Technological Disparity and Its Impact on Market Quality*

Kiseo Chung†
Texas Tech University

Seoyoung Kim‡
Santa Clara University

Abstract

Technological investments made by speed-sensitive market participants are increasingly frequent and have thus been a focal point of recent research. We examine an important, but unexplored facet of this trend: the technological disparity between the fastest market participants and the exchange itself. Using a proprietary dataset of a high-frequency market maker's limit orders and order acknowledgments timestamped to the nanosecond, we explore the consistency and reliability of an exchange's ability to discern the correct sequence of orders when messages are submitted in rapid (sub-microsecond) succession. We find a high degree of variability in acknowledgment times, and the proportion of times in which the first order entered is also first to be acknowledged is surprisingly low when consecutive orders are placed at very high frequencies. Furthermore, we provide evidence of impaired market quality as a result. These issues remain pertinent even following substantial technological improvements made by the exchange, because of the ongoing technological disparity between the exchange and the fastest market participants, who continue to competitively invest in technological improvements.

JEL Classification: G1; G2
Keywords: technological disparity; price/time priority; queuing uncertainty; order imbalance; market microstructure.

* We thank Robert Bartlett, Rick Cooper, Sanjiv Das, David Denis, Diane Denis, Prachi Deuskar, Patrick Flannery, Joel Hasbrouck, Frank Hatheway, Terrence Hendershott, Sam Lee, Scott Richard, Walter Tackett, Daniel Trepanier, Ingrid Werner, and seminar participants at 2019 Allied Korea Finance Associations, Baruch University, CFMR 2020, Emory University, IFABS 2019 Medellin, Indian School of Business, the Journal of Investment Management (JOIM) 2017 Fall Conference, Purdue University, Santa Clara University, Showcasing Women in Finance 2018 (@ Miami University), and Xangrila Inc (formerly, Xambala Inc).We especially thank Daniel Trepanier for technical support and insights without which this paper could not come together.
† Rawls College of Business, Texas Tech University; 2500 Broadway; Lubbock, TX 79409. Email: kiseo.chung@ttu.edu
‡ Leavey School of Business, Santa Clara University; 500 El Camino Real; Santa Clara, CA 95053. Email: srkim@scu.edu

1. Introduction

Technological innovations have dramatically transformed the landscape of equity trading in the last decade, engaging academics and regulators alike. Given the ever-increasing speed at which high-frequency traders (HFTs) and liquidity providers are able to submit orders, numerous studies have naturally explored the direct ramifications and externalities arising from the speed-arms race *among* market participants. However, a particularly prominent yet unattended issue in this high-frequency era concerns not only the differing speeds among high-frequency players but also the technological disparity between the exchange itself and its fastest market participants. That is, although exchanges are financially motivated to reduce latency (Li, Ye, and Zheng, 2023), an exchange's infrastructure in place may not be able to keep up with the increasing frequency at which orders are submitted.

For the speed-sensitive liquidity provider, one critical consideration in providing immediacy is predicated on the important assumption that a given exchange or trading venue can correctly acknowledge orders in a timely and methodical fashion. Specifically, in a continuous market based on price-time priority, limit orders of the same price should be ranked (and filled) based on the time at which they are submitted. However, software architectures that achieved adequate levels of performance in a less competitive era now appear to induce randomness to what should ideally be a deterministic process of acknowledging orders in the same sequence in which they are sent. Thus, a natural and highly important question arises as to (i) whether, empirically, time priority is being violated and (ii) whether there are negative consequences for other participants in the trading ecosystem as a result. Our purpose is to document this heretofore unexplored phenomena and to provide evidence of the resulting market externalities.

Using a proprietary dataset of a specific high-frequency market maker's limit orders and acknowledgements, which are timestamped to the nanosecond, we are able to first document this timely and important issue regarding the consistency and reliability of an exchange's ability to acknowledge and correctly rank orders in practice. Because we definitively know the sequence in which we place

orders from our own co-located, dedicated OUCH ports, we can track how often the orders placed first are actually first to be acknowledged and queued by the exchange.

We find that when orders are placed in rapid succession, the proportion of orders placed first that are also first to be acknowledged (i.e., the *FIFO ratio*) can be surprisingly low depending on the time deltas we allow between consecutive messages. For instance, when consecutive orders are placed at least 16 microseconds (μs) apart in time, 99% of orders sent first are also first to be acknowledged and queued. However, at lower latencies of less than one microsecond (μs) between orders, only half of the orders sent first are also first to be acknowledged (i.e., the *FIFO ratio* is 59%). Interestingly, we observe that the *FIFO ratio* for time deltas of two μs improves significantly after a major technological overhaul by the exchange,[1] but the *FIFO ratio* for time deltas of one μs remains dismally low. That is, our evidence suggests that technological disparity persists between the exchange and its fastest participants, who continue to outpace the exchange as they competitively invest in technological improvements. This ongoing disparity, in turn, perpetuates queuing uncertainty and randomness in time priority for high-frequency liquidity providers and hence, increases their risks and costs of providing immediacy to other market participants.

To examine potential market externalities arising from this phenomenon, we begin by exploring the implications for perceived liquidity/depth in the limit-order book, specifically as it pertains to excess messaging and rapid order cancellations. That is, liquidity providers must race for queue position to add liquidity to the continuously evolving limit-order book upon each new price formation. With the added uncertainty arising from queuing uncertainty and violations in price-time priority, we expect liquidity providers to strategically submit orders in excess of the liquidity they actually intend to provide at a given price level, since (in the continuous market) they can rapidly

[1] Specifically, we leverage the launch of the Nasdaq Financial Framework (NFF) on May 26, 2016. See Nasdaq Press Release accessed on <https://www.nasdaq.com/about/press-center/nasdaq-debuts-groundbreaking-nasdaq-financial-framework-enhancing-operations>.

cancel a number of these orders once their respective queue positions in the limit-order book have been assigned.

Focusing on time deltas of one μs between consecutive orders, which proxies for the ongoing technological disparity between market participants and the exchange for our sample period, we find that the *FIFO ratio* is a substantial predictor of the percentage of rapid-fire order cancellations. For instance, an increase in the *FIFO ratio* from 60% to 90% is associated with a 1.81% decline in order cancellations occurring within 50 μs of a price-formation speed race to add liquidity, which represents a 9.64 percent decline from the average rapid-fire order cancellations of 18.78%. Similarly, we find that an increase in the *FIFO ratio* is associated with a significant decline in the ratio of the total quantity of shares placed at the onset of a new price-formation relative to the quantity of shares available halfway throughout the life of a newly established price level. Overall, the excess messaging in response to queueing uncertainty persists even after a major technological upgrade by the exchange, suggesting that the exchange has not contemporaneously matched the latest technological improvements made by its fastest market participants.

Given that liquidity providers in the continuous market can mitigate the consequences they face from queuing uncertainty (i.e., through excess messaging and subsequent order cancellations), we now turn to explore whether this uncertainty has palpable ramifications on other market participants outside of the fastest, first-in-line liquidity providers who directly experience the costs. A particularly risky but potentially lucrative time for liquidity providers on Nasdaq falls within the last ten minutes of the trading day, wherein the closing cross is held alongside the continuous market. During this ten-minute period, the closing-cross active interest alerts market participants to the evolving demand for liquidity at the close (i.e., the liquidity-seeking on-close orders to buy or sell a particular security at the official closing price of the day). In response, liquidity providers can enter Imbalance Only (IO) orders, which is a type of limit order that offsets the unmatched on-close orders placed by other market participants. Given that the closing cross is an important market mechanism that sets the daily Nasdaq

Official Closing Price (NOCP) for each security, the resulting demand to purchase or sell shares at the NOCP provides a potentially substantial reward to liquidity providers who are first in line with their IO orders.

However, in contrast to orders placed in the continuous market, liquidity providers cannot cancel their IO orders once placed after 3:50 PM ET, and moreover, do not see their queue position in the closing cross until settlement at 4:00 PM ET. Thus, liquidity providers are faced with additional risks when participating in the closing cross due to the uncertainty regarding the extent to which they should pre-emptively accumulate an offsetting position in the final minutes of the continuous market. That is, if a liquidity provider intends to absorb an on-close buy imbalance, then he/she would ideally accumulate an offsetting position by purchasing shares or otherwise creating an offsetting hedge prior to the end of the trading day. If, after these efforts, his/her IO order is ultimately left unfilled, then he/she is forced to hold inventory overnight and is subject to overnight price volatility.

To complicate matters, the fastest liquidity providers are unable to iteratively learn their predictive fill rates relative to their high-frequency competitors, if the queueing uncertainty they face arises from the exchange's technological inability to consistently acknowledge and queue orders in the correct sequence (rather than from uncertainty as to how fast other players are). As a result, the uncertainty and risks faced by liquidity providers who place IO orders in the closing cross are greatly exacerbated by randomness in time priority, because the market features of the closing cross preclude liquidity providers from engaging in the excess messaging and rapid-fire cancellations they employ in the face of queuing uncertainty in the continuous market. Thus, a natural question arises as to whether greater randomness in time priority results in more on-close orders being left unfilled.

We find that the aggregate end-of-day order imbalance (i.e., the percentage of unabsorbed on-close orders at the end of the trading day) and the number of tickers with unabsorbed on-close orders has increased substantially over time. Interestingly, after accounting for other fundamental factors contributing to order imbalance, we find that the *FIFO ratio* is a substantial predictor of the unabsorbed

imbalance at market close. For instance, an increase in the *FIFO ratio* from 60% to 90% is associated with a 0.345% decline in aggregate order imbalance, which represents a 35.9 percent decline from the average market-wide percentage order imbalance of 0.96%. Moreover, we observe that smaller-cap stocks, which are inherently more difficult to hedge in after-hours trading, suffer disproportionately relative to large-cap stocks. That is, a similar increase in the *FIFO ratio* is associated with a 1.21% decline in the aggregate order imbalance of stocks outside of the top quintile with respect to market capitalization. These results suggest that greater uncertainty surrounding whether IO orders will be acknowledged in the correct sequence in which they are placed makes liquidity providers more reluctant to submit IO orders in the first place.

Overall, our paper is the first to provide evidence of a heretofore unexplored byproduct of disproportionate technological advances made by market participants versus those made by the exchange, leading to ever lower latencies in equity trading that cannot be appropriately distinguished. Because this aspect of queuing uncertainty arises from the technological disparity between the exchange and the fastest market participants, it is not something that can be alleviated unless the exchange contemporaneously matches the pace of technological improvements made by market participants. Moreover, we provide evidence that this other (and more harmful) cause of queuing uncertainty that arises from violations in price/time priority has greater implications for market liquidity and daily market clearing. This aspect of queuing uncertainty is critical to ongoing discussions pertaining to optimal market design, as it will persist as long as the fastest market participants not only outpace other participants but also outpace the capability of the exchange to distinguish who is faster.

This paper is organized as follows. In Section 2, we provide a literature review along with an overview of the relevant technical details of modern market structure. In Section 3, we describe the data and methodology, and we provide summary statistics. In Section 4, we present the empirical analyses as to the *FIFO ratio* and aspects of market quality. In Section 5, we discuss potential solutions and policy implications, and in Section 6, we provide concluding remarks.

2. Background on Market Structure: Literature Review and Technical Details

In this section, we begin with a review of prior work studying market quality and structure in a high-frequency era. We then provide an overview of technical details pertaining to market design, specifically as it pertains to differences between the continuous market and closing cross, as well as details regarding an exchange's technological structure.

2.1 Literature Review

The Securities and Exchange Commission (SEC) has identified high frequency trading as "one of the most significant market structure developments in recent years",[2] citing estimates that more than half of total trading volume is attributable to high frequency trades. For reference, each trading day presents at least 8 million speed races to market participants, who seek to either take liquidity, provide liquidity, or cancel posted orders throughout the day. To win these speed races, rapid advances in technology have dramatically increased the rate at which both liquidity providers and liquidity takers (i.e., traders) can enter orders, with consecutive orders oftentimes placed microseconds, and now, even nanoseconds apart. Accordingly, a plethora of work has emerged to study the ramifications and potential policy implications arising from the ascent of high frequency trading and the speed-arms race among market participants. [3]

On one hand, many studies have identified negative externalities from increasing HFT activity.[4] For instance, HFTs can create mispricing that disadvantages ordinary investors (Jarrow and Protter, 2012), and a HFT's ability to gain advance access to information imposes adverse selection costs on other slower market participants (Biais, Foucault, and Moinas, 2015). Theoretical models also

[2] See the Securities and Exchange Commission's Concept Release on Equity Market Structure (17 CFR Part 242), released on Thursday, January 21, 2010.
[3] See Biais and Foucault (2014) and O'Hara (2015) for an overview of issues arising with the advent of high frequency traders (HFTs).
[4] Note, however, that liquidity measurement problems arising from HFT activity can materially affect inferences drawn from empirical work (Holden and Jacobsen, 2014).

demonstrate that increased speed by HFTs can lead to lower information production and price informativeness (Baldauf and Mollner, 2020; Huang and Yueshen, 2021). Moreover, the rise of high-frequency trading may adversely impact liquidity for other players, as suggested by the decline in the consolidated depth across trading venues associated with a greater presence of HFTs (Kervel, 2015) and the lower adverse selection and trading costs associated with exogenous disruptions to traders' speed advantages (Shkilko and Sokolov, 2020). Overall, competition for queue position encourages the high-frequency arms race (Yao and Ye, 2018), and generally, there appears to be an overall "socially excessive" investment in speed (Hoffmann, 2014; Biais, Foucault, and Moinas, 2015).

On the other hand, studies have also found positive ramifications. For instance, Brogaard, Hendershott, and Riordan (2014) provide evidence that HFTs improve price discovery and market efficiency, and Conrad, Wahal, and Xiang (2015) provide evidence that high-frequency quotation leads to lower trading costs and price paths that better resemble a random walk. Similarly, studies on colocation services provided by exchanges, which further reduce latency, provide evidence of lower spreads (Boehmer, Fong, and Wu, 2015; Frino, Mollica, and Webb, 2014) and increased overall liquidity (Brogaard, Hagströmer, Nordén, and Riorda, 2015) once colocation services are introduced. Consistent with these empirical observations, Foucault, Hombert, and Roşu (2016) present a model in which HFTs not only contribute to short term increases in trading volume but also long-term price discovery.

