On-demand market-access, fast trading and the role of decentralized exchanges

Michael Brolley[*]
Wilfrid Laurier University

Marius Zoican[†]
University of Toronto

May 27, 2020

Abstract

Exchanges acquire excess processing capacity to accommodate trading activity surges associated with zero-sum high-frequency trader (HFT) "duels." The idle capacity's opportunity cost is an externality of low-latency trading. We build a model of a decentralized exchange (DEX) with flexible capacity. On DEXs, HFTs acquire speed in real-time from an elastic supply of trade facilitators. The price of speed surges as HFTs compete during activity bursts. Relative to centralized exchanges, HFTs acquire more speed on DEXs, but for shorter timespans. Low-latency "sprints" speed up price discovery without harming liquidity. Overall, speed rents decrease and fewer resources are locked-in to support zero-sum HFT trades.

Keywords: high-frequency trading, FinTech, decentralized exchanges, market design
JEL Codes: G10, G14, G23

[*] Michael Brolley is affiliated with Wilfrid Laurier University, Lazaridis School of Business. E-mail: mbrolley@wlu.ca.

[†] Marius Zoican (corresponding author) is affiliated with University of Toronto Mississauga and Rotman School of Management. Marius can be contacted at marius.zoican@rotman.utoronto.ca. Address: Rotman School of Management, 105 St. George Street, Toronto, Ontario; Canada M5S 3E6. We have greatly benefited from discussion on this research with Carole Comerton-Forde and Nicolas Inostroza. Marius gratefully acknowledges the Connaught Fund for a New Researcher Award and the Global Risk Institute for financial support.

On-demand market-access, fast trading and the role of decentralized exchanges

Abstract

Exchanges acquire excess processing capacity to accommodate trading activity surges associated with zero-sum high-frequency trader (HFT) "duels." The idle capacity's opportunity cost is an externality of low-latency trading. We build a model of a decentralized exchange (DEX) with flexible capacity. On DEXs, HFTs acquire speed in real-time from an elastic supply of trade facilitators. The price of speed surges as HFTs compete during activity bursts. Relative to centralized exchanges, HFTs acquire more speed on DEXs, but for shorter timespans. Low-latency "sprints" speed up price discovery without harming liquidity. Overall, speed rents decrease and fewer resources are locked-in to support zero-sum HFT trades.

Keywords: high-frequency trading, FinTech, decentralized exchanges, market design
JEL Codes: G10, G14, G23

1 Introduction

Electronic markets are driven by technology. Unlike humans, computers make trading decisions with extremely low latency, approaching the speed of light. If algorithms react simultaneously to a trading opportunity, their orders arrive at the market at the same time in so-called "micro-bursts." Indeed, using nanosecond-level message data from Nasdaq, Menkveld (2018) documents that 20% of trades cluster in sub-millisecond intervals. However, micro-bursts are associated with higher adverse selection costs for liquidity providers, consistent with high-frequencies "races" between algorithms reacting to the same trading signal.

Budish, Cramton, and Shim (2015) document that the median duration of an arbitrage opportunity dropped from 97 milliseconds in 2005 to 7 milliseconds in 2015, whereas the dollar profit per arbitrage trade did not change during the same period. In markets with time priority, a trader needs to be at least as fast as its fastest competitor to capture such opportunities. However, since short-term "duels" between high-frequency traders (HFTs) are essentially zero-sum games, they may even harm liquidity (Menkveld and Zoican, 2017). Biais, Foucault, and Moinas (2015) argue that the "arms race" for speed is a socially costly investment.

The low-latency arms race between HFTs also impacts the trading venues themselves. Higher demand for trading speed generates a need for better exchange infrastructure. Trading platforms have to be able to tackle surges in market activity, that is, "micro-bursts," as well as normal market conditions. To avoid order delays, exchanges invest in excess processing capacity, faster computer chips, and bigger data buffers (Yesalavich, 2010). Throughout 2007, at the start of the financial crisis, the New York Stock Exchange (NYSE) increased processing capacity twice: first in February, from 17 to 38 thousand messages a second, and again in August from 38 to 64 thousand messages per second. In 2011, spurred on by the emergence of high-frequency trading, the NYSE platform was able to process four times as much: 250,000 orders every second.[1] More recently, Nasdaq reports in May 2019 that the Securities Information Processor (SIP), which links the U.S. trading venue data into a single feed, has a capacity of 5.6 million messages per second.[2]

Exchanges have an incentive to cater to the low-latency arms race and invest in excess capacity as they can partly extract HFT rents. Trading venues have a local monopoly over their own infrastructure: for instance, you have to pay NYSE a fee for locating your computer next to the NYSE server. Budish, Lee, and Shim (2019) estimate that co-location fees amount to $874M-1024M in 2018, which is of the same order of magnitude as transaction fees charged by trading venues. Therefore, technology is an important source of revenue for exchanges. When trading off cost (i.e., building expensive computer capacity) against performance (message delays for traders),

[1] Sources for this paragraph: After Crash, NYSE Got the Message(s), Wall Street Journal, October 16, 2007; and How Linux Mastered Wall Street, PC World, August 15, 2011.

[2] See: Time is Relative: Where Trade Speed Matters, and Where It Doesn't, Nasdaq, May 30, 2019.

exchanges are biased towards the latter, since HFTs are willing to bear the cost.

Since processing power is a scarce resource, maintaining idle capacity is costly. Menkveld (2018) partitions trading days into microseconds, and finds that multi-trade microseconds occur only in 10% of the sample. MarketDataPeaks – a website that aggregates message rates across all U.S.-based exchanges on a daily basis – documents that the average throughput on a random day is of 3.21M messages per second, surging to 25.2M messages per second within micro-bursts. Therefore, we estimate that as much as 90% of the high-throughput exchange infrastructure is idle for 90% of the typical trading day.

In this paper, we study dynamic pricing of trading infrastructure and the role it can play toward reducing the social costs associated with the low-latency arms race. We build a model of high-frequency trading following Menkveld and Zoican (2017), to which we introduce a real-time market for computer power (CPU) capacity. Currently, demand-based dynamic pricing is successfully implemented in other settings, for example in electricity markets, airline reservation systems, or ride-sharing platforms. Financial markets, however, are unique as the demand for excess capacity is chronically generated by a zero-sum game between fast traders, rather than by fundamental consumer demand shocks. Further, from a technological standpoint, computer processing power for trading needs to be re-priced much more frequently, ideally at sub-second intervals.

