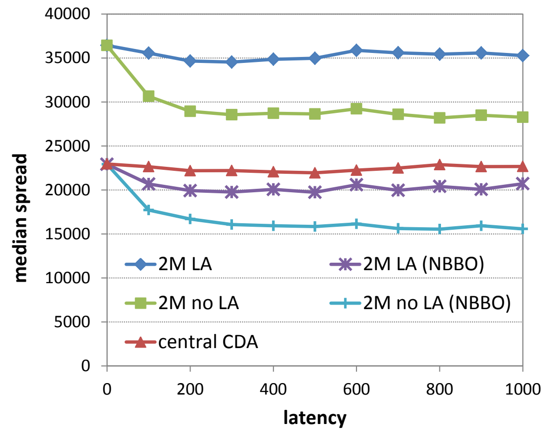
discounting produces the same result. The equality of surplus at zero latency is verified for all four curves in Figure 5, which represent various CDA models simulated in parallel for the same order streams.

We use resampling to compute p-values for: (1) the pairwise differences between the centralized markets and the two-market model, both with and without LA; and (2) the mean difference between the two-market model with and without LA. These results are shown in Table 1. The p-values represent the probability of obtaining surplus differences at least as extreme as those observed if the actual distributions were identical. At zero latency, the p-values between continuous markets are approximately 0.5 because the market configurations behave identically in that setting. The call market surplus at zero latency is significantly lower (hence $p \approx 1.0$), due to the differential effect of discounting noted above. For latencies greater than zero, we find that the differences between the top three curves shown in Figure 5 are all statistically significant. LA degrades efficiency in the two-market model, and centralizing the markets in a combined CDA outperforms the fragmented market with LA.

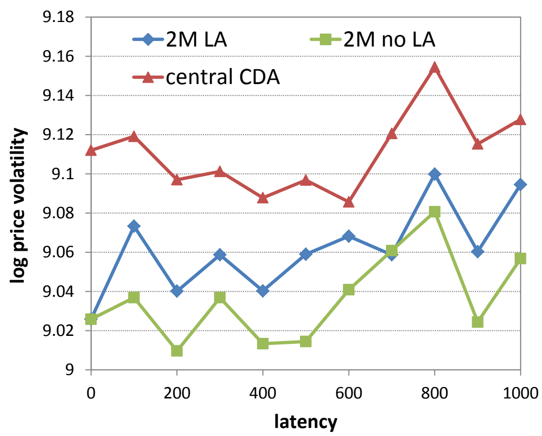
It may seem counterintuitive that the two-market model without LA is significantly better than the centralized CDA. It turns out that for continuous markets, fragmentation can actually provide a benefit, as the separated markets are less likely to admit inefficient trades (i.e., where both traders' values fall on the same side of the longer-term equilibrium price) that arise due to the vagaries of arrival sequences. LA defeats this benefit by ensuring that any orders that would match in the central CDA also trade in the fragmented case, albeit with LA rather than with a counterpart investor.