Finally, in contrast to the studies examining the impact of accelerated technological upgrades made by market participants, others have focused on the impact of technological upgrades made by the exchange itself. Riordan and Storkenmaier (2012) provide evidence that an exchange's technological upgrades to increase speed led to reduced quoted spreads and enhanced price discovery, and Kemme, McInish, and Zhang (2022) provide evidence that an exchange's speed improvements additionally led to less manipulative trading behavior. Similarly, Pagnotta and Philippon (2018) theoretically model the impact of competition among trading venues to attract traders, finding that imposing a minimum

speed requirement across venues reduces market-wide inefficiencies whereas imposing a maximum speed requirement does not. Moreover, Li, Ye, and Zheng (2023) provide evidence that slower exchanges are more costly to traders and provide worse execution, which in turn disincentivizes exchanges from slowing down trades.

Overall, though, amidst discussions of optimal market design in a high-frequency world and the purported impact of a technological arms race, the potential mismatch in technological capabilities between the exchange and its high-frequency market participants (and the ensuing market externalities) has, heretofore, remained empirically unattended. Studies thus far take for granted that queuing uncertainty is an unavoidable consequence of technological advances and market design (e.g., Yueshen, 2014; Budish, Cramton, and Shim, 2015). That is, these studies assume that multiple traders react simultaneously to new information, and thus are prioritized randomly by the exchange, which cannot process messages simultaneously. However, in reality, orders are virtually impossible to be placed simultaneously. Rather, the speed at which high-frequency traders react to new information has dramatically accelerated over time, which renders outdated technology unable to discern the fine but distinctive difference in origination times between orders placed just nanoseconds apart.

Thus, in stark contrast to prior literature in this space, our focus is to: (i) provide evidence of ongoing technological disparity between an exchange and its market participants using proprietary data from a high-frequency market maker, and (ii) provide suggestive evidence of the resulting impairment in market quality. Along this regard, Menkveld and Zoican (2017) is the closest paper to ours as it theoretically derives how exchange speed affects market quality through the actions of high-frequency market participants. In particular, the authors demonstrate that a faster exchange, in certain cases, might not lead to better liquidity. Our paper bridges an important gap by empirically documenting metrics that are indicative of the technological disparity that persists between an exchange and its market participants even following substantial speed improvements made by the exchange.

2.2. Overview of Market Design: Distinctions between the Continuous Market and Closing Cross

In this section, we provide critical details in an exchange's market design to highlight key features contributing to the various risks faced by liquidity providers, which in turn, result in less liquidity for other market participants. Consistent with our data (presented in Section 3) and empirical analyses (presented in Section 4), we focus on the specifics of the National Association of Securities Dealer Automated Quotations Sock Market, often simply referred to as NASDAQ or Nasdaq.

In the continuous market, often simply referred to as "the stock market," trading hours run from 9:30 AM to 4:00 PM Eastern Time on Monday through Friday, except on specially designated holidays and half days. Liquidity providers have the flexibility to continuously add and remove quotes from the central limit order book throughout the trading day. Furthermore, by subscribing to the Nasdaq TotalView-ITCH data feed, liquidity providers see the evolving limit order book and trade executions continuously throughout the trading day. Importantly, in the continuous market, liquidity providers also continuously learn their queue positions (typically within 43 μs during our sample period) on orders submitted throughout the trading day and are free to cancel orders upon receiving the order acknowledgement. Thus, in the continuous market, liquidity providers can strategically place more limit orders than they plan to fulfill, since they have an easy way to mitigate the risks born from queueing uncertainty.

On the other hand, the closing cross, which occurs alongside the continuous market during the last ten minutes of the trading day from 3:50 PM to 4:00 PM Eastern Time, is a specific market setting that determines the Nasdaq Official Closing Price (NOCP) each day for each Nasdaq-listed security. For reference, a security's daily closing price is used as the reference price for determining index valuations, net asset values (NAVs) for funds, to mark brokerage accounts, and to determine account margins. Thus, the closing cross represents a critical market mechanism in which a substantial portion

of daily volume occurs,[5] providing a potentially lucrative but particularly risky opportunity for liquidity providers.

Because the on-close demand to buy versus sell shares of a particular security results in either a buy-side or sell-side order imbalance, liquidity providers can respond to the unmet demand for liquidity at the close by entering Imbalance Only (IO) orders, which is a type of limit order that offsets the unabsorbed on-close orders placed by liquidity-seeking market participants. To facilitate this process, the closing-cross active interest is disseminated every five seconds, between 3:50 and 4:00 PM Eastern Time, in the form of Net Order Imbalance Indicator (NOII) messages.[6]

However, in stark contrast to orders placed in the continuous market, where liquidity providers learn their queue positions and can cancel orders accordingly, IO orders placed in the closing cross cannot be canceled and queue positions are not revealed until settlement at 4:00 PM Eastern Time. That is, in the closing cross, liquidity providers are unable to employ their self-protective strategy of excess messaging followed by rapid order cancellations. Thus, a natural question arises as to whether greater randomness in time priority due to technological disparity causes liquidity providers to be less willing to absorb on-close market demand.

The risks to liquidity providers arising from queuing uncertainty are particularly pronounced in the closing cross, because they are forced to hold an unexpected position overnight if an IO order is ultimately left unfilled at the close. For instance, if a liquidity provider intends to absorb an on-close buy imbalance, then he/she will pre-emptively accumulate an offsetting position by purchasing shares or otherwise creating an offsetting hedge in the final minutes of the continuous market. If his/her IO order is ultimately left unfilled, then he/she is forced to hold inventory overnight and is subject to

[5] For instance, in the first quarter of 2017, the average volume demanded at the close was approximately 40,000 shares per ticker, and the average proportion of total daily volume filled at the close was 4.2 percent.

[6] The timing of messages has since changed, with NOII information now being broadcasted every 10 seconds from 3:50 to 3:55 PM ET then every second from 3:55 to 4:00 PM ET. Thus, we expect a greater incidence of concentrated IO orders to occur in the recent period over the last five minutes of the closing cross.

overnight price volatility. This risk is particularly magnified when a speed-sensitive liquidity provider encounters randomness in time priority, which creates difficulties in learning his/her predictive fill rate relative to other liquidity providers.

2.3. Overview of Market Design: Opting for a Continuous Versus Discretized Market

Recent studies have questioned the merits of keeping a continuous market design in a high-frequency world, advocating instead that orders be processed batch by batch at larger time intervals. That is, to reduce the ever-increasing investments for seemingly marginal improvements in speed and to attenuate the asymmetry arising from these sub-microsecond-level differences in speed across players, one advocated solution is to treat all orders arriving within the same discrete-time bucket equally with respect to time priority (see, for instance, Budish, Cramton, and Shim, 2015). However, under this discretized market design, traders have no incentive to submit new orders or to withdraw older ones at the beginning of the trading interval, since all orders of the same price receive equal priority as long as they are placed within the same time bucket, $(T_i, T_i + h]$. As a result, although frequent batch auctions prevent an arms race to be first, they instead promote a reverse race to be last (i.e., as feasibly close to $T_i + h$), since traders prefer to process as much information as possible prior to submitting or canceling orders.[7]

Ultimately, this critical time period within the batch interval will be determined by the technological capabilities of the traders, which continues to encourage investments in speed since the fastest traders will be able to execute orders closest to $T_i + h$. That is, faster traders will have a greater information set on which to base their orders for a given trading interval *i* than slower traders who are forced to execute orders earlier in the trading interval. Moreover, faster traders will still be able to

[7] See, for instance, Haas and Zoican (2016). We note that, even in a discretized market design, technological disparity poses an issue, if the exchange is incapable of distinguishing which orders were placed sub-microseconds prior to the end of each interval.

"snipe" stale quotes from slower traders who are unable to cancel a quote based on information that is received too close to $T_i + h$. Thus, frequent batch auctions are unlikely to attenuate issues of adverse selection or "excessive" investment in speed, since information still arrives continuously rather than in discrete intervals, and the high-frequency arms race remains irrespective of whether an exchange chooses a continuous or discretized market design. Overall, discretized markets are not immune to problems arising from technological disparity between the exchange and its fastest players.

2.4. Overview of Design Topology: Technical Details Regarding Order Gateways and Matching Engines

The preferred structural design used by the U.S. equities markets, including Nasdaq, is a hub-and-spoke architecture, where the matching engine sits at the center of equidistant gateways which manage individual customer connections. The matching engine is responsible for keeping track of limit orders to buy and sell each stock and matches them against orders priced at a market (or marketable) price. The activity going through the matching engine is tracked for clearing and settlement purposes. The exchange also provides a summary of market activity in the form of market data for all participants to use in real time.

Order gateways manage individual client sessions. Each gateway has multiple order ports, and, in turn, order ports are assigned to clients for their use in managing communication with the exchange. The order ports allow clients to send outbound messages to the exchange, such as add-order requests or cancel-order requests. The ports also provide communication from exchanges in the form of order acknowledgements, order rejections, cancel confirmations, and executions. Exchanges perform checks on their gateways to ensure that incoming messages are properly formed. The gateway also throttles messages as necessary by either outright rejecting or decreasing the pace of orders when message rates

exceed pre-determined thresholds, and passes on properly formed (i.e., compliant) orders to the matching engine as it receives them.

This design topology presents two important benefits: the distributed architecture is both (i) reliable, and (ii) highly scalable. As the number of symbols or message traffic increases, an exchange can add gateways and/or matching engines as necessary. Thus, if one gateway fails, clients can migrate to an alternate gateway. Exchanges also have disaster recovery plans which allow them to switch market activity to a backup facility located a significant distance away from their primary trading location.

However, despite its distributed nature, a significant problem remains in this architectural design, whereby multiple locations where messages can encounter a bottleneck in processing. As these queues lengthen, the processing time for each order increases and the client experience departs from the ex-ante expectation of the price-time priority that should be honored in an ideal (i.e., first come, first serve) setup. For instance, consider a situation in which a single client places an order to a port on a gateway server. The gateway checks the order and passes it on to the exchange's matching engine, which in turn responds with an order acknowledgement for a seamlessly executed roundtrip of order placement to acceptance. In contrast, consider a speed race in which multiple clients rapidly send orders, causing congestion at the gateway's outbound network connection to the exchange's matching engine. As a result, a queue forms at the incoming matching engine connection, and orders may no longer be acknowledged by the matching engine in the sequence in which they were placed.

2.5. A Tale of Two Olympians: "FIFO" in a competitive sports setting

We close out this section with an analogy whereby the 100M race at the Olympics corresponds to a speed race on an exchange, the Olympic referees represent the exchange's gateways and matching engine, and the fastest swimmers correspond to high-frequency market participants. That is, the

referees employ stopwatches/cameras to rank swimmers, which represent the technology employed by an exchange to accept and queue orders.

Much like the competition among high-frequency liquidity providers has led to ever decreasing time deltas between orders, the competition between Olympic swimmers has led to vast technological improvements in suits, training, and dietary plans in order to shave off fractions of seconds in performance times. However, if the quality of the stopwatch/camera is insufficient to differentiate performance times with very small time deltas, the faster swimmer may not be appropriately recognized as such. In other words, the rank ordering of swimmers could be different from the true order at which they finish if their time deltas are too fine to be properly captured by less sophisticated stopwatches and cameras with lower frame rates. In these circumstances, trying to reverse-engineer capabilities based on reported performances times could mislead the strategic planning of athletes and their trainers for subsequent competitions. Below is an excerpt from an article regarding poor FIFO ratios in the 1960 Olympics:

> "Watches capable of discerning hundredths of a second were in regular use in the Olympics by 1948. But what good is such refinement if, when an athlete crosses the finish line, the judge drops a tenth of a second or more merely clicking the stopwatch? (Human thought takes time to propagate and enact, too.) The weakness of this link became terribly apparent during the 1960 Summer Olympics, in Rome, when two swimmers, the American Lance Larson and the Australian John Devitt, seemingly tied in the hundred-metre freestyle. A half-dozen judges, peering through the waves at the finish, reached a stalemate: three declared Larson the winner, the other three Devitt. Though Omega's stopwatches indicated that Larson had the faster time, by at least a tenth of a second, a referee broke the tie and awarded Devitt gold." (Burdick, 2018)

3. Data

In this section, we describe our main empirical metrics and data sources. Appendix A1 provides a comprehensive and consolidated account of all variables and corresponding sources.

3.1. Sources

Our sample consists of 3,740 unique tickers and spans the first trading day of January 2014 through the last trading day of April 2017, which totals 2,911,692 ticker-days. We obtain time-stamped orders and acknowledgements from a proprietary dataset created and culled with a specific high-frequency market maker on Nasdaq -- which we describe in further detail below -- to calculate order-to-accept latencies and, ultimately, the first-in-first-out ratio of orders submitted in rapid succession.

We augment our proprietary dataset with the raw exchange data feed from Nasdaq (i.e., TotalView–ITCH), which provides the entire evolving limit order book for each stock (timestamped to the nanosecond), Net Order Imbalance Indicator (NOII) message data, reference prices, near prices, closing prices, and continuous-market trading volume as well as the information required to determine speed races, order cancellations, and the half-life quantity of a given price formation from the raw exchange data feed from Nasdaq (i.e., TotalView–ITCH).