Recent FinTech innovations allow for low-latency dynamic pricing of exchange infrastructure. In particular, Section 6 describes the architecture of decentralized exchanges (DEX) currently functioning on the Ethereum blockchain. The limit order book data and the trade matching engine are distributed as computer code in a peer-to-peer network. Each participant in the network ("miner") has a copy of the exchange and may "rent out" spare computer power to process incoming orders, for a fee. When a trading surge occurs, excess demand pushes up the price of computing power, and miners allocate more CPU resources to the exchange. In normal market times, miners can re-route idle computer power towards other, more productive goals.

We compare two market design choices (as in, for example Boulatov and George, 2013). First, we consider a *centralized exchange* where HFTs invest in low-latency technology before trading starts. This setup closely corresponds to the prevailing paradigm in which traders buy co-location services from the exchange on a subscription basis: the investment is sunk at the time of trading. Therefore, at centralized exchanges the equilibrium speed investment and price of CPU power is determined ex ante, and does not depend on real-time market conditions. Second, we consider a *decentralized exchange* where HFTs compete to acquire low-latency in real-time, that is, conditional on observing a profitable trading opportunity. This setup corresponds to cloud- or blockchain-based distributed exchanges as described in Section 6.

We find that a market with dynamic pricing for speed improves the allocation of idle resources. While dynamic pricing does not eliminate high-frequency "duels" altogether, it shortens the time during which resources are committed to these zero-sum HFS races, relative to contemporary markets where capacity is rented and idle until an opportunity arrives. Further, dynamic pricing of trading speed does not harm liquidity. As in Budish, Cramton, and Shim (2015), low-latency races in our model are a zero-sum game. Consequently the bid-ask spread depends only on the relative speed of traders (i.e., the probability of winning the race) and not on the absolute latency. In a symmetric equilibrium, high-frequency snipers invest equal amounts in trading speed and have equal probabilities to win the race: therefore, the bid-ask spread is the same in markets with pre-commitment and on-demand speed.

Further, dynamic pricing of trading speed does not harm liquidity and may improve the speed of price discovery. As in Budish, Cramton, and Shim (2015), low-latency races in our model are a zero-sum game. Consequently the bid-ask spread depends only on relative speed of traders (i.e., the probability of winning the race) and not on the absolute latency. In a symmetric equilibrium, HFTs invest equal amounts in trading speed and have equal probabilities to win the race: therefore, the bid-ask spread is the same in both centralized and decentralized market.

An on-demand market may also improve price discovery. At pre-commitment markets, high-frequency snipers must rent CPU power continuously, for example via co-location fees, even when there are no trading opportunities. Since on-demand markets couple CPU resources to order execution, HFS can spend similar resources to acquire more CPU power for shorter periods of time. In this way, if trading at a per-commitment market resembles a marathon where runners need to pace themselves, trading at an on-demand market mirrors a sprint where speed can surge higher over short intervals. Consequently, the expected time from news until a price update (either via a speculator trade or quote update from the market-maker) is lower in on-demand markets. The result is, of course, conditional on the caveat that HFS can reach the gateway of either exchange with the same latency. Finally, HFTs earn lower rents in on-demand versus pre-commitment markets. Even if the aggregate CPU power used at on-demand markets is lower, high-frequency traders face higher prices during micro-bursts when speed competition intensifies, reducing their rents during speed races around news events. Thus, on-demand markets transfer value from zero-sum HFT races to the suppliers of exchange infrastructure.

We acknowledge that in a market with speed "on-demand", the core assumption that unused capacity can be routed to other tasks when not in use is incompatible within the current market paradigm, where latency-reduction technology is committed on a fixed subscription basis to individual traders or firms, and therefore excluded from alternative productive uses by the exchange. We note, however, that recent FinTech innovations have led to low-latency dynamic pricing of exchange infrastructure in the cryptoasset sphere. Using decentralized exchanges (DEX) functioning on the Ethereum blockchain, limit order book data and the trade matching engine are distributed as computer code in a peer-to-peer network. Each participant in the network ("miner") has a copy of the exchange and may "rent out" spare computer power to process incoming orders, for a fee.

When a trading surge occurs, excess demand pushes up the price of computing power, and miners allocate more CPU resources to the exchange. In normal market times, miners can re-route idle computer power towards other goals. We provide a detailed discussion of this environment in Section 6.

2 Related literature

Our paper contributes to an active discussion on the role of technology in financial market design. It relates to several branches of literature in market microstructure and financial innovation (FinTech).

High-frequency trading. Foucault and Moinas (2019) provide a comprehensive review of the recent literature on fast trading. In recent years, an arms race emerged for ever-faster markets: Baldauf and Mollner (2018) report an average exchange-to-trader latency as low as 31 microseconds for highly liquid New York Stock Exchange instruments.

Closest to our paper, Biais, Foucault, and Moinas (2015) and Budish, Cramton, and Shim (2015) explicitly model investment in low-latency infrastructure on centralized exchanges. They argue that the speed arms race generates excessive investment in low-latency technology, but does not necessarily improve market quality. In both studies, speed investment is a binary decision (i.e., a trader can be either "fast" or "slow") and is sunk before trading starts. Further, the cost of low-latency technology is fixed and corresponds, for instance, to the cost of colocation. In our paper, we model the market for trading infrastructure. We allow for real-time acquisition and pricing of low-latency technology, allowing for speed investment to be contingent on contemporaneous trading data. We argue that such a market design, made possible by peer-to-peer decentralized platforms, leads to a better allocation of limited computer resources and limits the wasteful aspect of the low-latency trading arms race.

Recent empirical evidence suggests that benefits from faster markets flattened out. For example, Chao, Yao, and Ye (2017) find that a drop in exchange latency on NASDAQ from microseconds to nanoseconds reduced market depth. Shkilko and Sokolov (2019) find that when rain disrupts the microwave network connection between Chicago and New York, slowing down the market, liquidity actually improves. However, traders may *individually* benefit from being fast. Baron, Brogaard, Hagströmer, and Kirilenko (2019) document that fast traders earn short-term profits on market orders, consistent with order sniping. Foucault, Kozhan, and Tham (2016) find evidence that high-frequency trading generates toxic cross-market arbitrage in foreign exchange markets. Daian, Goldfeder, Kell, Li, Zhao, Bentov, Breidenbach, and Juels (2019) empirically document that traders on decentralized exchanges engage in bidding wars on fees to obtain time priority.

How does the exchange infrastructure impact market quality? Exchanges may use latency investment strategically: Pagnotta and Philippon (2018) show that trading platforms may invest in speed as a horizontal differentiation tool to relax competition. Such strategic considerations are limited in the case of decentralized exchanges, where the infrastructure is provided competitively by atomistic participants in a peer-to-peer network. Menkveld and Zoican (2017) argue that low-latency exchanges promote zero-sum "duels" between informed HFTs and can lead to lower liquidity.