(a) Mean execution time



(b) Median spread



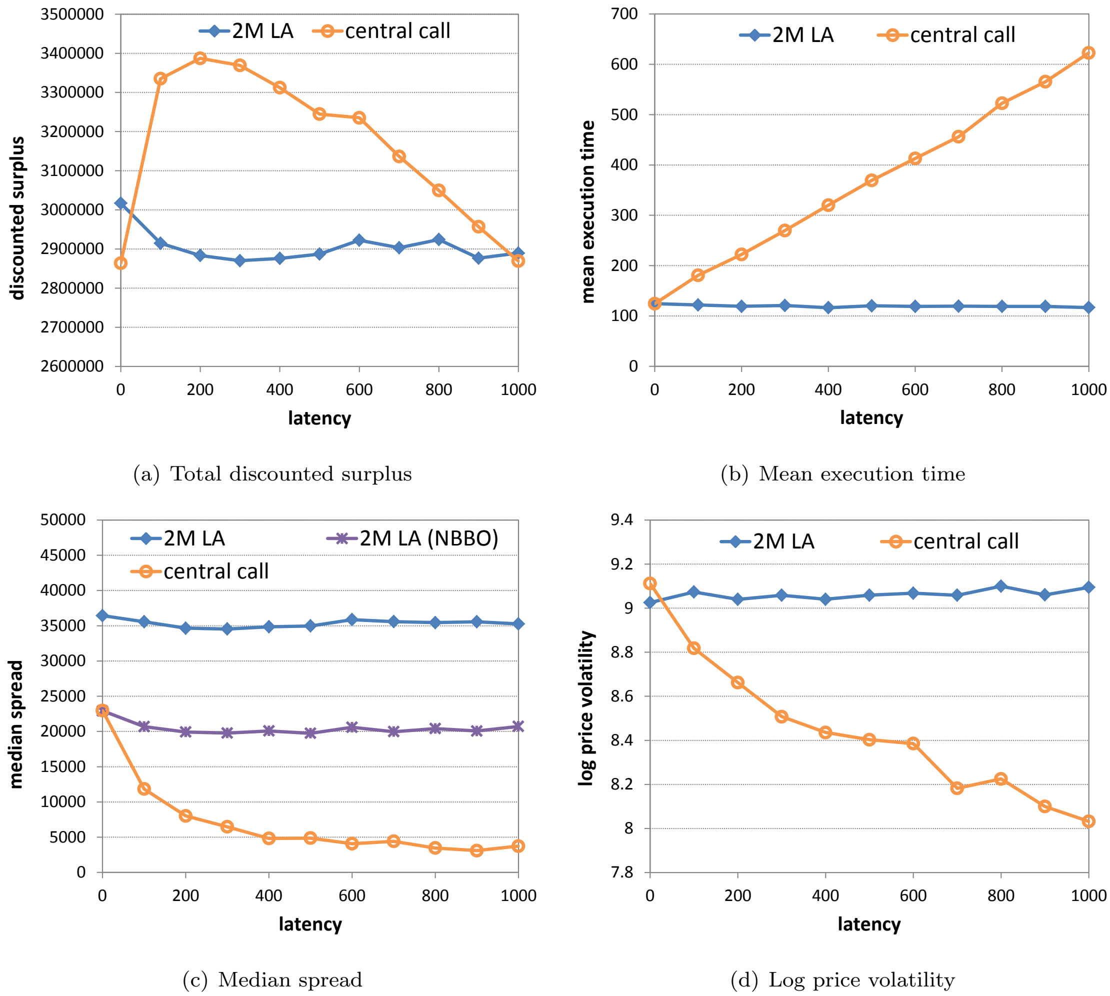
(c) Log price volatility

Figure 6: Mean execution time, median spread, and volatility. Execution time is the difference between bid submission and transaction times, and spread is the amount by which ASK exceeds BID. The spreads in the two-market models (2M) are the average of the individual markets. Price volatility is based on the standard deviation of midquote prices sampled every 250 time steps. Spreads and volatility are measured over a time period of length 3000. Each point reflects the average over all observations for each latency.

6.2 Effect of LA on liquidity and volatility

We also evaluate the effect of latency arbitrage on market liquidity, as measured via execution times and BID-ASK spreads. Figure 6(a) shows that execution time is highest for the two-market model without LA. The fastest trade execution is achieved in the two-market model with LA, which is qualitatively consistent with findings in the literature that trading at lower latencies improves overall execution time [Angel et al., 2011, Garvey and Wu, 2010, Riordan and Storkenmaier, 2012]. The improvement in execution time is at best approximately 30 milliseconds, however, which is generally unobservable by non-HF traders.

Figure 6(b) shows that the highest spreads are those in the two-market model with LA. LA also slightly exacerbates NBBO spreads, which are smaller than spreads of individual markets. The impact of latency arbitrage and market fragmentation on volatility (Figure 6(c)) is minimal, as the differences across the three market configurations are not statistically significant. Overall, LA reduces trading delay at the cost of somewhat widened spreads. The increase in spread could reflect an implicit transaction cost responsible for part of the significant surplus reduction observed above.