The main advantage of using the Nasdaq ITCH data over commonly used TAQ data is that the ITCH data contains true prevailing quotes, which allows us to properly track the life of an order. In contrast, Daily TAQ Quotes data provides the inner-most quotes and associated depths as provided by each participating exchange aggregated to the millisecond (or nanosecond) and cannot distinguish arrival rates and cancellation times of specific limit orders submitted to an exchange.[8] Thus, Daily TAQ Quotes provides the aggregated real-time information seen by market participants, such as brokerages and other financial institutions, who seek best execution across the highly fragment equity-

[8] Despite decreasing the latency at which quotes are aggregated to nanoseconds, this inherent difference persists between the two data sources.

trading space, whereas the Nasdaq ITCH feed provides more finely tuned real-time information critical to the functioning of market makers and high-frequency liquidity providers on Nasdaq. Moreover, unlike the ITCH feed, TAQ does not provide ongoing NOII information throughout the closing cross.

3.2. Measuring O-A Latencies and FIFO Ratios

Leveraging our vantage point from a specific high-frequency market maker, we are able to place orders in rapid succession and capture when orders are acknowledged and queued. These orders are placed from co-located, dedicated OUCH ports, which are not shared with other market makers. This proprietary data-collection process allows us to definitively capture the proportion of orders placed first that are also first to be acknowledged based on the time deltas between order placements.

The *order-to-accept (O-A) latency* captures the distance in time from when an order is placed to when it is officially acknowledged by the exchange (in this case, by Nasdaq). The O-A latency is measured by a server clock that is synchronized to precision clocks to ensure reliability in employing the server clock to time stamp the data. On average, we capture O-A latencies across 1.5 million messages daily.

The *first-in-first-out (FIFO) ratio* captures the percentage of cases in which orders are accepted by the exchange in the correct sequence. That is, the *FIFO ratio* represents the proportion of times in which the first order placed is also first to be acknowledged. To calculate this ratio, we begin by examining pairs of orders originating from two different dedicated OUCH ports.[9] Depending on the distance in time between orders (measured in nanoseconds), we organize the order pairs along the following time-delta bins of interest: (0, 100), (100, 250), (250, 500), (500, 1000), (1000, 2000), (2000, 4000), (4000, 8000), (8000, 16000), (16000, 32000), (32000, 64000).

[9] Queueing uncertainty predominantly arises from orders placed across ports rather than from orders placed within the same port. That is, time priority is easily maintained when sending multiple orders from the same order port, but the question remains as to whether time priority is maintained across ports, which is the focal point of our study.

For each order pair within a given bin, we ascertain the time at which each order is sent and the time stamp at which the exchange acknowledges the orders. Specifically, time stamping occurs at the exchange's matching-engine, which determines the time priority and queue position for orders in the resulting limit-order book. Finally, within each time-delta bin, we count how many orders were acknowledged in the same time sequence as the order in which they were sent. We call this metric the *FIFO count*, and we divide the *FIFO count* by the total number of orders sent to arrive at the *FIFO ratio*.

3.3. Determining Speed Races and Excess Messaging by Way of Rapid Order Cancellations

Each trading day presents millions of speed races by traders who seek to either take, provide, or remove liquidity from the limit-order book. To focus on the impact of queuing uncertainty on the behavior of market participants, we take the perspective of a liquidity provider who must race for queue position to add liquidity to the limit-order book upon each new price formation. Queue position in the limit-order book is an important consideration to liquidity providers, and significant resources are deployed to secure a front position in the queue each time a new price level forms. With the added uncertainty arising from violations in price-time priority, we expect liquidity providers to submit orders in excess of the liquidity they ultimately intend to provide at a given price level, since (in the continuous market) they can rapidly cancel a number of these orders once their actual queue position in the limit-order book has been acknowledged.

To identify the number of such liquidity-adding speed races each day, we focus on securities priced over $1.00 with a minimum tick size of one cent,[10] and we identify instances in which the formation of a new bid (or offer) is at a price that was previously an offer (or a bid), where the bid-ask

[10] A Tick Size Pilot Program was approved for a two-year time frame beginning October 3, 2016, wherein a test group of tickers would be quoted and traded in minimum increments of five cents. For additional details, refer to the FINRA website: https://www.finra.org/industry/tick-size-pilot-program

spread is one cent. Intuitively, because stocks with a spread greater than one cent can be price-improved, such price formations are not considered "races". We also do not include instances where a bid or offer volume disappears and subsequently backfills at the same price, since a motivated participant could have posted more volume at that price in the first place. We provide a graphical example of these scenarios in **Figure 1**.

For each speed race to provide liquidity, we count the number of orders and quantity of shares added upon formation of each price level. We then measure the natural quantity that dealers are willing to provide by the quantity of shares for each price level halfway through the life of the price level (see **Figure 2** for a graphical representation). We use this quantity to calculate an *Inverse Half-Life Ratio*, which represents the total quantity of shares placed at the onset of the new price-formation speed race scaled by the half-life quantity of shares. In addition, we track the average percentage of orders placed at a new price-formation speed race that are cancelled within 50 μs to measure the extent of excess messaging that occurs at each speed race. We focus on order cancellations within 50 μs since the median O-A latency (i.e., acknowledgement time) during our sample period is 43 μs, and orders canceled long after the queue position is acknowledged may be information-based liquidity withdrawal (rather than a withdrawal due to intentional, ex-ante excess messaging at price formation). To set ideas, in **Figure 3**, we provide an example of an actual price and depth formation for symbol INTC (i.e., Intel Corporation) on May 31, 2017.

3.4. Summary Statistics

In **Table 1**, we present summary statistics on the basic characteristics of our sample. The average *FIFO Ratio* at a time delta of less than one μs is 59.31% (i.e., the average rate at which the first order placed is also first to be acknowledged by the exchange when consecutive orders are placed less than one μs apart is approximately 59%). There are approximately 2.8 million speed races per day to add liquidity at a new price formation, and on average, 18.8% of liquidity-adding orders placed at the onset of a new

price-formation speed race are cancelled within 50 μs of placement. In addition, we observe an average *Inverse Half-Life Ratio* of 3.52, which suggests that approximately 3.5 times as many orders are typically placed at the onset of a new price-formation speed race than the natural steady-state depth of the book for that price level. Furthermore, we observe that the average market-wide (i.e., aggregate) end-of-day order imbalance is 0.96%,[11] with an average total on-close demand at approximately 36,600 shares per ticker. The average daily share volume based on trades executed in the continuous market is approximately 935,000 shares per ticker.

As demonstrated in **Table 2**, a substantial portion of daily volume occurs at the closing cross. On a typical (i.e., median) day, the average portion of daily volume that occurs at the close is 4.88% (per ticker) for Nasdaq-listed securities. This average is as high as 34.84% across all tickers during our sample period. **Table 2** also reports the average ticker-level end-of-day imbalance on a typical (i.e., median) day; specifically, on a typical day, the average portion of on-close order imbalance that is unabsorbed by the final NOII message is 4.45% (when comparing percentage order imbalances across individual tickers, as opposed to calculating an aggregate market-wide percentage order imbalance as in **Table 1**).

4. Empirical Results

4.1. Order-to-Accept (O-A) Latency and FIFO Ratio over Time

To examine fluctuations in the order-to-accept (O-A) latency, we begin by plotting the daily median O-A latency from January 2014 through May 2017. The results, which we present in **Figure 4**, show notable intra-day variation (i.e., jitter) over time. Although the jitter is less dramatic in the later part of

[11] The aggregate end-of-day order imbalance is calculated based on the total unabsorbed orders from the last NOII message of the day as a percentage of the total on-close demand (across all tickers with an initial imbalance as of the first NOII message at 3:50:00 PM).

the sample, we continue to observe non-negligible differences in O-A times on both a day-to-day and intra-day basis.

To examine the practical consequences of the observed jitter, we also examine the proportion of times in which the first order entered is also first to be acknowledged by the exchange's matching engine (i.e., the *FIFO Ratio*). The results, which we plot in **Figure 5a**, shows that the *FIFO Ratio* is surprisingly low when orders are placed in the rapid succession. Specifically, when consecutive orders are placed less than one μs (i.e., one microsecond) apart in time, only 59% of the orders sent first are also first to be acknowledged. However, at greater latencies of ten μs between orders, the *FIFO Ratio* in price/time priority is roughly 96%, and reaches 99% at latencies of at least 16 μs between orders.

In **Figures 5b** and **5c**, we also plot the five-day moving average *FIFO Ratio* over time, and we observe substantial variation in the proportion of times in which the first order entered is also first to be acknowledged. For instance, in examining the *FIFO Ratio* over time when consecutive orders are placed up to two μs apart (**Figure 5b**), we observe that the average FIFO Ratio ranges from 48.07% to 91.58% [untabulated]. Furthermore, we observe that the *FIFO Ratio* markedly improves over time, with a distinct upward trajectory following the launch of the Nasdaq Financial Framework (NFF) on May 26, 2016, which entailed a particularly intense series of technological upgrades.[12] In contrast, in examining the *FIFO Ratio* over time when consecutive orders are placed less than one μs apart (**Figure 5c**), we observe that the average FIFO Ratio ranges from 47.50% to 76.33% [untabulated]. Most notably, for these time deltas of < 1 μs between consecutive orders, the *FIFO Ratio* does not appear to improve materially with the passage of time, hovering at an average *FIFO Ratio* of 59.56% in 2017

[12] This "groundbreaking" technological overhaul differed significantly from routine improvements made by Nasdaq as it entailed major upgrades to its architecture with a pronounced focus on the technical capabilities of the algorithmic matching, processing, and execution of orders by its Matching Engine (*Nasdaq Debuts Groundbreaking Nasdaq Financial Framework, Enhancing Operations for Over 100 Market Operators Globally*; May 16, 2016; Nasdaq Press Release accessed on <https://www.nasdaq.com/about/press-center/nasdaq-debuts-groundbreaking-nasdaq-financial-framework-enhancing-operations>).

[untabulated]. To explore the implications of this structural break while accounting for other contemporaneous factors in a multivariate setting, we estimate the following OLS regressions:

$$Dep\,Var_t = \alpha + \beta_1 \cdot Post\,Upgrade_t + X_t \cdot \gamma + \varepsilon_t \quad \textbf{(1)}$$

We explore three different dependent variables: (i) *Median OA Latency* $_t$, which represents the daily median distance in time from when an order is placed to when it is officially acknowledged by Nasdaq, and (ii) *FIFO Ratio (< 2 μs)* $_t$ and (iii) *FIFO Ratio (< 1 μs)* $_t$, which represent the five-day moving average rate at which the first order placed is also first to be acknowledged by the exchange when consecutive orders are placed less than two μs or less than one μs apart, respectively. Our independent variable of interest, *Post Upgrade* $_t$, equals one for days following the launch of the NFF on May 26, 2016, and zero otherwise. X_t is a vector of the following control variables: *Total Demand at Close*$_t$, which is the total shares requested at close across all tickers on day t; *Close-to-Close Volatility*$_t$, which is the average absolute percentage change in closing price across all tickers from day t to day $t+1$; *End-of-Day Volatility*$_t$, which is the average absolute percentage change in price from 3:50 PM to market close across all tickers on day t; *Trading Volume*$_t$, which is the average share volume across all tickers based on trades executed in the continuous market on day t; *Rebalance Day Flag*$_t$, which equals one on trading days that fall on (i) an index-rebalancing day, (ii) the last trading day of the month; (iii) the last trading day of the quarter; or (iv) the third Friday of the month; and *Half Day Flag*$_t$, which is equals one on trading days closing at 1:00 PM ET, and zero otherwise. T-statistics are calculated using Newey-West standard errors with five lags to account for potential heteroskedasticity and serial correlation.

The results, which we present in **Table 3**, show that the median O-A latency drops significantly following the major tech overhaul by the exchange. Specifically, we observe a coefficient estimate of -4.461 (t-statistic = -15.21) on the *Post Upgrade* indicator variable (Column 1), which translates to a

4.46 μs decline in the median O-A latency following the technological upgrade and represents roughly a 10% decline in the daily median O-A latency which, on average, was 44.07 μs prior to the upgrade and declines to 39.61 μs following the upgrade (untabulated). Moreover, consistent with **Figure 5c**, the *FIFO Ratio* for time deltas of two µs significantly increases following the upgrade, with a coefficient estimate of 0.0957 (*t*-statistic = 8.80) on the *Post Upgrade* indicator variable (Column 2), suggesting a substantial improvement in the exchange's ability to properly queue consecutive orders placed two μs apart. However, consistent with **Figure 5b**, the *FIFO Ratio* for time deltas of one µs does not materially change. Specifically, we observe a coefficient estimate of -0.0052 (*t*-statistic = -0.84) on the *Post Upgrade* indicator variable (Column 3), which suggests that the technological upgrades by the exchange were insufficient in contemporaneously matching the technological progress of the fastest players.

Together, **Figures 4** and **5** alongside with **Table 3** suggest that the NFF (i) had a meaningful impact on improving the time it takes Nasdaq to acknowledge an order and (ii) substantially improved the exchange's ability to distinguish the timing of orders placed up to two μs apart, but was nonetheless (iii) insufficient in keeping up with the even lower latencies at which the fastest market participants operate, as evidenced by the exchange's inability to reliably distinguish the correct timing of orders placed less than one μs apart.

To provide historical context, the newer commercially available network adapters in 2016 offered an average tick-to-trade latency of 1.538 μs,[13] as shown by a marketing brochure (presented in Appendix A2) released during that time. Thus, improvements in the *FIFO Ratio* for time deltas of two μs, made possible by the launch of the NFF, were insufficient to keep up with the most up-to-date hardware available to market participants. That is, although the exchange invests in technological advancements over time, the fastest players continue to advance at an even faster pace.

[13] See, for instance, a 2016 CSPi marketing brochure for the CSPi ARC Series E-Class network adapter (http://www.cspi.com/wp-content/uploads/2016/06/Tick-to-Trade-Latency_FINAL-2.pdf).

Thus, during our sample period, the ongoing technological disparity between the exchange and its market participants is best proxied by *FIFO Ratio* (*< 1 μs*), which remains dismally low even after Nasdaq's major upgrades went live.[14] However, we note that benchmarks for technological disparity will change over time based on the prevailing capability of the exchange relative to the most advanced high-frequency traders and liquidity providers.