This paper proposes decentralized markets as a solution to the low-latency arms race. There are several alternative proposals available. Budish, Cramton, and Shim (2015) argue that a discrete-time market with frequent batch auctions eliminates the advantage of being marginally faster than competitors and stimulates price competition between high-frequency traders. Kyle and Lee (2017) propose, at the opposite end of the spectrum, a fully-continuous exchange where traders submit buy or sell trade rates over time, rather than quantities. Speed bumps, that is intentional delays to order execution, are a solution particularly favored by exchanges (see Baldauf and Mollner, 2019, for a list of trading platforms that implemented speed bumps); however, their effectiveness depends on implementation details. Aoyagi (2019) argues that a random delay can incentivize investment in trading speed and worsen adverse selection. Brolley and Cimon (2019) caution that exchanges may choose the design of speed bumps for their trading platforms to maximize exchange profits, which may not coincide with the elimination of the high-frequency arms race. Existing proposals focus therefore either on removing traders' incentives to be fast, or limiting speed directly. We argue for a market-based solution, where speed is priced in real-time, and therefore more costly within trading micro-bursts. Decentralized markets do not eliminate speed races, but rather "localize" them only to instances when speed is required, thereby reducing the time that technology remains idle.

Our model follows Budish, Cramton, and Shim (2015) and Menkveld and Zoican (2017) in that there is no exogenous information asymmetry between high-frequency market-makers and speculators. Indeed, Brogaard, Hendershott, and Riordan (2014) find that fast traders both consume and provide liquidity. Adverse selection is generated by asynchronous arrival at an exchange with time-priority rules, where trading messages are processed in the order in which they reach the exchange. This complements a large body of literature where adverse selection stems from asymmetric information (for example, Glosten and Milgrom, 1985; Foucault, Hombert, and Rosu, 2016).

Financial technology and innovation. The paper also relates to a growing body of research studying the impact of financial technology (FinTech) on trading and market structure.

A majority of Blockchain-driven trading protocols feature a decentralized infrastructure where transaction settlement is implemented by "miners." Cong, He, and Li (2019) argue that in proof-of-work blockchains, where only the fastest miner receives the settlement reward, there are strong incentives to form coalitions (or *mining pools*) to share risk. The ensuing competition between ever larger mining pools inefficiently increases the energy consumption of proof-of-work-based blockchains. Basu, Easley, O'Hara, and Sirer (2019) study the optimal design of miner fees on decentralized markets. They argue for a second-price auction system, which would reduce traders' incentives to bid strategically on transaction fees, and therefore lead to more predictable transaction costs. Biais, Bisière, Bouvard, and Casamatta (2019) argue that information delays and software upgrades may lead to situations where miners "fork" a blockchain, leading to less reliable transaction records. Easley, O'Hara, and Basu (2019) show that transaction fees are not necessarily Pareto-improving. They worry that, in the long run, waiting times and equilibrium fees could be high enough to discourage user participation. We complement these papers by focusing on distributed *trading*: Miners in our model post orders to a P2P-maintained limit book rather than registering already matched trades.

A number of papers focus on the role of Blockchain in market design. Khapko and Zoican (2019) study the impact of distributed ledger on post-trade infrastructure, and find that real-time settlement can reduce liquidity as it increases inventory costs for intermediaries. Chiu and Koeppl (2019) estimate that Blockchain-based settlement could lead to savings of 1-4 basis points per transaction in the U.S. corporate debt market. Malinova and Park (2017) argue that Blockchain can improve trading transparency while limiting front-running risk, and therefore generate welfare gains.

Dynamic pricing. Dynamic pricing, where prices reflect the time-varying demand and marginal costs, is a regular occurrence in other infrastructure exchanges, such as the market for electricity (Joskow and Wolfram, 2012). Cramer and Krueger (2016) use the example of the ride-sharing company Uber to show how a peer-to-peer platforms with dynamic pricing increase the efficiency of car usage, leading to lower costs for riders. Azevedo and Weyl (2016) discuss how advances in information technology can ameliorate the functioning of matching markets. In the same spirit, our paper argues that the benefits of a nimble, real-time, market for trading infrastructures can reduce costs associated with the low-latency arms race.

3 Model

Asset. A single risky asset is traded on a limit order book with price-time priority. At the start of the trading game ($t = 0$), the asset has fundamental value v, which is common knowledge to all market participants. Innovations to the fundamental value ("news") arrive at a Poisson rate δ. Conditional on news, the fundamental value is equally likely to be $v + \sigma$ (i.e., good news) or $v - \sigma$ (i.e., bad news), where $\sigma > 0$ is the news size.

Trading environment. A matching engine operates the limit order book, where it accepts limit orders and "marketable" limit orders submitted by market participants. A limit order is defined as a price quote to either buy (a bid) or sell (an ask) a given amount of the asset. Any unexecuted limit orders are stored in the order book. A "marketable" limit order (either a limit buy order with a price higher than the lowest ask in the book or a limit sell order with a price below the highest bid) is immediately matched with the corresponding bid or ask price, resulting in a trade. We refer to such marketable limit orders simply as market orders.

Traders. There are two types of traders: $H = 2$ high-frequency traders (HFTs) and a large number of liquidity investors (LI). Both trader types are risk-neutral. High-frequency traders submit both marketable and non-marketable limit orders: they compete to supply liquidity, as well as trade on fundamental value innovations. There is no fee for submitting orders to the exchange. HFTs possess a monitoring technology that allows them to perfectly and instantaneously observe changes to the security's fundamental value. Further, they can invest in technology (computer processors) allowing them to access the market faster.

Liquidity investors receive private value shocks at a Poisson rate μ. Conditional on a shock, the private value of a liquidity investor is either η (translating to a buy intention) or $-\eta$ (sell intention), with equal probabilities, where $\eta > \sigma$. Liquidity investors are infinitely impatient and submit only market orders as soon as they are hit with a private value shock. Finally, liquidity investors do not have access to either the monitoring or market-access technologies.