(a) Total discounted surplus

(b) Mean execution time

(c) Median spread

(d) Log price volatility

Figure 7: Total discounted surplus, mean execution time, median spread, and log price volatility for the centralized call market and the two-market (2M) model with LA. Each point reflects the average over all observations for each latency setting.

6.3 Discrete-time call market

Lastly, we evaluate the effect of switching to a discrete-time market. Figure 7(a) shows that the total surplus in the centralized call market far exceeds that of the two-market model with LA. By aggregating orders over time, call markets essentially perform a more informed clear, and thus they are less prone to executing inefficient trades than CDAs [Economides and Schwartz, 1995]. Recall that in our call market simulation, the latency setting dictates the clearing period. From the figure, we can see that the surplus of the call market increases dramatically between clearing periods 0 and 100, then peaks at latency 200 before declining steadily. This behavior is a reflection of discounting, which we apply at a high rate ($\rho = 0.0006$) in order to bias against periodic clearing. We select the smallest discount rate such that we obtain lower surplus—within the range of latencies evaluated—in the centralized call market than in the two-market model with LA. At this discount rate, there is an approximately 45% decline in utility for a fixed amount of trade profit, for every additional second of execution time. In other words, an extremely strong preference for small improvements in execution time is necessary before the welfare of the two-market model with LA approaches that of the centralized call market. Even with such steep discounting, the call market significantly outperforms the two-market model with LA for latencies between 100 and 900, dipping to no significant difference at latency 1000 (Table 1). Recall that the difference in total discounted surplus at zero latency is because the call market selects a uniform price for each clear, thereby increasing the incumbent (earlier) bid's share of surplus and consequently reducing total discounted surplus.

As shown in Figure 7(b), the mean execution time in the centralized call market is much higher than that of the two-market model with LA. Unsurprisingly, we find a linear relationship between latency and execution time in the call market model. As market clears occur less frequently in the call market, it takes longer for a bid to match and be removed from the order book. Moreover, as latency increases and the NBBO gets progressively out of date, submitted orders are more prone to be routed to the inferior market. As a result, submitted bids may linger in the order book for a while before a matching order arrives.

In Figure 7(c), we observe that the tightest spread is realized in the centralized call market. The median spread decreases with latency due to the accumulation of bids in the order book, which is indicative of greater liquidity in the market. The temporal aggregation in the centralized call market is also responsible for decreased volatility relative to the two-market model with LA (Figure 7(d)).

6.4 Relationship between transactions and surplus

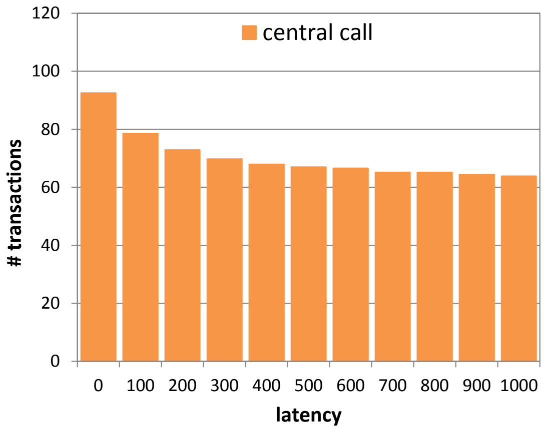
Figure 8 shows the total number of transactions for each market model, averaged over all observations at a given latency. In Figure 8(a), the number of transactions in the centralized call market declines as latency increases; this corresponds to the slowdown in surplus gains at higher latencies for any additional delay. The number of transactions in the centralized CDA (Figure 8(b)) and the two-market model without LA (Figure 8(c)) are comparable, though slightly lower in the latter. This is consistent with our observations of surplus patterns in Figure 5. The two-market model without LA results in higher surplus despite a reduction in number of transactions, indicating that each transaction in the fragmented model is associated with more surplus on average than in the centralized CDA. Figure 8(d) shows a breakdown of the number of transactions attributed to the background ZI agents (light bar) and LA (dark bar). The number of LA transactions increases with latency, as the number of arbitrage opportunities grows as the NBBO update delay increases. Whereas the total number of transactions with LA is higher than those in the other market models, the average surplus per transaction is considerably lower. This provides further evidence that trades in the fragmented markets with LA are often outside the efficient set.