Overall, the evidence thus far suggests that violations in time priority are a real and ongoing risk incurred by a time-sensitive liquidity provider and points to a paradoxical byproduct of the competition among high-frequency market participants – namely, their never-ending quest to be faster than one another has resulted in an ongoing technological disparity between market participants and the exchange itself. However, a new question now arises as to whether these violations in time priority, born from a difference of less than one microsecond, has palpable ramifications on other market participants outside of the first-in-line liquidity providers who directly experience the costs. We now proceed to examine the implications for market quality arising from the uncertainty in the path from order placement to official exchange acknowledgment.

4.2. Implications for Market Quality: Excess Messaging and Rapid Order Cancellations

To test the market externalities arising from the queuing uncertainty experienced by liquidity providers, we begin by exploring the implications for perceived liquidity/depth in the limit-order book, specifically as it pertains to excess messaging and rapid order cancellations. That is, in the continuous market, liquidity providers are free to cancel their orders, and thus, are likely to initially submit more liquidity-adding orders than they plan to fulfill, with the intent to cancel a proportion of their orders

[14] We also note that what we uncover and document, by way of the FIFO ratio, is not simply a technological glitch, which would cause the FIFO ratio for all time deltas to be low for distinct periods (i.e., during a glitch) and should subsequently recover for all time deltas once the glitch is cleared. To the contrary, we find that larger time deltas between orders uniformly result in greater FIFO ratios, and conversely that smaller time deltas uniformly result in lower FIFO ratios.

upon learning their queue positions. For our sample period, liquidity providers are typically notified of their queue position within 43 μs of placing an order (as reported in **Table 1**). Thus, we begin by plotting the time-series trend in the daily average percentage of order cancellations that occur within 50 μs of placing a liquidity-adding order at the onset of a price-formation speed race for queue position. The results, which we present in **Figure 6**, show a substantial increase in rapid-fire order cancellations in recent years, suggesting that excess messaging has increased substantially over time.

To account for other contemporaneous factors of order cancellations in a multivariate setting, we estimate the following OLS regression:[15]

$$\%\ Order\ Cancellations_t = \alpha + \beta_1 \cdot Post\ Upgrade_t + \beta_2 \cdot FIFO\ Ratio(< 1\ \mu s)_t + X_t \cdot \gamma + \varepsilon_t \quad \textbf{(2)}$$

Our dependent variable, *% Order Cancellations*$_t$, is the daily average percentage of orders placed at a new price formation speed race that are cancelled within 50 μs,[16] where speed races for queue position are identified by the formation of a new bid (or offer) at a price that was previously an offer (or bid) where the bid-ask spread is one cent. Our two independent variables of interest are: (i) *FIFO Ratio (< 1 μs)*$_t$, which is the five-day moving average rate for day *t* at which the first order placed is also first to be acknowledged by the exchange when consecutive orders are placed less than one μs apart, and (ii) *Post Upgrade* $_t$, which equals one for days following the launch of the NFF on May 26, 2016, and zero otherwise. X_t is a vector of the following control variables: *Close-to-Close Volatility*$_t$, which is the average absolute percentage change in closing price across all tickers from day *t* to day *t+1*; *Trading Volume*$_t$, which is the average share volume across all tickers based on trades executed in the continuous market on day *t*; *Rebalance Day Flag*$_t$, which equals one on trading days that fall on (i) an index-rebalancing day, (ii) the last trading day of the month; (iii) the last trading day of the quarter; or (iv) the third Friday of the month; and *Half Day Flag*$_t$, which is equals one on trading days

[15] We choose OLS estimation for ease of exposition. A double-censored Tobit model yields very similar results.
[16] Our results are robust to counting the orders cancelled within 100 μs.

closing at 1:00 PM ET, and zero otherwise. *T*-statistics are calculated using Newey-West standard errors with five lags to account for potential heteroskedasticity and serial correlation.[17]

The results, which we present in **Panel A** of **Table 4**, show that *% Order Cancellations* is substantially and significantly associated with both the *FIFO Ratio* (Column 1) and *Post Upgrade* indicator (Column 2), with coefficient estimates of -0.0714 (*t*-statistic = -2.93) and 0.0295 (*t*-statistic = 8.92), respectively. That is, in examining the percentage of orders cancelled within 50 μs of placement (which is when queueing position is revealed to liquidity providers), we see that the percentage of rapid order cancellations continues to *increase* in the period following a major technological upgrade by the exchange, which is consistent with the evidence that queueing uncertainty and violations in time priority has exacerbated over time due to the continued technological disparity between the exchange and its market participants. Overall, in the regression specification including both variables of interest (**Column 3**), the *FIFO Ratio* has a coefficient estimate of -0.0604 (*t*-statistic = -2.50), which suggests that an increase in the *FIFO Ratio* from 60% to 90% translates to a 1.81% decline in the rapid-fire order cancellations occurring within 50 μs of placement. For reference, such a decline represents a 9.64 percent decrease in the average order cancellation rate of 18.78% (as reported in **Table 1**).

As an additional exploration of violations in time priority and their impact on excess messaging, we explore how the *FIFO ratio* relates to the *Inverse Half Life Ratio*, which measures the total quantity of shares added to the limit-order book at the onset of a new price-formation speed race scaled by the half-life quantity of shares, providing an indication of the extent of excess messaging relative to the natural steady-state depth that dealers are willing to provide at the newly formed price level. Thus, we estimate the following OLS regression:

$$Inverse\ Half\ Life\ Ratio_t = \alpha + \beta_1 \cdot Post\ Upgrade_t + \beta_2 \cdot FIFO\ Ratio(< 1\ \mu s)_t + X_t \cdot \gamma + \varepsilon_t \quad \textbf{(3)}$$

[17] Our results are robust to ten-day and 20-day lags.

FIFO Ratio (< 1 μs) $_t$ and *Post Upgrade* $_t$ are, again, the independent variables of interest, and X_t is a vector of the same control variables as specified in **regression equation (2)**. As before, *t*-statistics are calculated using Newey-West standard errors with five lags to account for potential heteroskedasticity and serial correlation.[18]

The results, which we present in **Panel B** of **Table 4**, show that the *Inverse Half-Life Ratio* is also substantially and significantly associated with both the *FIFO Ratio* (Column 1) and *Post Upgrade* indicator (Column 2) with the same qualitative interpretations as our results from **Panel A**. Overall, in the regression specification including both variables of interest (Column 3), the *FIFO Ratio* has a coefficient estimate of -2.8825 (*t*-statistic = -3.03), which suggests that an increase in the *FIFO Ratio* from 60% to 90% translates to a 0.86 decline in the total quantity of shares placed at the onset of a new price-formation relative to the half-life quantity of shares. For reference, this ratio is typically around 3.5 (as reported in **Table 1**). These numbers indicate that 3.5 times as many shares are typically added at the onset of a new price-formation speed race than the quantity remaining half-way throughout the life of the new price formation and that a 30% increase in the *FIFO Ratio* would bring this ratio down to 2.64 times.

Overall, the results suggest that technological disparity and violations in time priority, as measured by *FIFO Ratio (< 1 μs)*, pose a substantial concern for high-frequency liquidity providers. In turn, these liquidity providers submit an excess of order messages in rapid succession, many of which are then cancelled within 0.000050 seconds of placement and do not remain on the limit-order book throughout the life of the price-formation. Moreover, these behaviors persist following major improvements made by the exchange because the improvements, though material, are insufficient to match the ever-lower latencies at which high-frequency players can operate. However, the question

[18] Our results are robust to ten-day and 20-day lags.

still remains as to whether randomness in time priority has a meaningful impact on other market participants. We now proceed to explore the implications for liquidity seekers in the market.

4.3. Implications for Market Quality: Unabsorbed Order Imbalance at the Close

To test the market externalities to liquidity seekers that result from the queuing uncertainty experienced by liquidity providers, we now explore the implications for market quality, specifically as it pertains to unabsorbed order imbalance at the closing cross on Nasdaq, which has been increasing over time (see **Figure 7**). As mentioned previously, the closing cross, which occurs alongside the continuous market for the last ten minutes of the trading day, is an important market mechanism that sets the Nasdaq Official Closing Price (NOCP), and, accordingly, a substantial portion of daily volume occurs at the closing cross. However, in stark contrast to orders placed in the continuous market, liquidity providers are prohibited from canceling their imbalance-only (IO) orders placed at the closing cross and do not see their queue positions until settlement at the end of the trading day. As a result, they are unable to employ the strategy of excess messaging and rapid order cancellations, as we have shown to occur in the continuous market. Thus, a natural question arises as to whether the ongoing technological disparity (and resulting randomness in time priority) causes liquidity providers to be less willing to absorb on-close market demand given the increased difficulty in estimating predictive fill rates relative to other liquidity providers.

To account for other contemporaneous factors in a multivariate setting, we estimate the following OLS regression:[19]

$$\%\ Order\ Imbalance_t = \alpha + \beta_1 \cdot Post\ Upgrade_t + \beta_2 \cdot FIFO\ Ratio(< 1\ \mu s)_t + X_t \cdot \gamma + \varepsilon_t \quad \textbf{(4)}$$

[19] We choose OLS estimation for ease of exposition. A double-censored Tobit model yields very similar results.

Our dependent variable, *% Order Imbalance*$_t$, is the average percentage of unabsorbed orders (across all tickers) from the final Net Order Imbalance Indicator (NOII) message on day *t*. Our independent variables of interest are *FIFO Ratio (< 1 µs)*$_t$ and *Post Upgrade* $_t$. X_t is a vector of the following control variables: *Total Demand at Close*$_t$, which is the total shares requested at close across all tickers on day *t*; *Close-to-Close Volatility*$_t$, which is the average absolute percentage change in closing price across all tickers from day *t* to day *t+1*; *End-of-Day Volatility*$_t$, which is the average absolute percentage change in price from 3:50 PM to market close across all tickers on day *t*; *Trading Volume*$_t$, which is the average share volume across all tickers based on trades executed in the continuous market on day *t*; *Rebalance Day Flag*$_t$, which equals one on trading days that fall on (i) an index-rebalancing day, (ii) the last trading day of the month; (iii) the last trading day of the quarter; or (iv) the third Friday of the month; and *Half Day Flag*$_t$, which is equals one on trading days closing at 1:00 PM ET, and zero otherwise. *T*-statistics are calculated using Newey-West standard errors with five lags to account for potential heteroskedasticity and serial correlation.[20]

The results, which we present in **Table 5**, show that *% Order Imbalance* is substantially and significantly associated with both the *FIFO Ratio* (Column 1) and *Post Upgrade* indicator (Column 2), with coefficient estimates of -0.0120 (*t*-statistic = -2.79) and 0.0011 (*t*-statistic = 2.25), respectively. That is, consistent with the idea that continued technological disparity has exacerbated violations in time priority, and accordingly, the risk to liquidity providers, we see that unfilled on-close demand for liquidity continues to *increase* in the period following a major technological upgrade by the exchange. Overall, an increase in the *FIFO Ratio* from 60% to 90% translates to a 0.345% decline in *% Order Imbalance* (Column 3), which represents a 35.9 percent decline from the daily average percentage order imbalance of 0.96% (as reported **Table 1**). Furthermore, *% Order Imbalance* moves as expected with observable fundamental factors that should contribute to end-of-day imbalances. For instance,

[20] Our results are robust to ten-day and 20-day lags.

end-of-day imbalance tends to be substantially greater when total on-close demand is greater (coefficient estimate = 0.0936; *t*-statistic = 4.93), and tends to be substantially higher on rebalance days (coefficient estimate = 0.0025; *t*-statistic = 1.68).

As a sanity check, we conduct a subsample analysis to examine the relation between the *FIFO Ratio* and unabsorbed order imbalance for large cap versus non-large cap stocks, whereby we bifurcate our sample between stocks in the top quintile based on market capitalization and those that are not (in the top quintile). Because large-cap inventory is inherently easier to hedge, liquidity providers should be less concerned with the prospect of unexpectedly carrying a position overnight, thereby making them less deterred by the prevailing queuing uncertainty when placing IO orders to absorb on-close demand for large-cap stocks.

The results, which we present in **Table 6**, are consistent with this expectation. Specifically, we observe that tickers at smaller market capitalizations stand to benefit more from improvements in properly determining time priority, whereby an increase in the *FIFO ratio* from 60% to 90% is associated with a 1.21% decline (coefficient estimate = -0.0403; *t*-statistic = -2.95) in the aggregate percentage order imbalance among stocks outside of the top quintile with respect to market capitalization (Column 1). In comparison, a similar increase in the *FIFO Ratio* is associated with a 0.27% decline (coefficient estimate = -0.0089; *t*-statistic = -2.24) in the aggregate percentage order imbalance among stocks in the top quintile with respect to market capitalization (Column 2).

Overall, prior literature has focused on technological improvements made by either the market participants or the exchange, but the technological gap between these two groups has not been studied until now. We argue that this disparity is also a very important part of discussions of optimal market design. Our analyses not only document violations in time priority but also provide suggestive evidence that this uncertainty borne by speed-sensitive liquidity providers poses a material risk, which, in turn, is passed on to liquidity seekers who are more likely to suffer end-of-day order imbalances for their unfilled, on-close orders when randomness in time priority is high.

5. Discussion and Policy implications

In this paper, we provide evidence that competition among speed-sensitive market participants to gain a technological advantage over one another has led to technological imbalance between the fastest participants and the exchange itself. Specifically, we find that the proportion of times in which the first order entered is also first to be acknowledged by the exchange is quite low when consecutive orders are placed at very high frequencies. We also provide suggestive evidence of impaired market quality in the form of: (i) increased excess messaging in the continuous market, and (ii) greater unabsorbed order imbalance at the closing cross as a result of the randomness in honoring price-time priority. Most importantly, we provide evidence that these issues do not improve following a major technological upgrade by the exchange, in part, because of the continued technological disparity between the exchange and its high-frequency players.