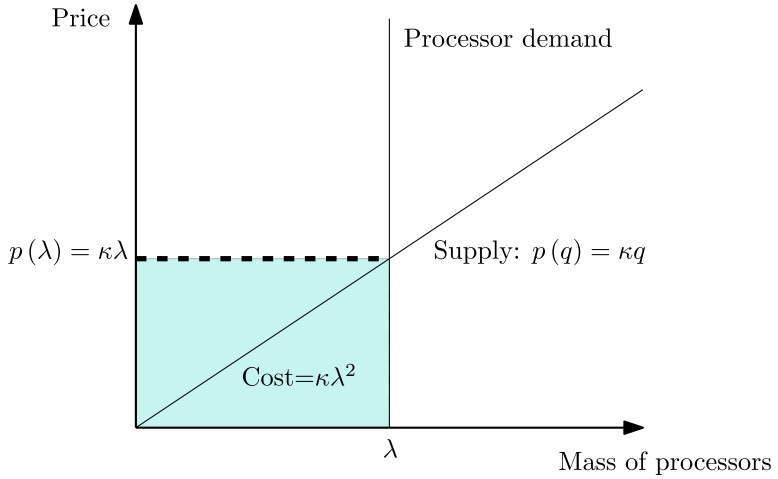
Market Access Technology. High frequency traders have access to a low-latency market access technology provided by suppliers of computer processing power, which we refer to as *processors.*

There is a continuum of competitive processors indexed by j: By lending its resources to HFTs, processor j incurs a linear opportunity cost of $K_j = \kappa \times j$ per unit of time. The parameter κ measures the elasticity of computer power supply: a higher κ translates to a higher marginal cost for technology.

The Poisson arrival rate of HFT messages is proportional to the amount of computer power she controls at a given point in time. In particular, if an HFT rents a mass λ of processors, any submitted orders arrive at the matching engine as a Poisson process with intensity λ.

The technology supply schedule corresponds to the marginal cost function of processors. Therefore, as Figure 1 illustrates, if HFTs demand a processing rate λ, the clearing price for processing power is $k\lambda$, and consequently HFTs spend $k\lambda^2$ on low-latency technology.

Figure 1: **Processor market clearing**



The technology market clears at a uniform price, and quantities are allocated pro rata. In particular, if the two HFTs, i and $-i$, demand order arrival rates λ_i and λ_{-i}, the unique equilibrium technology price is $p(\lambda_i + \lambda_{-i}) = \kappa(\lambda_i + \lambda_{-i})$. It follows that individual HFT investment in processing speed is

$$\begin{aligned} \mathcal{C}_i &= \kappa\lambda_i(\lambda_i + \lambda_{-i}) \text{ and, respectively,} \\ \mathcal{C}_{-i} &= \kappa\lambda_{-i}(\lambda_i + \lambda_{-i}). \end{aligned} \tag{1}$$

Timing. The timing of our setup closely follows that of Menkveld and Zoican (2017). At $t = 0$, the two high-frequency traders compete to provide quotes. Since any liquidity investor demands at most one unit, the HFTs submit limit orders for one unit of the asset. The highest-price buy quote (bid) and the lowest-price sell quote (ask) prevail and are posted to limit order book. We denote the HFT that succeeds in posting their quotes as the high-frequency market-maker (HFM). The other HFT acts as a quote sniper, which we refer to as the high-frequency "bandit" (HFB).

We define a "trigger event" to be the first arrival of either news or a liquidity investor. The trigger event time is random, with an expected value of $\mathbb{E}t_{\text{trigger}} = \frac{1}{\delta+\mu}$. The HFTs immediately react to the trigger event and send out orders to the exchange. With probability $\frac{\mu}{\mu+\delta}$, a liquidity investor consumes the appropriate quote that matches the direction of their trading intentions. Otherwise, with probability $\frac{\delta}{\mu+\delta}$, news arrives. The HFB rushes to submit a market order to the matching engine to snipe the stale quote, while the HFM rushes to cancel the stale quote. Upon the fill or cancellation of the stale quote, the game ends and market participants realize their payoffs.

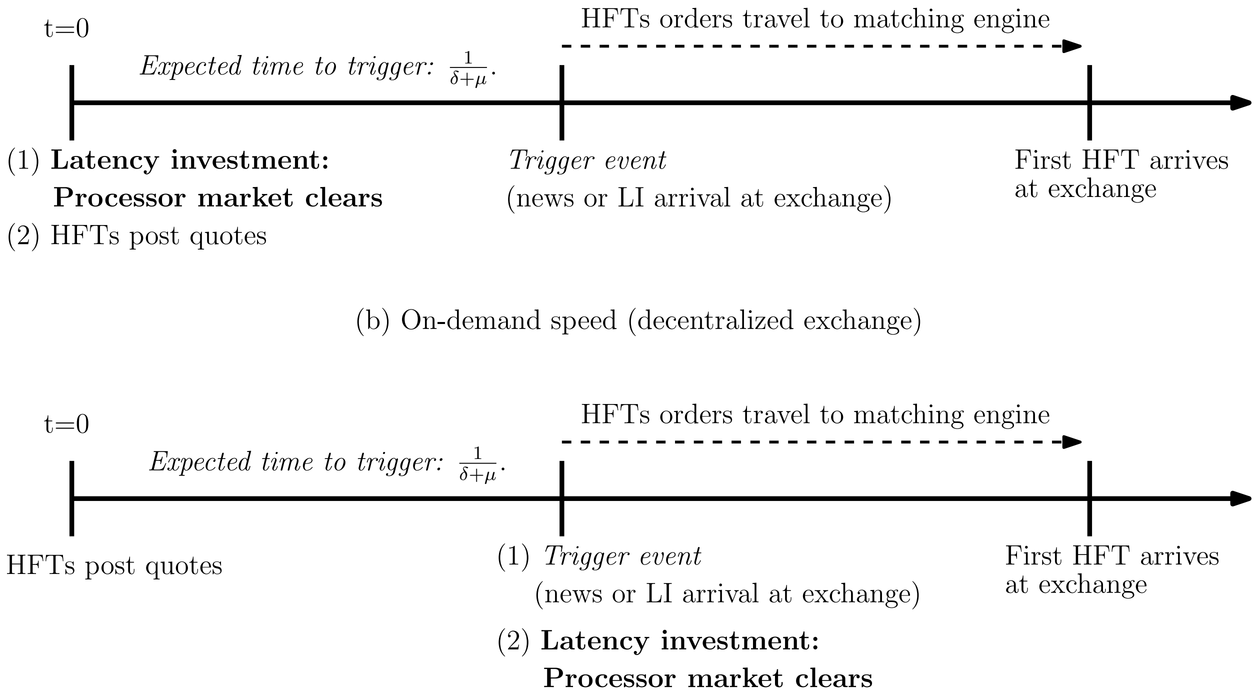
In the spirit of Boulatov and George (2013), we compare two alternative exchange environments corresponding to different clearing times for the technology market. That is, we consider both

(a) a centralized exchange with speed pre-commitment and,

(b) a decentralized exchange with on-demand speed.

The timing of the trading game, in both cases, is illustrated below.