7 Conclusion

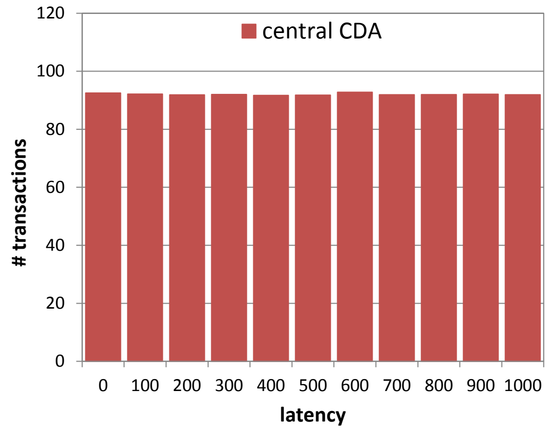
To understand an important phenomenon in high-frequency trading, we introduced a two-market model of latency arbitrage. We implemented this model in a system combining agent-based modeling and discrete-event simulation in order to evaluate the interplay of latency arbitrage, market fragmentation, and market design, as well as their consequences for market performance. Our results demonstrate that market efficiency is negatively affected by the actions of a latency arbitrageur, with no countervailing benefit in liquidity or any other measured characteristic. This key finding holds even given an extreme temporal discount factor, which we imposed in order to tip the scales to favor shorter execution intervals. Taking into consideration the substantial operational costs of the latency arms race would only amplify our conclusions about the harmful implications of this practice.

Virtually all modern financial markets employ continuous trading, which enables speed-advantaged traders to make risk-free profits over fragmented markets and which degrades overall efficiency. Our proposed alternative is a discrete-time call market, which eliminates latency arbitrage opportunities and improves efficiency. A call market prevents high-frequency traders from gaining a latency advantage, thereby increasing surplus for background traders. Aggregating orders over small, regular time intervals provides additional efficiency gains, and in fact these benefits appear to overshadow the gains attributable specifically to neutralizing latency arbitrage.

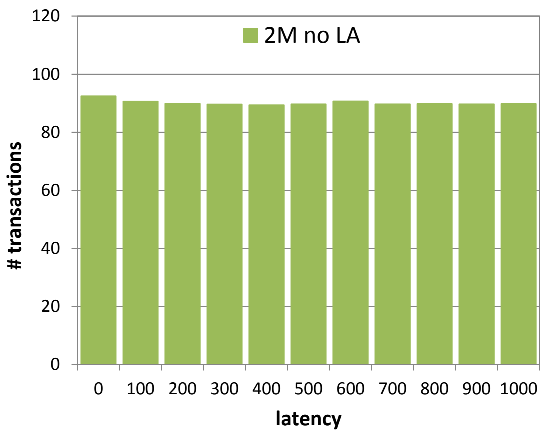
Our model offers a tool to policymakers and other researchers to more rigorously evaluate financial market rules. We believe it can play a constructive role in the debate around HFT and market structure, and we invite others—including those who may have reservations regarding the conclusions here—to propose either alternative scenarios or structural elements that could be incorporated within our general framework. As for any simulation model, our results are valid only to the extent our assumptions capture the essence of real-world markets, and we are eager to explore extensions that would test the limits of our conclusions. For example, the current model relies on an exceedingly simple characterization of trader behavior, and it



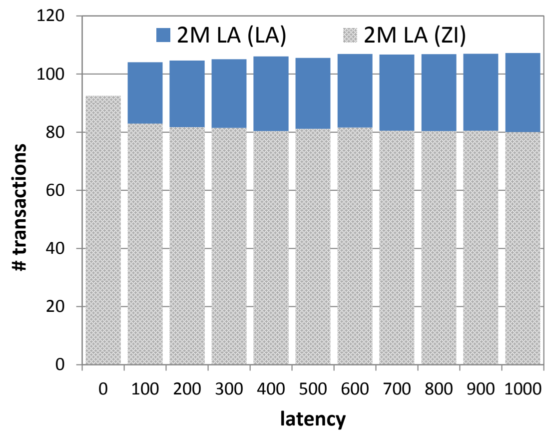
(a) Centralized call



(b) Centralized CDA



(c) 2M (no LA)