With respect to potential solutions to the issues we have empirically documented, we propose a few alternative market designs that could alleviate an exchange's perceived ambiguity in time priority. For instance, queuing problems can be mitigated, to an extent, with finer and finer tick sizes. That is, finer tick sizes first promote price-based priority, making time priority less important, as evidenced by the lower depth at each (now, finer) quote following decimalization (SEC, 2012). However, there are limits to how finely tick size can be meaningfully reduced, and relatedly, Werner, Rindi, Buti, and Wen (2023) theoretically and empirically demonstrate that tick-size reductions may not necessarily lead to improved market quality. Thus, an ever-decreasing tick size is not the first-order solution to effectively address technological disparity.

With respect to potential infrastructure-related solutions, one option is to time stamp each order message as it arrives at the gateway server. Each gateway could then send time-stamped messages to the matching engine server, allowing the messages to be properly sorted and prioritized. This approach has the benefit of enhancing time priority, though at the cost of increased delays in message processing. There is a more serious risk that orders could become more severely out of sequence, particularly under

periods of heavy trading volume. However, dynamic buffer window lengths could potentially address these issues.

Another option is to split the message traffic across multiple matching engines to allow market participants to send orders directly to the matching engine (without first accessing the gateway server). This solution would reduce the queuing of messages and allow the networking technology to accurately determine time priority, as evidenced by the improvement in FIFO ratios (for time deltas of < 2 μs between consecutive orders) following the implementation of the NFF on May 26, 2016. Thus, we anticipate that subsequent updates would lead to further improvements in the FIFO ratio even at sub-microsecond time deltas between consecutive orders.

Overall, as we have demonstrated, market participants continue to also improve (and quite rapidly) as the exchange advances. Thus, solutions with the most longevity must account for the fact that technological disparity is likely to persist.

6. Concluding Remarks

Our study bridges an important gap in the literature, which has heretofore taken queuing uncertainty for granted without exploring the source and severity of this phenomena. Our study also sheds light on a phenomena often referred to as "quote stuffing" (Gai, Yao, and Ye, 2013; Egginton, Van Ness, and Van Ness, 2016) by offering evidence that queuing uncertainty itself contributes to both excess messaging and impaired market quality.

To be clear, much work is required before we can arrive at more definitive conclusions as to the extent of the benefits versus costs of suggested improvements to market structure and design. Overall, the evidence we present suggests an unintended but costly byproduct of the asymmetry in the technological advances of high-frequency liquidity providers relative to that of the exchange, and broaches an important and ever-present issue to consider in market design in a high-frequency era.

References

Baldauf, M. and Mollner, J., 2020. High-frequency trading and market performance. *The Journal of Finance*, 75(3), 1495-1526.

Biais, B., and T. Foucault, 2014. HFT and Market Quality. *Bankers, Markets & Investors* 128, 5-19.

Biais, B., T. Foucault, and S. Moinas, 2015. Equilibrium Fast Trading. *Journal of Financial Economics* 116, 292-313.

Boehmer, E., Fong, K. and Wu. J. 2015. International evidence on algorithmic trading. AFA 2013 paper

Brogaard, J., Hagströmer, B., Nordén, L. and Riordan, R., 2015. Trading fast and slow: Colocation and liquidity. *Review of Financial Studies*, 28(12), 3407-3443.

Brogaard, J., T. Hendershott, S. Hunt, and C. Ysusi, 2014. High-Frequency Trading and the Execution Costs of Institutional Investors. *The Financial Review* 49, 345-369.

Brogaard, J., T. Hendershott, and R. Riordan, 2014. High-Frequency Trading and Price Discovery. *Review of Financial Studies* 27, 2267-2306.

Budish, E., P. Cramton, and J. Shim, 2015. The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response. *Quarterly Journal of Economics* 130, 1547-1621.

Burdick, Alan. "The Olympics' Never-Ending Struggle to Keep Track of Time", *The New Yorker,* 8 February, 2018

Conrad, J., S. Wahal, and J. Xiang, 2015. High-Frequency Quoting, Trading, and the Efficient of Prices. *Journal of Financial Economics* 116, 271-291

Egginton, J. F., B. F. Van Ness, and R. A. Van Ness, 2016. Quote Stuffing. *Financial Management* 45 (3), 583-608.

Foucault, T., Hombert, J. and Roşu, I., 2016. News trading and speed. *The Journal of Finance*, 71(1), 335-382.

Frino, A., Mollica, V. and Webb. R. I. 2014. The impact of co-location of securities exchanges' and traders' computer servers on market liquidity. *Journal of Futures Markets* 34:20–33.

Gai, J., Yao, C. and, Ye, M., 2013. The externalities of high frequency trading. WBS Finance Group Research Paper, (180).

Hoffmann, P., 2014. A dynamic limit order market with fast and slow traders. *Journal of Financial Economics*, 113(1), 156-169.

Huang, S. and Yueshen, B.Z., 2021. Speed acquisition. *Management Science*, 67(6), 3492-3518.

Holden, C. and S. Jacobsen, 2014. Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions. *Journal of Finance* 69, 1747-1785.

Jarrow, R., and P. Protter, 2012. A Dysfunctional Role of High Frequency Trading in Electronic Markets. *International Journal of Theoretical and Applied Finance* 15, 2-15.

Kemme, D.M., McInish, T.H. and Zhang, J., 2022. Market fairness and efficiency: Evidence from the Tokyo Stock Exchange. *Journal of Banking & Finance*, 134.

Kervel, V., 2015. Competition for Order Flow with Fast and Slow Traders. *Review of Financial Studies* 28, 2094-2127.

Li, S., Ye, M. and Zheng, M., 2023. Refusing the Best Price?. *Journal of Financial Economics,* 147(2), 317-337.

Menkveld, A.J. and Zoican, M.A., 2017. Need for speed? Exchange latency and liquidity. *Review of Financial Studies*, 30(4), 1188-1228.

Nasdaq Press Release, May 16, 2016. Nasdaq Debuts Groundbreaking Nasdaq Financial Framework, Enhancing Operations for Over 100 Market Operators Globally, accessed on <https://www.nasdaq.com/about/press-center/nasdaq-debuts-groundbreaking-nasdaq-financial-framework-enhancing-operations>.

Riordan, R., and Storkenmaier. A. 2012. Latency, liquidity and price discovery. *Journal of Financial Markets*15:416–37.

O'Hara, M., 2015. High Frequency Market Microstructure. *Journal of Financial Economics* 116, 257-270.

Pagnotta, E.S. and Philippon, T., 2018. Competing on speed. *Econometrica*, 86(3), pp.1067-1115.

Securities and Exchange Commission, 2010. Concept Release on Equity Market Structure; Proposed Rule. 17 CFR Part 242, Vol. 75, No. 14, 3694-3614.

Securities and Exchange Commission, 2012. Report to Congress on decimalization. Required by Section 106 of the JOBS Act.US Government Printing Office, Washington, DC.

Shkilko, A. and Sokolov, K., 2020. Every cloud has a silver lining: Fast trading, microwave connectivity, and trading costs. *The Journal of Finance*, 75(6), 2899-2927.

Upson, J., McInish, T. and Johnson IV, B.H., 2021. Order based versus level book trade reporting: An empirical analysis. *Journal of Banking & Finance*, 125

Werner, I. M., Rindi, B., Buti, S., and Wen, Y., 2023. Tick size, trading strategies, and market quality. *Management Science*, 69(7), 3818-3837.

Yao, C. and Ye, M., 2018. Why trading speed matters: A tale of queue rationing under price controls. *The Review of Financial Studies*, 31(6), 2157-2183.

Yueshen, B. Z., 2014. Queuing Uncertainty in Limit Order Market. *Working Paper Series*.

Table 1. Summary Statistics

This table presents descriptive statistics for our sample of 3,740 unique tickers spanning the first trading day of January 2014 through the last trading day of April 2017. We begin with a total of 2,911,692 ticker-days. We then aggregate all statistics to a daily frequency, and we report summary statistics on these daily averages for the following measures: *FIFO Ratio (< 1 μs)*, *FIFO Ratio (< 2 μs)*, *Order-to-Accept (O-A) Latency*, *Intra-Day Jitter*, *Speed Races for Queue Position*, *Inverse Half-Life Ratio*, *% Order Cancellations*, *% Order Imbalance*, *% Tickers with Order Imbalance*, *% Tickers with ≥ 20% Order Imbalance*, *Total Demand at Close* (in millions), and *Trading Volume* (in millions). Definitions of these variables are described in detail in Appendix Table A1.

	Mean	(Stdev.)	P25	P75
FIFO Ratio (< 1 μs)	0.5931	(0.052)	0.5568	0.6216
FIFO Ratio (< 2 μs)	0.6440	(0.080)	0.5888	0.6961
Order-to-Accept (O-A) Latency (μs)	42.89	(5.84)	39.94	43.81
Intra-Day Jitter (μs)	10.15	(4.15)	6.65	13.50
Speed Races for Queue Position	2,828,673	(892,079)	2,269,080	3,196,352
Inverse Half-Life Ratio	3.52	(1.27)	2.96	3.57
% Order Cancellations (within 50 μs)	0.1878	(0.021)	0.1730	0.1977
% Order Imbalance	0.0096	(0.014)	0.0051	0.0100
% Tickers with Order Imbalance	0.2444	(0.068)	0.2009	0.2777
% Tickers with ≥ 20% Order Imbalance	0.0675	(0.028)	0.0481	0.0807
Total Demand at Close	0.0366	(0.034)	0.0256	0.0364
Trading Volume	0.9351	(0.173)	0.8319	1.0228
Number of observations	819	---	---	---

Table 2. Summary Statistics for Select Tickers

This table reports, for select tickers at various quantiles: (i) the percentage of orders cancelled within 50 μs of placement; (ii) the average ratio of the total quantity of shares placed at the onset of a new price-formation speed race scaled by the half-life quantity of shares; (iii) the number of liquidity-adding speed races for queue position, identified by the formation of a new bid (or offer) at a price that was previously an offer (or bid) where the bid-ask spread is one cent; (iv) the percentage of daily volume filled at the close; and (v) the percentage order imbalance from the final Net Order Imbalance Indicator (NOII) message of the day. We also report these metrics for the daily mean across all tickers during this timeframe. Overall, our sample consists of 3,740 unique tickers spanning the first trading day of January 2014 through the last trading day of April 2017.

	P25	Median	P75	Max
Daily mean across all tickers				
(i) % order cancellations	0.1730	0.1854	0.1977	0.2541
(ii) Excess Messaging Ratio	2.96	3.21	3.57	17.04
(iii) Speed races for queue position	488	650	669	1,423
(iv) % of volume at close	0.0413	0.0488	0.0582	0.3484
(v) % end of day imbalance	0.0354	0.0445	0.0602	0.1590
"MSFT"				
(i) % order cancellations	0.5213	0.5518	0.6053	0.8945
(ii) Excess Messaging Ratio	4.23	4.90	5.80	12.91
(iii) Speed races for queue position	2,336	3,415	5,014	41,647
(iv) % of volume at close	0.0570	0.0784	0.1051	0.2997
(v) % end of day imbalance	0.0000	0.0000	0.0000	0.2439
"YHOO"				
(i) % order cancellations	0.4842	0.5294	0.5820	0.7535
(ii) Excess Messaging Ratio	3.54	4.37	5.00	7.77
(iii) Speed races for queue position	2,474	3,544	5,550	43,795
(iv) % of volume at close	0.0209	0.0349	0.0736	0.4985
(v) % end of day imbalance	0.0000	0.0000	0.1500	0.3993
"AAAP"				
(i) % order cancellations	0.0000	0.0000	0.0000	0.5000
(ii) Excess Messaging Ratio	1.00	1.50	2.00	12.49
(iii) Speed races for queue position	2	4	8	117
(iv) % of volume at close	0.0018	0.0041	0.0089	0.3791
(v) % end of day imbalance	0.0000	0.0000	0.2125	1.0000
"CUR"				
(i) % order cancellations	0.0481	0.1401	0.2136	0.7897
(ii) Excess Messaging Ratio	2.05	2.60	3.62	28.49
(iii) Speed races for queue position	38	86	173	1,106
(iv) % of volume at close	0.0007	0.0030	0.0083	0.1631
(v) % end of day imbalance	0.0000	0.1826	0.6831	1.0000

Table 3. Technological Disparity Pre-Versus-Post Exchange Upgrade

This table presents estimates from the following time-series OLS regressions:

$$Dep\ Var_{\ t} = \alpha + \beta \cdot Post\ Upgrade_{\ t} + X_t \cdot \gamma + \varepsilon_t$$

The dependent variable in Column (1), *Median OA Latency* $_t$, is the median distance in time from when an order is placed to when it is officially acknowledged by Nasdaq for each day. The dependent variable in Columns (2) and (3), *FIFO Ratio (< 2 μs)* $_t$ *and FIFO Ratio (< 1 μs)* $_t$, capture the five-day moving average rate at which the first order placed is also first to be acknowledged by the exchange when consecutive orders are placed less than two *μs* apart or less than one *μs* apart , respectively. *Post Upgrade* $_t$ equals one for days following the major tech overhaul by Nasdaq (NFF) on May 26, 2016, and zero otherwise. X_t is a vector of the following control variables, which are described in Appendix Table A1: *Total Demand at Close* (in millions), *Close-to-Close Volatility*, *End-of-Day Volatility*, *Trading Volume* (in millions), *Rebalance Day Flag*, and *Half Day Flag*. All variables have been aggregated at a daily frequency from January 2014 through May 2017, from a sample of 4,034,086 ticker-days. *T*-statistics are calculated using time-clustered standard error (column 1) and Newey-West standard errors with 5 lags (Columns 2 and 3). Statistical significance at the 10%, 5%, and 1% levels are denoted by *, **, and ***, respectively.