Figure 2: **Model timing**



(a) Speed pre-commitment (centralized exchange)

(b) On-demand speed (decentralized exchange)

t=0
Expected time to trigger: $\frac{1}{\delta+\mu}$.
HFTs orders travel to matching engine
(1) **Latency investment:**
Processor market clears
(2) HFTs post quotes
Trigger event
(news or LI arrival at exchange)
First HFT arrives at exchange
t=0
Expected time to trigger: $\frac{1}{\delta+\mu}$.
HFTs orders travel to matching engine
HFTs post quotes
(1) *Trigger event*
(news or LI arrival at exchange)
(2) **Latency investment:**
Processor market clears
First HFT arrives at exchange

In a *speed pre-commitment* environment, the market for latency infrastructure clears at $t = 0$, before trading starts. The setup echoes the prevailing market practice where centralized exchanges offer traders a menu of (usually monthly) server co-location and direct market access subscriptions. By contrast, in the *on-demand speed* environment, the market for latency infrastructure clears in real-time. High-frequency traders acquire processing power "as-needed," after observing the trigger event. The setup echoes the implementation Ethereum-based distributed exchanges such as EtherDelta. Traders can attach a higher gas fee to their order submissions, offering incentives for miners to prioritize their message over others.

4 Equilibrium

We search for Nash Equilibria that are symmetric in processing speed investment ($\lambda_i = \lambda_{-i}$).

Definition 1 (Equilibrium)**.** An *equilibrium* of the trading game consists of (i) a choice λ_i of low-latency investment for each HFT$_i$, (ii) an HFT quoting strategy at $t = 0$, that is a bid price at which each HFT is ready to buy the asset and an ask price at which each HFT is ready to sell the asset, (iii) HFT orders following the trigger event, conditional on whether the order book features his quotes at the end of $t = 0$, and (iv) a market clearing price for computer processors.

In Section 4.1, we study the equilibrium in a centralized exchange with speed pre-commitment; Section 4.2 focuses on decentralized exchanges with on-demand speed acquisition.

Order book symmetry. For simplicity of exposition, we focus throughout the paper on HFT ask quotes. This is without loss of generality, as buy and sell orders are symmetric as in, for example, Budish, Cramton, and Shim (2015).

To fix intuition, assume first good news arrived and the asset value is $v + \sigma$. A high-frequency bandit will attempt to buy the asset for ask and realize a profit of $v + \sigma - \mathsf{ask}$. Since HFTs engage in a zero-sum game, the corresponding market maker loss is $\mathsf{ask} - \sigma$. Alternatively, assume bad news and therefore the asset value is $v - \sigma$. In this case, the HFB attempts to sell the asset at the bid and make an identical profit of $v - \mathsf{ask} - (v + \sigma) = \sigma - \mathsf{ask}$. The two scenarios mirror each other: a successful snipe yields a profit of $\sigma - \mathsf{ask}$ to the HFB and a loss of $\mathsf{ask} - \sigma$ to the market-maker. Similarly to this intuition, the public value of the security v at $t = 0$ cancels our in our analysis that follows in Sections 4.1 and 4.2. Thus, for simplicity, we normalize $v = 0$ in our subsequent analysis. Finally, we use subscripts 'M' and 'B' denote variables and functions pertaining to HFM and HFB, respectively.

4.1 Centralized exchange environment (pre-committed speed investment)

In this section, we assume that investment in processing speed by HFT$_i$ given by $\mathcal{C}_i(\lambda)$ in equation (1) is made at the beginning of $t = 0$, just prior to the start of the trading game where HFTs rush to post quotes. We view this pre-committment to processing speed–denoted *PC* –as similar to the current centralized exchange paradigm: firms pay co-location subscription fees to access the exchange with reduced latency, where access persists for a length of time such that speed investment is a sunk cost with respect to any one trading day that begins at $t = 0$.

We solve this game via backward induction. At the quoting race stage following speed investment at $t = 0$, HFTs race to post their quotes to the matching engine to become the HFM. Because HFTs i and $-i$ are both rushing to post quotes but do not yet know their roles in the game, the expected payoff from participating as the HFM is written as,

$$
\begin{aligned}
\pi_0^{\mathrm{M}} &= \mathsf{Pr}(\text{news}) \times \mathsf{Pr}(\text{quote sniped}) \times (\mathsf{ask} - \sigma) + (1 - \mathsf{Pr}(\text{news})) \times \mathsf{ask} \\
&\quad - \mathbb{E}[\text{processor rental time} \mid \lambda_i] \times \mathcal{C}_i(\lambda_i), \\
&= -\frac{\delta}{\delta+\mu}\frac{\lambda_{-i}}{\lambda_i+\lambda_{-i}}(\sigma - \mathsf{ask}) + \frac{\mu}{\delta+\mu}\mathsf{ask} - \mathbb{E}[\text{processor rental time} \mid \lambda_i] \times \mathcal{C}_i(\lambda_i). \qquad (2)
\end{aligned}
$$

where the term $\frac{\lambda_{-i}}{\lambda_i+\lambda_{-i}}$ reflects the probability that should HFT$_i$ become the HFM, HFT$_i$ loses the race to cancel the stale quote ask following the arrival of news, and $\mathbb{E}[\text{processor rental time} \mid \lambda_i] \times \mathcal{C}_i(\lambda_i)$ reflects the total expected processing expenditure (i.e., duration multiplied by speed intensity). With probability $\frac{\mu}{\delta+\mu}$, a liquidity investor arrives to the market before news, buys the asset, and the HFM receives the ask price.

If HFT$_i$ becomes the HFB, they will seek to snipe stale quotes in the event of news arrival. As such, their expected payoff is given by,

$$
\begin{aligned}
\pi_0^{\mathrm{B}} &= \mathsf{Pr}(\text{news}) \times \mathsf{Pr}(\text{quote sniped}) \times (\sigma - \mathsf{ask}) - \mathbb{E}[\text{processor rental time} \mid \lambda_i] \times \mathcal{C}_i(\lambda_i), \\
&= \frac{\delta}{\delta+\mu}\frac{\lambda_i}{\lambda_i+\lambda_{-i}}(\sigma - \mathsf{ask}) - \mathbb{E}[\text{processor rental time} \mid \lambda_i] \times \mathcal{C}_i(\lambda_i). \qquad (3)
\end{aligned}
$$

In equation (3), since HFT$_i$ is the bandit, the snipe probability is $\frac{\lambda_i}{\lambda_i+\lambda_{-i}}$, that is, the probability that HFT$_i$ is the first to arrive at the market after news.