(d) 2M (LA)

Figure 8: Total number of transactions in each market model. In Figure 8(d), the light bar represents the total number of ZI transactions, and the dark bar represents LA transactions. Each bar reflects the average over all observations for each latency setting.

considers a limited range of regulatory mechanisms and responses. Additional avenues for further study include examining the impact of more sophisticated HFT and background trader strategies (such as those using historical information or responding to LA price signals) as well as the effect of introducing other types of traders such as market makers. It would also be interesting to examine interactions between multiple HFTs employing differing strategies and to evaluate this model from a game-theoretic perspective.

Acknowledgments

An early version of the two-market model was conceived in collaboration with Akshat Kaul. This work was supported in part by Grant CCF-0905139 and an IGERT Fellowship from the U.S. National Science Foundation.

References

J. Adler. Raging bulls: How Wall Street got addicted to light-speed trading. *Wired Magazine*, Aug. 2012.

Y. Amihud, H. Mendelson, and B. Lauterbach. Market microstructure and securities values: Evidence from the Tel Aviv Stock Exchange. *Journal of Financial Economics*, 45(3):365–390, 1997.

Y. Amihud, B. Lauterbach, and H. Mendelson. The value of trading consolidation: Evidence from the exercise of warrants. *Journal of Financial and Quantitative Analysis*, 38(4):829–846, 2003.

J. J. Angel, L. E. Harris, and C. S. Spatt. Equity trading in the 21st century. *Quarterly Journal of Finance*, 1(1):1–53, 2011.

S. L. Arnuk and J. C. Saluzzi. *Broken Markets: How High Frequency Trading and Predatory Practices on Wall Street are Destroying Investor Confidence and Your Portfolio.* FT Press, 2012.

M. Avellaneda and S. Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3): 217–224, 2008.

J. Banks, J. S. Carson II, B. L. Nelson, and D. M. Nicol. *Discrete-Event System Simulation.* Prentice Hall, fourth edition, 2005.

M. Baron, J. Brogaard, and A. Kirilenko. The trading profits of high frequency traders. Technical report, Commodity Futures Trading Commission, 2012.

P. Bennett and L. Wei. Market structure, fragmentation, and market quality. *Journal of Financial Markets*, 9(1):49–78, 2006.

M. E. Blume. Competition and fragmentation in the equity markets: The effect of Regulation NMS. *SSRN Electronic Journal*, pages 1–18, 2007.

G. Bowley. U.S. markets plunge, then stage a rebound. *The New York Times*, 2010.

M. Buchanan. Meltdown modelling. *Nature*, 460(August):680–682, 2009.

S. N. Cohen and L. Szpruch. A limit order book model for latency arbitrage. *Mathematics and Financial Economics*, 6:211–227, 2012.

S. Das. The effects of market-making on price dynamics. In *Seventh International Conference on Autonomous Agents and Multi-Agent Systems*, pages 887–894, Estoril, Portugal, 2008.

N. Economides and R. A. Schwartz. Electronic call market trading. *Journal of Portfolio Management*, 21 (3):10–18, 1995.

J. D. Farmer and D. Foley. The economy needs agent-based modelling. *Nature*, 460(7256):685–686, Aug. 2009.

J. D. Farmer, P. Patelli, and I. I. Zovko. The predictive power of zero intelligence in financial markets. *Proceedings of the National Academy of Sciences*, 102(6):2254–2259, 2005.

D. Friedman. The double auction market institution: A survey. In Friedman and Rust [1993], pages 3–25.

D. Friedman and J. Rust, editors. *The Double Auction Market: Institutions, Theories, and Evidence.* Addison-Wesley, 1993.

R. Garvey and F. Wu. Speed, distance, and electronic trading: New evidence on why location matters. *Journal of Financial Markets*, 13(4):367–396, 2010.

D. K. Gode and S. Sunder. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy*, 101(1):119–137, 1993.

R. L. Goettler, C. A. Parlour, and U. Rajan. Informed traders and limit order markets. *Journal of Financial Economics*, 93(1):67–87, 2009.