	OA Latency (Median)	*FIFO Ratio (< 2 μs)* $_t$	*FIFO Ratio (< 1 μs)* $_t$
	(1)	(2)	(3)
Post Upgrade $_t$	-4.4614***	0.0957***	-0.0052
	[-15.21]	[8.80]	[-0.84]
Total Demand at Close $_t$	-9.7152	0.0994	0.0748
	[-1.45]	[1.07]	[1.21]
Close-to-Close Volatility $_t$	-0.0380***	0.0010***	0.0005***
	[-5.45]	[5.18]	[3.85]
End-of-Day Volatility $_t$	-0.2638***	-0.0152***	-0.0126***
	[-2.78]	[-9.67]	[-11.73]
Trading Volume $_t$	2.5092	-0.0242	-0.0098
	[1.44]	[-0.92]	[-0.52]
Rebalance Day Flag $_t$	0.6297	-0.0158*	-0.0111*
	[0.73]	[-1.72]	[-1.72]
Half-Day Flag $_t$	1.1549	-0.0178	-0.0136
	[0.74]	[-0.76]	[-0.67]
Intercept	42.1789***	0.6376***	0.6014***
	[28.13]	[28.32]	[35.65]
Adjusted R-squared	0.1361	0.2981	0.0051
Number of observations	808	808	808

Table 4. FIFO Ratio and Excess Messaging

This table presents estimates from the following time-series OLS regressions:

$$Dep\ Var_{\ t} = \alpha + \beta_1 \cdot Post\ Upgrade_{\ t} + \beta_2 \cdot FIFO\ Ratio\ (< 1\ \mu s)_t + X_t \cdot \gamma + \varepsilon_t$$

The dependent variable in ***Panel A***, *% Order Cancellations*$_t$, is the average percentage of orders placed at a new price-formation speed race that are cancelled within 50 μs, where liquidity-adding speed races for queue position are identified by the formation of a new bid (or offer) at a price that was previously an offer (or bid) where the bid-ask spread is one cent. The dependent variable in ***Panel B***, *Inverse Half-Life Ratio* $_t$, is the average ratio of the total quantity of shares placed at the onset of a new price-formation speed race scaled by the half-life quantity of shares, where the half-life quantity is identified as the quantity of shares available halfway through the life of a given price level. *Post Upgrade* $_t$ equals one for days following the major tech overhaul by Nasdaq (NFF) on May 26, 2016, and zero otherwise. *FIFO Ratio(< 1 μs)* $_t$ is the five-day moving average rate at which the first order placed is also first to be acknowledged by the exchange when consecutive orders are placed less than one μs apart. X_t is a vector of the following control variables, which are described in Appendix Table A1: *Close-to-Close Volatility*, *Trading Volume* (in millions), *Rebalance Day Flag*, and *Half Day Flag*. All variables have been aggregated at a daily frequency from January 2014 through May 2017, from a sample of 4,034,086 ticker-days. *T*-statistics are calculated using Newey-West standard errors with 5 lags. Statistical significance at the 10%, 5%, and 1% levels are denoted by *, **, and ***, respectively.

Panel A. Dependent Variable = % Order Cancellations $_t$

	(1)	(2)	(3)
FIFO Ratio (< 1 μs) $_t$	-0.0714***		-0.0604**
	[-2.93]		[-2.50]
Post Upgrade $_t$		0.0295***	0.0292***
		[8.92]	[9.13]
Close-to-Close Volatility $_t$	0.0001***	0.0002***	0.0002***
	[5.68]	[4.81]	[5.82]
Trading Volume $_t$	-0.0073	0.0041	0.0036
	[-1.26]	[0.89]	[0.78]
Rebalance Day Flag $_t$	-0.0013	0.0000	-0.0003
	[-0.57]	[0.02]	[-0.18]
Half-Day Flag $_t$	-0.0207***	-0.0113**	-0.0121**
	[-2.58]	[-2.03]	[-2.13]
Intercept	0.2352***	0.1744***	0.2108***
	[14.03]	[37.85]	[13.33]
Adjusted R-squared	0.0314	0.4284	0.4480
Number of observations	808	808	808

Table 4 continued.

Panel B. Dependent Variable = Inverse Half-Life Ratio $_t$

	(1)	(2)	(3)
FIFO Ratio (< 1 μs) $_t$	-3.0044***		-2.8825***
	[-4.27]		[-4.03]
Post Upgrade $_t$		0.3364***	0.3241***
		[3.01]	[2.99]
Close-to-Close Volatility $_t$	0.0012	0.0004	0.0019
	[0.79]	[0.26]	[1.23]
Trading Volume $_t$	-0.4565*	-0.3137	-0.3356
	[-1.92]	[-1.23]	[-1.39]
Rebalance Day Flag $_t$	-0.1296*	-0.1027	-0.1190*
	[-1.85]	[-1.41]	[-1.68]
Half-Day Flag $_t$	-0.7033***	-0.5677***	-0.6078***
	[-4.63]	[-3.74]	[-4.46]
Intercept	5.6632***	3.6587***	5.3922***
	[13.24]	[16.18]	[12.84]
Adjusted R-squared	0.0160	0.0177	0.0310
Number of observations	808	808	808

Table 5. Technological Disparity and End-of-Day Order Imbalance

This table presents estimates from the following time-series OLS regression:

$$\%\ Order\ Imbalance_t = \alpha + \beta_1 \cdot Post\ Upgrade_t + \beta_2 \cdot FIFO\ Ratio\ (< 1\ \mu s)_t + X_t \cdot \gamma + \varepsilon_t$$

The dependent variable, *% Order Imbalance* $_t$, is the percentage of unabsorbed orders from the final Net Order Imbalance Indicator (NOII) message of the day (across all tickers with a starting imbalance as of the first NOII message at 3:50:00 PM). *Post Upgrade* $_t$ equals one for days following the major tech overhaul by Nasdaq (NFF) on May 26, 2016, and zero otherwise. *FIFO Ratio (< 1 μs)* $_t$, captures the five-day moving average rate at which the first order placed is also first to be acknowledged by the exchange when consecutive orders are placed less than one μs apart. X_t is a vector of the following control variables, which are described in Appendix Table A1: *Total Demand at Close* (in millions), *Close-to-Close Volatility*, *End-of-Day Volatility*, *Trading Volume* (in millions), *Rebalance Day Flag*, and *Half Day Flag*. All variables have been aggregated at a daily frequency from January 2014 through April 2017, from a sample of 2,911,692 ticker-days. *T*-statistics are calculated using Newey-West standard errors with 5 lags. Statistical significance at the 10%, 5%, and 1% levels are denoted by *, **, and ***, respectively.

	(1)	(2)	(3)
FIFO Ratio (< 1 μs) $_t$	-0.0120***		-0.0115***
	[-2.79]		[-2.61]
Post Upgrade $_t$		0.0011**	0.0011**
		[2.25]	[2.15]
Total Demand at Close $_t$	0.0963***	0.0927***	0.0936***
	[5.16]	[4.84]	[4.93]
Close-to-Close Volatility $_t$	0.0000	-0.0000	0.0000
	[0.20]	[-0.01]	[0.33]
End-of-Day Volatility $_t$	-0.0001*	0.0000	-0.0001
	[-1.91]	[0.93]	[-1.31]
Trading Volume $_t$	0.0012	0.0018	0.0017
	[0.98]	[1.48]	[1.41]
Rebalance Day Flag $_t$	0.0023	0.0027*	0.0025*
	[1.55]	[1.78]	[1.68]
Half-Day Flag $_t$	0.0071*	0.0076**	0.0074**
	[1.89]	[2.09]	[1.98]
Intercept	0.0106***	0.0027**	0.0097***
	[3.84]	[2.04]	[3.38]
Adjusted R-squared	0.3422	0.3440	0.3481
Number of observations	808	808	808

Table 6. Technological Disparity and End-of-Day Imbalance for Large-Cap Vs. Non-Large Cap Stocks

This table presents estimates from the following time-series OLS regression:

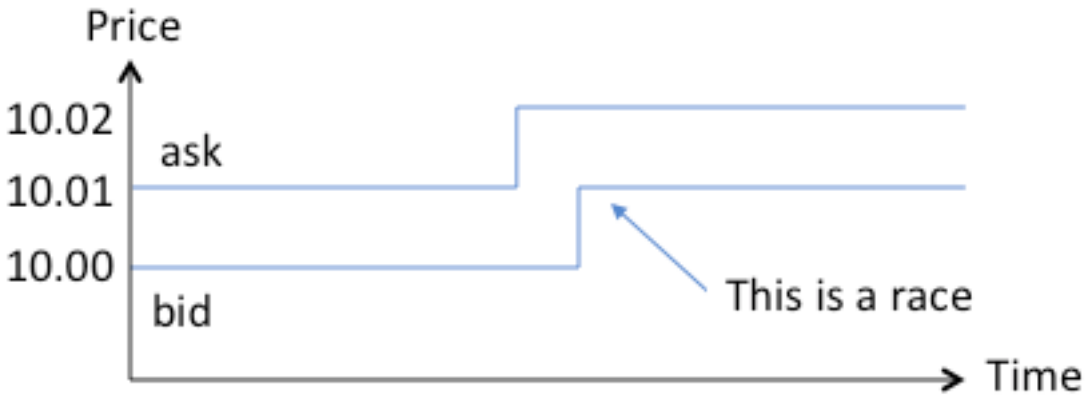
$$\%\ Order\ Imbalance_t = \alpha + \beta_1 \cdot FIFO\ Ratio\ (< 1\ \mu s)_t + X_t \cdot \gamma + \varepsilon_t$$

The dependent variable, *% Order Imbalance* $_t$, is the percentage of unabsorbed orders from the final Net Order Imbalance Indicator (NOII) message of the day (across all tickers with a starting imbalance as of the first NOII message at 3:50:00 PM). Columns (1) and (2) present the results separated by non-large-cap and large-cap tickers, respectively, where the large-cap tickers are determined by the top quintile based on market capitalization. *FIFO Ratio (< 1 μs)* $_t$ is the five-day moving average rate at which the first order placed is also first to be acknowledged by the exchange when consecutive orders are placed less than one *μs* apart. X_t is a vector of the following control variables, which are described in Appendix Table A1: *Total Demand at Close* (in millions), *Close-to-Close Volatility*, *End-of-Day Volatility*, *Trading Volume* (in millions), *Rebalance Day Flag*, and *Half Day Flag*. All variables have been aggregated at a daily frequency from January 2014 through April 2017, from a sample of 2,911,692 ticker-days. *T*-statistics are calculated using Newey-West standard errors with 5 lags. Statistical significance at the 10%, 5%, and 1% levels are denoted by *, **, and ***, respectively.

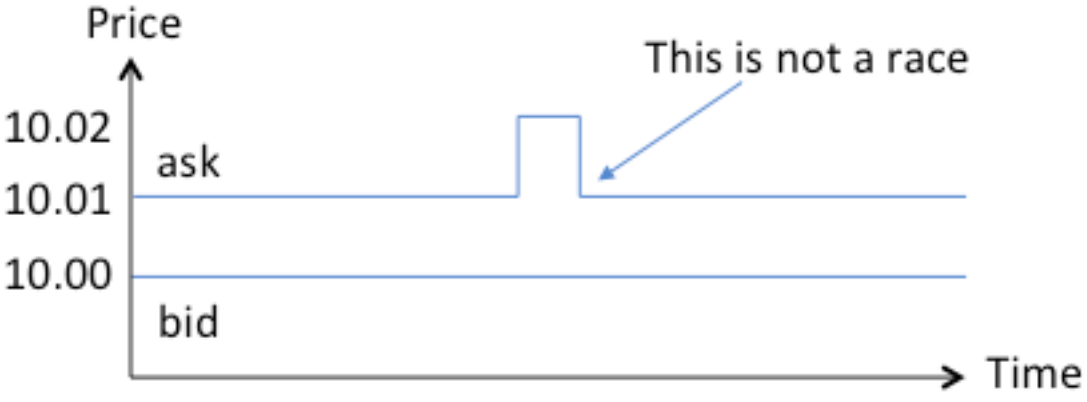
	Non-Large Cap Tickers	*Large-Cap Tickers*
	(1)	(2)
FIFO Ratio (< 1 μs) $_t$	-0.0403***	-0.0089**
	[-2.95]	[-2.24]
Total Demand at Close $_t$	0.1580***	0.0343***
	[2.96]	[12.62]
Close-to-Close Volatility $_t$	-0.0000	0.0000
	[-1.48]	[0.14]
End-of-Day Volatility $_t$	-0.0013***	-0.0001
	[-3.85]	[-0.93]
Trading Volume $_t$	-0.0408***	0.0007
	[-7.28]	[1.90]
Rebalance Day Flag $_t$	0.0016	0.0027***
	[0.77]	[3.12]
Half-Day Flag $_t$	-0.0035*	0.0082***
	[-0.85]	[3.69]
Intercept	0.0664***	0.0062**
	[6.85]	[2.40]
Adjusted R-squared	0.065	0.332
Number of observations	808	808

Figure 1. Race Scenario for Queue Position to Add Liquidity upon New Price Formation

This figure graphically demonstrates the kind of price formations that predicate a speed race to add liquidity to the limit-order book. Specifically, a speed race is defined by the formation of a new bid (or offer) at a price that was previously an offer (or a bid), where the bid-ask spread is one cent. For our analyses, we focus on securities priced over $1.00 with a minimum tick size of one cent.