What is the equilibrium spread? To ensure both HFTs are indifferent to becoming an HFM or an HFB, HFTs will quote a price $\mathsf{ask}^\star$ such that the expected profit functions in (2) and (3) are equal. Intuitively, if HFT$_i$ instead posts $\mathsf{ask}^+ > \mathsf{ask}^\star$, then HFT$_{-i}$ always becomes the market-maker since she quotes the tighter spread. Therefore, HFT$_i$ earns the HFB expected profit for a spread of $\mathsf{ask}^\star$ and is consequently indifferent between deviating or not.[3] Similarly, if HFT$_i$ quotes a tighter spread, that is $\mathsf{ask}^- < \mathsf{ask}^\star$, she becomes the market maker for sure. However, the deviation is not profitable since,

$$
\pi_0^{\mathrm{M}}\left(\mathsf{ask}^-\right) < \alpha\pi_0^{\mathrm{M}}\left(\mathsf{ask}^\star\right) + (1-\alpha)\,\pi_0^{\mathrm{B}}\left(\mathsf{ask}^\star\right), \forall\alpha \in [0,1]\,. \qquad (4)
$$

Setting (2) and (3) equal and solving for $\mathsf{ask}^\star$, we obtain,

$$
\begin{aligned}
-\frac{\delta}{\delta+\mu}\frac{\lambda_{-i}}{\lambda_i+\lambda_{-i}}(\sigma - \mathsf{ask}) + \frac{\mu}{\delta+\mu}\mathsf{ask} &= \frac{\delta}{\delta+\mu}\frac{\lambda_i}{\lambda_i+\lambda_{-i}}(\sigma - \mathsf{ask}) \\
\iff \frac{\mu}{\delta+\mu}\mathsf{ask} &= \frac{\delta}{\delta+\mu}\frac{\lambda_i+\lambda_{-i}}{\lambda_i+\lambda_{-i}}(\sigma - \mathsf{ask}) \\
\iff \mathsf{ask}^\star &= \frac{\delta}{\delta+\mu}\sigma. \qquad (5)
\end{aligned}
$$

[3] The intuition here follows the proof of Proposition 1 from Menkveld and Zoican (2017), p. 1217.

Note that the cost of investment in speed from (2)-(3) cancels out when we compare the two payoff functions at the quoting stage, as the investment is sunk once the quoting game begins at $t = 0$. In fact, equation (5) is not only independent of the expected processing speed cost, but—as in Budish, Cramton, and Shim (2015)—the quoted spread $\mathsf{ask}^\star$ is independent of speed levels altogether. The HFT "duel" is essentially a zero-sum game. Any investment in speed that reduces the probability of adverse selection to HFT_i in their role as a market-maker also increases their ability to snipe a stale quote as a bandit at the same rate, $\frac{\lambda_{-i}}{(\lambda_i+\lambda_{-i})^2}$. Moreover, the quoting decision of HFT_i is impacted similarly by the speed investment of HFT_{-i}. If HFT_{-i} invests more in speed, then HFT_i's payoff as either market maker or bandit decreases at the same rate, that is $\frac{\lambda_i}{(\lambda_i+\lambda_{-i})^2}$. Hence, the equilibrium quoted price does not depend on the initial investment in speed by both HFTs.

Prior to the quoting game, HFTs commit resources to rent market access technology from processors at the beginning of $t = 0$. Market access technology is rented until the game ends following a trigger event. We denote the (random) trigger event time as $\tau_{\text{trigger}} = \min\{\tau_{\text{news}}, \tau_{\text{LI}}\}$. If the trigger event is an LI arrival, then the game stops immediately following the trigger event. Conversely, if the trigger event is news arrival, the game continues until the event time $\tau_{\text{HFT race}}$ that the stale quote is either consumed or updated. Formally, we write the expected processor rental duration for an HFT as:

$$\begin{aligned}\mathbb{E}\left[\tau_{\text{trigger}} + \mathbb{1}_{\tau_{\text{news}}\leq\tau_{\text{LI}}}\tau_{\text{HFT race}} \mid \lambda_i\right] &= \int_0^\infty \left[\int_0^y \left(x + \frac{1}{\lambda_i+\lambda_{-i}}\right)\delta e^{-\delta x}\,\mathrm{d}x + y\int_y^\infty \delta e^{-\delta x}\,\mathrm{d}x\right]\mu e^{-\mu y}\,\mathrm{d}y,\\ &= \frac{1}{\delta+\mu} + \frac{\delta}{\delta+\mu}\frac{1}{\lambda_i+\lambda_{-i}}. \end{aligned} \tag{6}$$

The expected duration of processor use simplifies to equation (6), which sums the expected duration of two events: the expected time until the first event $\left(\frac{1}{\delta+\mu}\right)$, and the expected duration of the HFT race following the arrival of news $\left(\frac{1}{\lambda_i+\lambda_{-i}}\right)$, which occurs with probability $\frac{\delta}{\delta+\mu}$.

Given the equilibrium quote $\mathsf{ask}^\star$ from (5) and the expected processor rental duration from (6), we write the HFT_i payoff function at $t = 0$ denoted π_0^{i} as,

$$\begin{aligned}\pi_0^{\text{i}}(\lambda_i) &= \mathsf{Pr}(i = M \mid \lambda_i)\pi_0^{\text{M}}(\lambda_i) + \mathsf{Pr}(i = B \mid \lambda_i)\pi_0^{\text{B}}(\lambda_i) - \mathbb{E}[\text{processor rental time} \mid \lambda_i] \times \mathcal{C}_i(\lambda_i)\\ &= \frac{\lambda_i}{\lambda_{-i}+\lambda_i}\frac{\delta\mu\sigma}{(\delta+\mu)^2} - \left(\frac{1}{\delta+\mu} + \frac{\delta}{\delta+\mu}\frac{1}{\lambda_i+\lambda_{-i}}\right) \times \lambda_i\kappa\left(\lambda_i+\lambda_{-i}\right). \end{aligned} \tag{7}$$

The simplification in (7) follows from the fact that $\pi_0^{\text{M}}(\mathsf{ask}^\star) = \pi_0^{\text{B}}(\mathsf{ask}^\star)$. To solve for the optimal processing speed intensity λ_i, we take the first-order condition of (7) and solve for the fixed point under the assumption of symmetric investment intensities ($\lambda_M = \lambda_B$). Proposition 1 describes the equilibrium that obtains.