A. Golub, J. Keane, and S.-H. Poon. High frequency trading and Mini Flash Crashes. *SSRN Electronic Journal*, pages 1–22, 2012.

T. A. Hanson. The effects of high frequency traders in a simulated market. In *Midwest Finance Association Annual Meeting*, 2012.

J. Hasbrouck and G. Saar. Low-latency trading. Technical Report 35-2010, Cornell University Johnson School of Management, 2012.

B. I. Jacobs, K. N. Levy, and H. M. Markowitz. Financial market simulation. *Journal of Portfolio Management*, 30(5):142–152, Jan. 2004.

M. Kearns, A. Kulesza, and Y. Nevmyvaka. Empirical limitations on high-frequency trading profitability. *Journal of Trading*, 5(4):50–62, 2010.

B. LeBaron. Building the Santa Fe artificial stock market, 2002.

B. LeBaron. Agent-based computational finance. In L. Tesfatsion and K. L. Judd, editors, *Handbook of Agent-Based Computational Economics*, pages 1187–1233. Elsevier, 2006.

B. LeBaron, W. B. Arthur, and R. Palmer. Time series properties of an artificial stock market. *Journal of Economic Dynamics & Control*, 23(1):1487–1516, 1999.

W. B. Lee, S.-F. Cheng, and A. Koh. Would price limits have made any difference to the 'Flash Crash' on May 6, 2010. *The Review of Futures Markets*, 9:55–93, 2011.

A. Madhavan. Exchange-traded funds, market structure and the Flash Crash. *SSRN Electronic Journal*, pages 1–33, 2011.

H. Mendelson. Consolidation, fragmentation, and market performance. *Journal of Financial and Quantitative Analysis*, 22(2):189–207, 1987.

J. Niu, K. Cai, S. Parsons, P. McBurney, and E. Gerding. What the 2007 TAC market design game tells us about effective auction mechanisms. *Autonomous Agents and Multiagent Systems*, 21:172–203, 2010.

M. O'Hara and M. Ye. Is market fragmentation harming market quality? *Journal of Financial Economics*, 100(3):459–474, June 2011.

M. Paddrik, R. Hayes, Jr., A. Todd, S. Yang, P. Beling, and W. Scherer. An agent based model of the E-Mini S&P 500 applied to Flash Crash analysis. In *IEEE Conference on Computational Intelligence for Financial Engineering and Economics*, pages 1–8, 2012.

R. G. Palmer, W. B. Arthur, J. H. Holland, B. LeBaron, and P. Tayler. Artifical economic life: A simple model of a stock market. *Physica D: Nonlinear Phenomena*, 75(1):264–274, 1994.

N. Popper. Flood of errant trades is a black eye for Wall Street. *The New York Times*, 2012.

N. Popper and C. Leonard. High-speed traders profit at expense of ordinary investors, a study says. *New York Times*, Dec. 2012.

M. Raberto and S. Cincotti. Modeling and simulation of a double auction artificial financial market. *Physica A: Statistical Mechanics and its Applications*, 355(1):34–45, Sept. 2005.

R. Riordan and A. Storkenmaier. Latency, liquidity and price discovery. *Journal of Financial Markets*, 15 (4):416–437, 2012.

D. Schneider. The microsecond market. *IEEE Spectrum*, pages 66–81, June 2012.

Securities and Exchange Commission. Regulation NMS, 2005. 17 CFR Parts 200, 201, 230, 240, 242, 249, 270.

C. Sparrow. The failure of continuous markets. *Journal of Trading*, 7(2):44–47, 2012.

The Government Office for Science, London. Foresight: The future of computer trading in financial markets, 2012.

S. Thurner, J. D. Farmer, and J. Geanakoplos. Leverage causes fat tails and clustered volatility. *Quantitative Finance*, 12:695–707, 2012.

C. Valetkevitch and C. Mikolajczak. Error by Knight Capital rips through stock market. *Reuters*, 2012.

M. P. Wellman. *Trading Agents.* Morgan & Claypool, 2011.