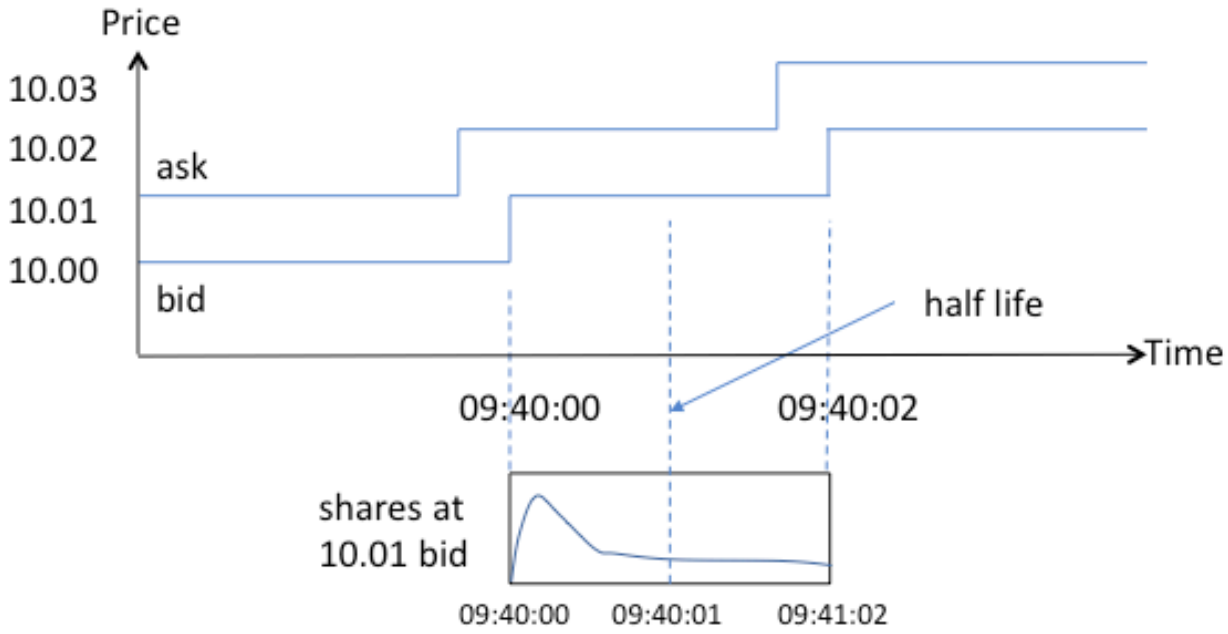
Here, the bbo of 10.00 / 10.01 opens up to 10.00 / 10.02. The formation of a new bid at 10.01 when the 10.01 price used to be an offer marks a new speed race for queue position to add liquidity.



Again, the bbo of 10.00 / 10.01 opens up to 10.00 / 10.02. However, the 10.01 price level backfills the previous 10.01 offer. Thus, the reformation of the price level at 10.01 is not considered a new race.

Figure 2. Depicting the Half-Life Quantity for a New Price Formation

This figure graphically demonstrates the half-life quantity of shares on the limit-order book for a given price formation. Specifically, we measure the steady-state depth that dealers are willing to provide by the quantity of shares for each price level halfway through the life of the price level. We then use this quantity to calculate a *Half-Life Ratio*, which represents the half-life quantity scaled by the total quantity of shares added at the onset of the new price-formation speed race.



Here the bbo of 10.00 / 10.01 opens up to 10.00 / 10.02. We mark the time stamp upon the start of the race to form the bid at 10.01. We also mark the time stamp at the start of a new race to form the 10.02 price level. We then look at the time stamp at the half-life of the price level to sample the number of shares on the book and the number of resting orders on the book.

Figure 3. Sample Price and Depth Formation

This figure plots a sample price and liquidity formation for INTC (Intel Corporation) on May 31, 2017. The top figure demonstrates a speed race to add liquidity to the limit-order book, where the best bid and ask quotes experience a one-cent shift upward. The bottom figures demonstrate the open-order count and resting share count, respectively, throughout the life of this given price formation, whereby the x-axis represents the logged time delta (in *ns*) since the onset of this speed race. The vertical gray line demonstrates the half-life quantity based on this logarithmic time scale (i.e., the quantity of shares for a given price formation halfway through the life of the price level).

Figure 4. Median Order-to-Accept (O-A) Latency over Time on a Single Port

This figure plots the daily median order-to-accept (O-A) latency in microseconds (μs) for the period spanning January 2014 through May 2017. The O-A latency refers to the distance in time from when an order is placed to when it is acknowledged by the matching engine. The 20th percentile and 80th percentile OA latencies for each day are plotted in gray.

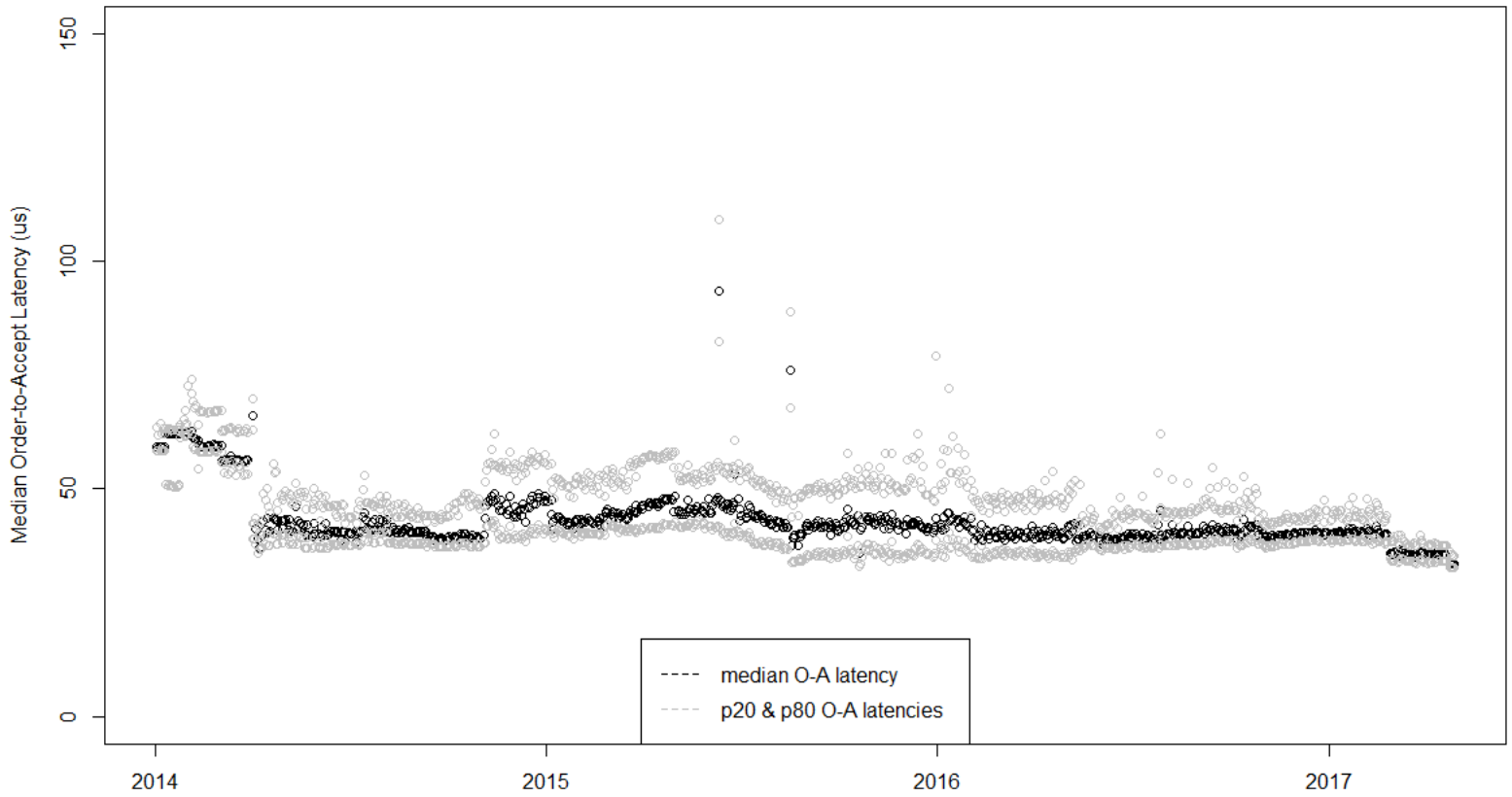


Figure 5a. FIFO Ratio Based on the Time between Consecutive Orders

This figure plots the *FIFO ratio* against the time (in nanoseconds, *ns*) between back-to-back orders. The FIFO ratio is the proportion of first orders placed that are first to be acknowledged by Nasdaq's matching engine. Here, we plot the average *FIFO* ratio across the varying time deltas for the period spanning January 2014 through May 2017.

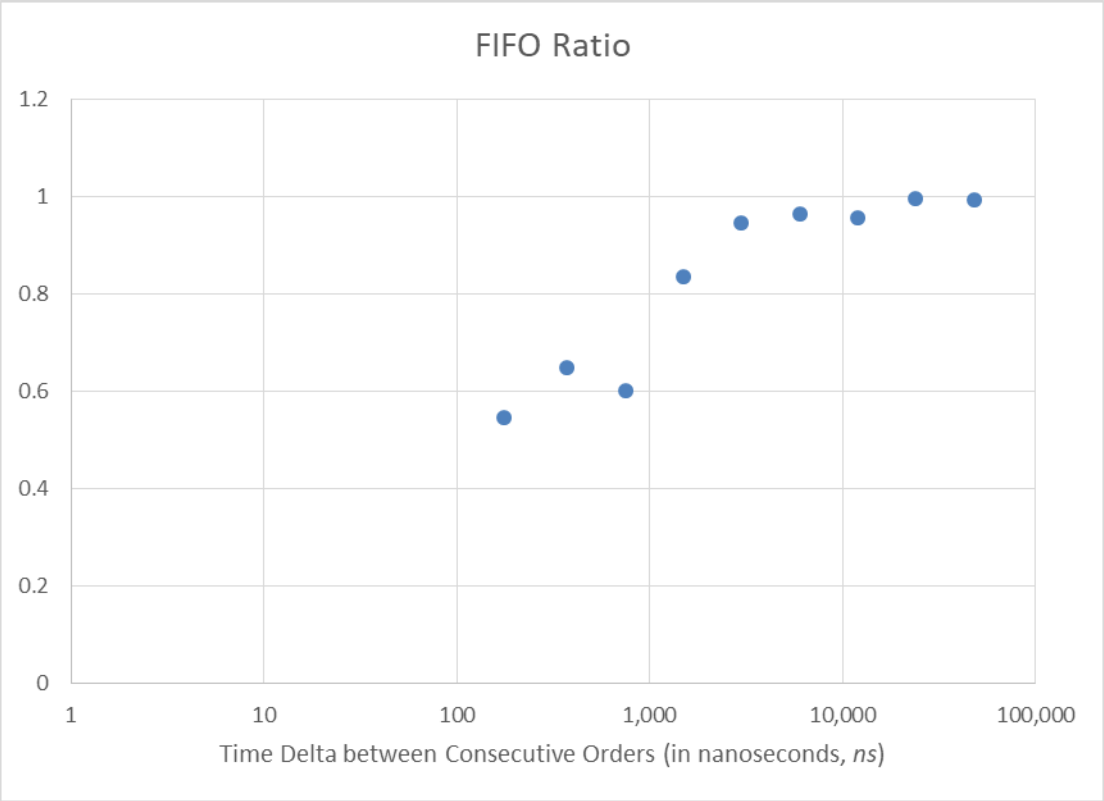


Figure 5b. FIFO Ratio over Time (for time deltas of < 2 μs)

This figure plots the five-day moving average *FIFO ratio* over time for time deltas of < 2 μs between back-to-back orders. The *FIFO ratio* is the proportion of first orders placed that are first to be acknowledged by Nasdaq's matching engine.

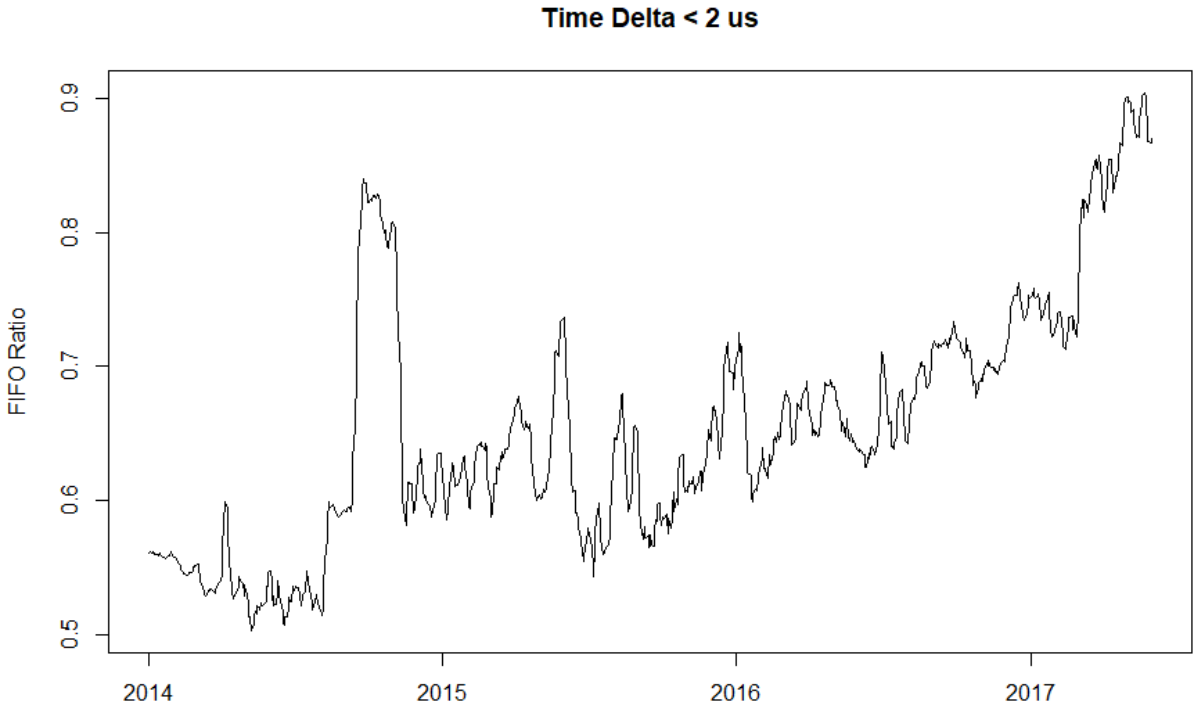


Figure 5c. FIFO Ratio over Time (for time deltas of < 1 μs)

This figure plots the five-day moving average *FIFO ratio* over time for time deltas of < 1 μs between back-to-back orders. The *FIFO ratio* is the proportion of first orders placed that are first to be acknowledged by Nasdaq's matching engine.