Proposition 1 (Symmetric Pre-Commitment Equilibrium). *Let* $(\delta, \mu, \sigma, \kappa) \in (0, \infty)^4$. *There exists a* *unique Nash Equilibrium in the sense of Definition 1 where* $\mathsf{ask}^{\star}$ *is as in Equation (5), and* $\lambda_i = \lambda_{-i} = \lambda^{\star}_{\mathrm{PC}} \in (0, \infty)$ *is the maximum of Equation (7). Moreover,* $\lambda^{\star}_{\mathrm{PC}}$ *is given by:*

$$\lambda^{\star}_{\mathrm{PC}} = \frac{-\delta + \sqrt{\delta^2 + \frac{3\delta\mu\sigma}{\kappa(\delta+\mu)}}}{6}. \tag{8}$$

Proof. The beginning of the proof follows from the discussion above. What remains is to take the first-order condition of equation (7) with respect to λ_i and invoking symmetry of speed intensity ($\lambda_i = \lambda_{-i} = \lambda^{\star}_{\mathrm{PC}}$) to show that the resulting solution $\lambda^{\star}_{\mathrm{PC}}$ is unique and positive.

$$\text{F.O.C (7): } \frac{\lambda_i\left(\delta\mu\sigma - (\lambda_i + \lambda_{-i} + \delta)\kappa(\lambda_i + \lambda_{-i})(\delta + \mu)\right)}{(\delta+\mu)^2(\lambda_i - \lambda_{-i})} = 0, \tag{9}$$

$$\iff \lambda^{\star}_{\mathrm{PC}} = \frac{-\delta + \sqrt{\delta^2 + \frac{3\delta\mu\sigma}{\kappa(\delta+\mu)}}}{6}. \tag{10}$$

where $\lambda^{\star}_{\mathrm{PC}}$ is the unique positive root, as the second root is negative and thus inadmissible. □

4.2 Decentralized exchange environment (on-demand speed investment)

Consider an environment in which HFTs invest in "on-demand" (*OD*) processing speed $\mathcal{C}_i(\lambda)$: HFTs submit a fee to rent processing speed at the same time that they submit an order to the exchange. This environment captures the essence of a decentralized exchange, where a group of trade facilitators—"processors" in the language of our model—provide order-processing services for a fee. In a decentralize exchange, an HFT submits an order to a processor with the requisite fee such that the order is ferried to the matching engine. In our setup, HFT_i will choose a Poisson arrival rate λ_i, where the intensity of processing speed reflects the mass of processors that the HFT pays to execute their order. The intuition is that the greater mass of processors that the HFT rents to submit their order, the greater the probability that their order is given priority over any competitors.

With on-demand speed investment, an investment decision is made following the quoting game, which happens only following the arrival of news. Similar to the centralized exchange case, the news arrival triggers a rush by the HFB to snipe the stale quote. To improve his chances of doing so, the HFB chooses the mass of processors λ_B to maximize the expected payoff from attempting to snipe the stale quote,

$$\begin{aligned} \pi_0^{\mathrm{B}}(\lambda_B \mid \text{news arrival}) &= \mathsf{Pr}(\text{quote sniped}) \times (\sigma - \mathsf{ask}) - \mathbb{E}[\text{processor rental time} \mid \lambda_B]\mathcal{C}_B(\lambda_B), \\ &= \frac{\lambda_B}{\lambda_B + \lambda_M}(\sigma - \mathsf{ask}) - \frac{1}{\lambda_B + \lambda_M}\lambda_B\kappa\left(\lambda_B + \lambda_M\right). \end{aligned} \tag{11}$$

Similarly, the arrival of news will trigger a reaction by the HFM to rush to cancel their stale quote. The HFM rents a mass of processors λ_M to minimize the cost of being sniped:

$$\begin{aligned}\pi_0^{\mathrm{M}}(\lambda_M \mid \text{news arrival}) &= \mathsf{Pr}(\text{quote sniped}) \times (\mathsf{ask} - \sigma) - \mathbb{E}[\text{processor rental time} \mid \lambda_M]\mathcal{C}_M(\lambda_M), \\ &= \frac{\lambda_B}{\lambda_B + \lambda_M}(\mathsf{ask} - \sigma) - \frac{1}{\lambda_B + \lambda_M}\lambda_M \kappa (\lambda_B + \lambda_M). \qquad (12)\end{aligned}$$

Taking the first-order conditions of (11) and (12), we obtain,

$$\text{F.O.C (11): } \frac{\lambda_M}{(\lambda_B + \lambda_M)^2}(\sigma - \mathsf{ask}) - \kappa = 0, \qquad (13)$$

$$\text{F.O.C (12): } \frac{\lambda_B}{(\lambda_B + \lambda_M)^2}(\sigma - \mathsf{ask}) - \kappa = 0. \qquad (14)$$

We note that first-order conditions (13) and (14) are symmetric and solve for the fixed point, $\lambda_M^\star = \lambda_B^\star = \lambda_{\mathrm{OD}}^\star$,

$$\lambda_{\mathrm{OD}}^\star = \frac{\sigma - \mathsf{ask}}{4\kappa}. \qquad (15)$$

Next, by backward induction, we analyze the quoting decision at $t = 0$, and solve for $\mathsf{ask}^\star$, taking into account the optimal speed investment at the trigger time τ_{trigger}. An HFT_i selects their quoting strategy conditional on the anticipated outcome of the sniping/cancelling game played by the HFM and HFB. Similarly to Section 4.1, the HFT_i chooses the quote $\mathsf{ask}^\star$ such that the expected payoff to becoming either an HFM or an HFB are equal. Evaluating the payoffs $\pi_0^{\mathrm{M}}(\lambda_i)$ and $\pi_0^{\mathrm{B}}(\lambda_i)$ at $\lambda_{\mathrm{OD}}^\star$ yields,

$$\pi_0^{\mathrm{M}}(\lambda_{\mathrm{OD}}^\star) = \frac{\delta}{\delta + \mu}\frac{(\mathsf{ask} - \sigma)}{2} + \frac{\mu}{\delta + \mu}\mathsf{ask} - \frac{\delta}{\delta + \mu}\frac{\sigma - \mathsf{ask}}{4\kappa}\kappa. \qquad (16)$$

$$\pi_0^{\mathrm{B}}(\lambda_{\mathrm{OD}}^\star) = \frac{\delta}{\delta + \mu}\frac{(\sigma - \mathsf{ask})}{2} - \frac{\delta}{\delta + \mu}\frac{\sigma - \mathsf{ask}}{4\kappa}\kappa. \qquad (17)$$

Solving for $\mathsf{ask}^\star$ such that $\pi_0^{\mathrm{M}}(\mathsf{ask}^\star; \lambda_{\mathrm{OD}}^\star) = \pi_0^{\mathrm{B}}(\mathsf{ask}^\star; \lambda_{\mathrm{OD}}^\star)$ we obtain,

$$\mathsf{ask}^\star = \frac{\delta}{\delta + \mu}\sigma. \qquad (18)$$

Taken together, ask^* into λ_{OD}^* yield the following proposition.