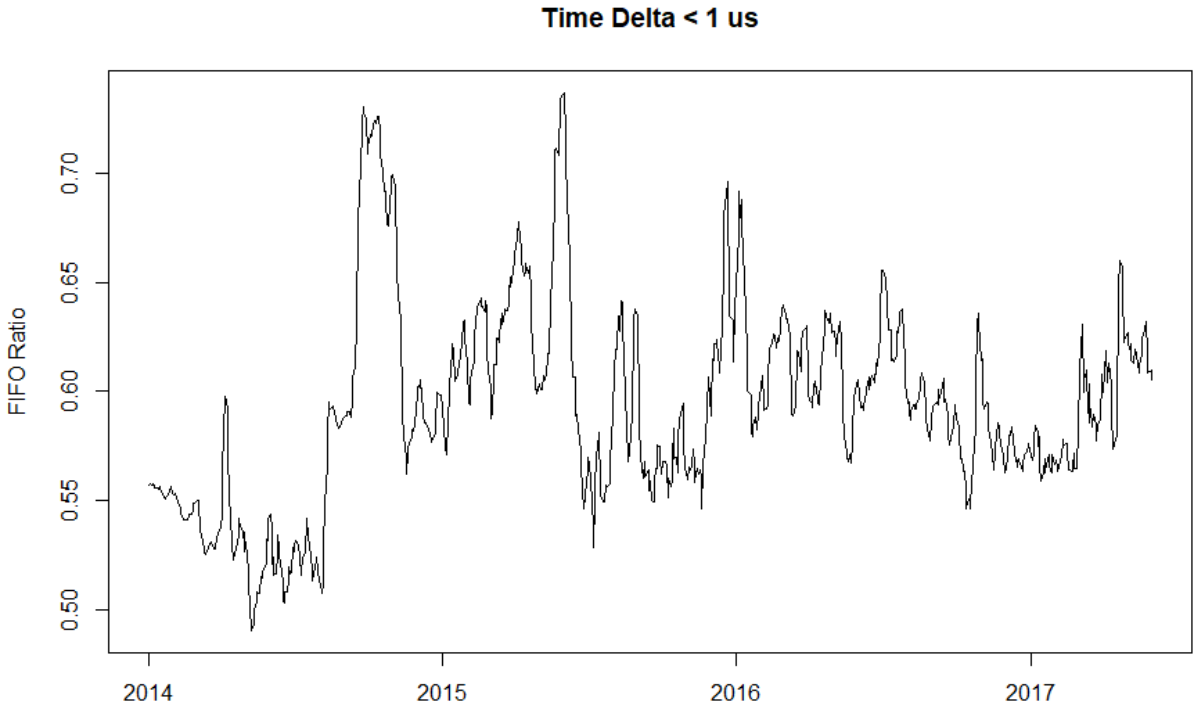


Figure 6. Percentage of Rapid Order Cancellations over Time

This figure plots, at a daily frequency from January 2014 through May 2017, the average percentage of orders placed at a new price-formation speed race that are cancelled within 50 μs. These liquidity-adding speed races for queue position are identified by the formation of a new bid (or offer) at a price that was previously an offer (or bid) where the bid-ask spread is one cent.

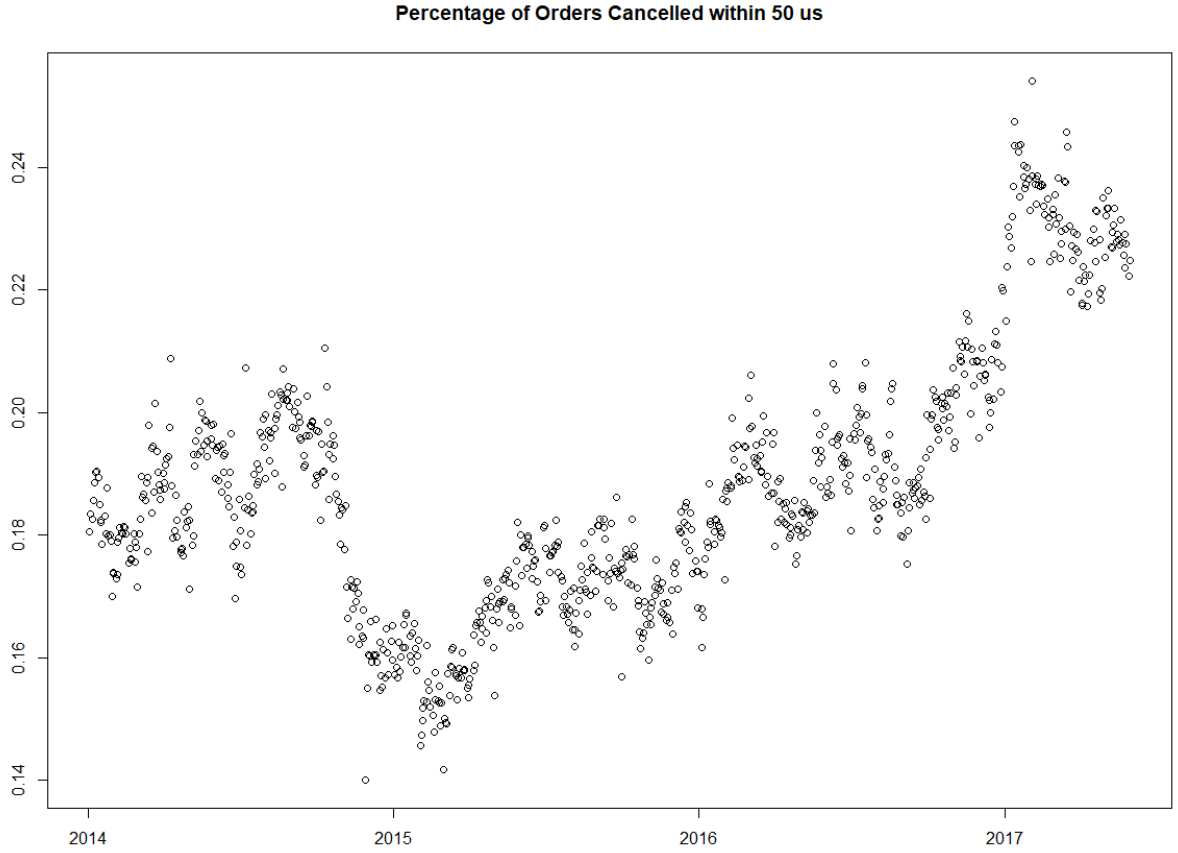
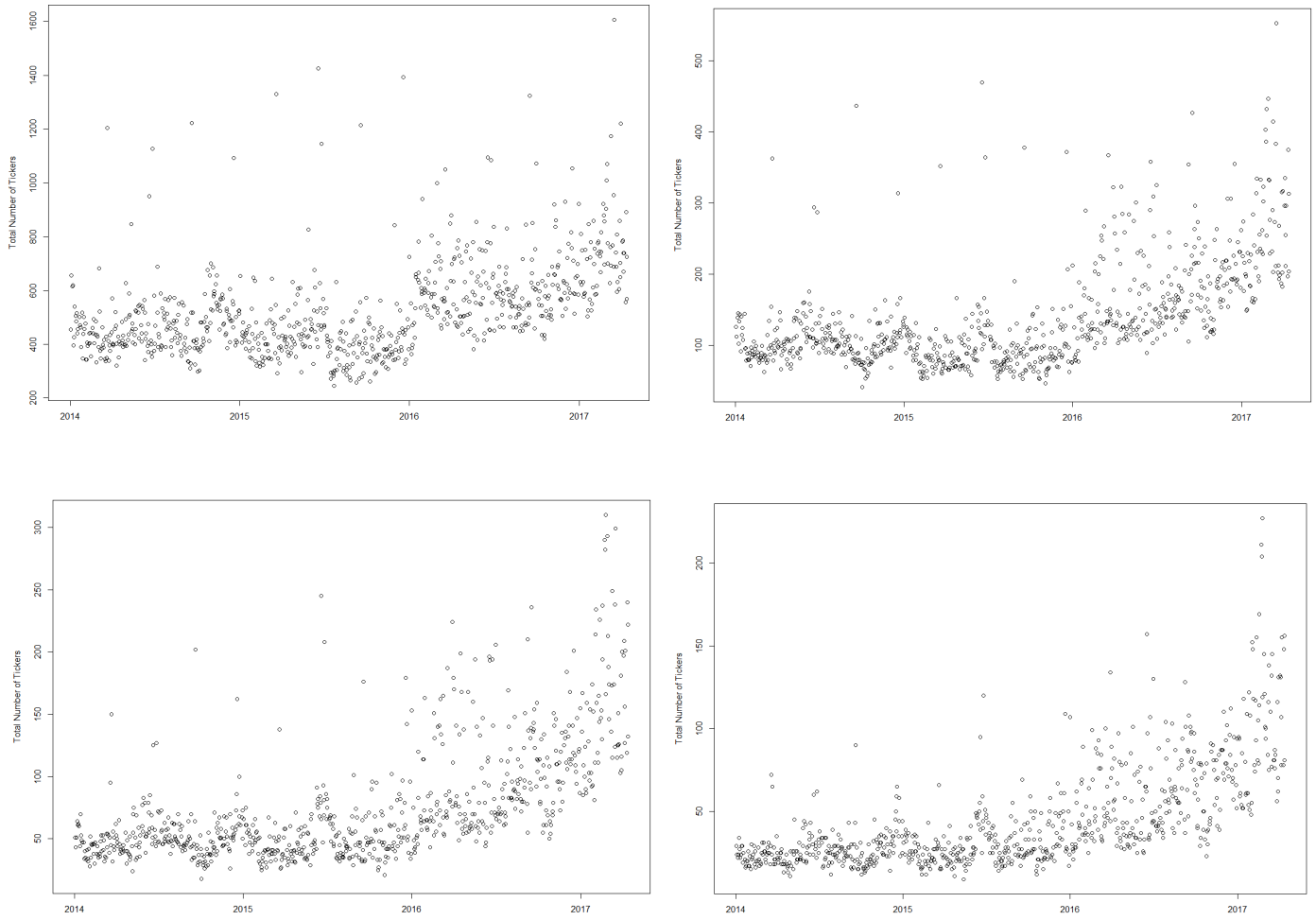


Figure 7. Total Daily Number of Tickers with End-of-Day Order Imbalance

This figure plots the total number of tickers, at a daily frequency, with minimum ending order imbalance >0% (upper left), ≥ 25% (upper right), ≥ 50% (lower left), or ≥ 75% (lower right), based on the final Net Order Imbalance Indicator (NOII) message of the day from January 2014 through April 2017.



Appendix A1: Variable Definition

This table defines and describes the variables used throughout the paper alongside their respective sources.

Variables	Definition	Source
Close-to-Close Volatility	Average absolute % change in closing price across all tickers	ITCH
End-of-Day Volatility	Average absolute % change in price from 3:50 PM to market close across all tickers	ITCH
FIFO Ratio (< 1 μs)	Daily average rate at which the first order placed is also first to be acknowledged by the exchange when consecutive orders are placed less than one *μs* apart	Proprietary
FIFO Ratio (< 2 μs)	Daily average rate at which the first order placed is also first to be acknowledged by the exchange when consecutive orders are placed less than two *μs* apart	Proprietary
Half Day Flag	Equals one on trading days closing at 1:00 PM ET, and zero otherwise	Nasdaq.com
Intra-Day Jitter (μs)	Daily difference between the 80^{th} and 20^{th} percentile of O-A latencies across all orders throughout the day	Proprietary
Inverse Half-Life Ratio	Daily average ratio of the total quantity of shares placed at the onset of a new price-formation speed race scaled by the half-life quantity of shares, where the half-life quantity is identified as the quantity of shares available halfway through the life of a given price level	ITCH
Order-to-Accept (O-A) Latency (μs)	Daily median distance in time (in *μs*) from when an order is placed to when it is acknowledged by the matching engine	Proprietary
Post Upgrade	Equals one on days following the major tech overhaul by Nasdaq (NFF) on May 26, 2016, and zero otherwise	Nasdaq.com
Rebalance Day Flag	Equals one on trading days that fall on (i) an index-rebalancing day, (ii) the last trading day of the month; (iii) the last trading day of the quarter; or (iv) the third Friday of the month	Nasdaq.com
Total Demand at Close	Daily average number of shares requested at close across all tickers	ITCH
Trading Volume	Daily average share volume across all tickers based on trades executed in the continuous market	CRSP

% Order Cancellations (within 50 μs)	Daily average percentage of orders placed at a new price-formation speed race that are cancelled within 50 *μs*	ITCH
% Order Imbalance	At a daily aggregate level, % Order Imbalance is measured as the total unabsorbed on-close orders (obtained from the final NOII message of the day) as a percentage of the total on-close demand (across all tickers with a starting order imbalance as of the first NOII message at 3:50:00 PM) At a ticker-day level, % Order Imbalance is measured as the individual ticker's unabsorbed on-close orders (obtained from the NOII message of the day) as a percentage of its total on-close demand.	ITCH

Appendix A2: Evidence of Continued Technological Disparity

The following figure is an excerpt from a 2016 CSPi marketing brochure for the CSPi ARC Series E-Class network adapter, originally accessed from their website on <http://www.cspi.com/wp-content/uploads/2016/06/Tick-to-Trade-Latency_FINAL-2.pdf>.

Measured Tick-To-Trade Latency in Microseconds, 25,000 Runs						
Network Adapter	Minimum	Mean	Median	99% Less Than	Maximum	Std Deviation
CSPi ARC Series E-Class*	1.487	1.538	1.530	1.603	3.912	.033
SolarFlare Flareon**	1.908	2.006	1.993	2.101	8.638	.054

Numbers you should look for – getting close to zero

As shown in Table 1, our most recent Myricom ARC network adapter is 500 ns or 24% faster than the competition's 99% result. The ARC Series FPGA-based architecture enables continual enhancements and our models show that new generations of silicon and firmware will drive this number down even further.