Proposition 2 (Symmetric On-Demand Equilibrium). *Let $(\delta, \mu, \sigma, \kappa) \in (0, \infty)^4$. There exists a unique Nash Equilibrium in the sense of Definition 1 where $\mathsf{ask}^\star$ is as in Equation (18), and $\lambda_i = \lambda_{-i} = \lambda_{OD}^\star \in (0, \infty)$ solves the system of first-order conditions (11)-(12). Moreover, $\lambda_{OD}^\star$ is given by:*

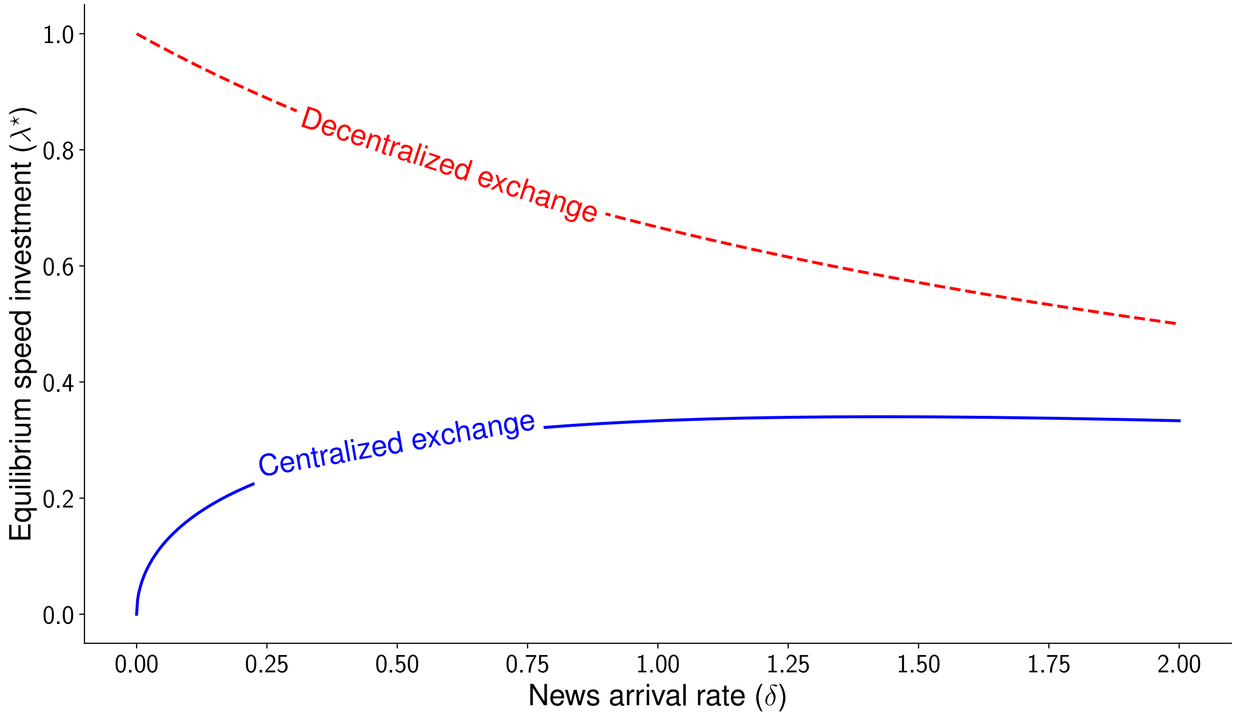
$$\lambda_{OD}^\star = \frac{\mu\sigma}{4\kappa(\delta + \mu)}. \qquad (19)$$

Proof. The proof follows the discussion above; equation (19) results from inputting $\mathsf{ask}^\star$ from equation (18) into the expression for $\lambda^\star_{\mathrm{OD}}$ in equation (15). □

Figure 3 illustrates the result in Propositions 1 and 2. First, HFTs acquire more processing power in decentralized relative to centralized exchanges. Second, HFT invest more (less) in speed on centralized (decentralized) markets as the news arrival rate increases. A higher news rate generates two effects. On the one hand, it increases the likelihood of a sniping opportunity, and therefore the value of trading speed. On the other hand, it translates to a wider bid-ask spread and lower sniping profits, conditional on news. On a centralized exchange, the first effect dominates and the value of speed increases in news frequency. In contrast, on a decentralized exchange, speed investment is incurred *conditional* on news arrival, and consequently only the second effect persists.

Figure 3: **HFT Speed Investment**

This figure illustrates the equilibrium HFT investment in low-latency technology measured by the Poisson arrival rate λ, as a function of the news arrival rate δ. Parameter values: $\mu = 2$, $\kappa = 0.25$, and $\sigma = 1$.



5 Impact on market quality

In this section, we compare the equilibrium outcomes of the centralized and decentralized environments to examine how changing the timing of HFT investment in low-latency technology translates to changes in liquidity, price discovery, and the total amount of resources allocated to the speed race.

Liquidity and Price discovery. In Sections 4.1 and 4.2, we obtain that both the centralized and decentralized markets lead HFTs to an identical equilibrium quoting strategy ($\mathsf{ask}^{\star}_{\text{PC}} = \mathsf{ask}^{\star}_{\text{OD}}$) that is independent of their investment in speed intensity λ.

Corollary 1 (Liquidity)**.** *The bid-ask spread* $\mathsf{ask} - \mathsf{bid}$ *is equal to* $2 \times \frac{\delta\sigma}{\delta+\mu}$ *in both the centralized and decentralized market setting.*

The result in Corollary 1 echoes the model of Budish, Cramton, and Shim (2015), who compare a continuous limit order market to frequent batch auctions. In both models, since the speed race is a zero-sum game between high-frequency traders, the bid-ask spread does not depend on absolute latency levels. Instead, the bid-ask spread depends solely on the magnitude of adverse selection costs.

Both the HFB and HFM spend resources to compete in the sniping/cancelling race that commences following the arrival of news. In a centralized market, resource commitment to processing speed takes place before quotes are posted, leaving the possibility that HFTs will commit resources to a speed race that never begins (i.e., if a liquidity trader arrives before news). HFTs in a decentralized market, however, invest in low-latency technology only after a speed race is triggered by news arrival. Using the equilibrium values for speed intensity across the two markets, we evaluate the difference in speed intensity $\lambda^{\star}_{\text{PC}} - \lambda^{\star}_{\text{OD}}$ from equations (8) and (19), as well as the difference in speed prices $p(\lambda^{\star}_{\text{PC}}) - p(\lambda^{\star}_{\text{OD}})$ to arrive at the following proposition.

Proposition 3 (Speed Intensity)**.** *The equilibrium speed intensity* $\lambda^{\star}$ *and price of speed* $p(\lambda^{\star})$ *are higher in the decentralized market than in the centralized market.*

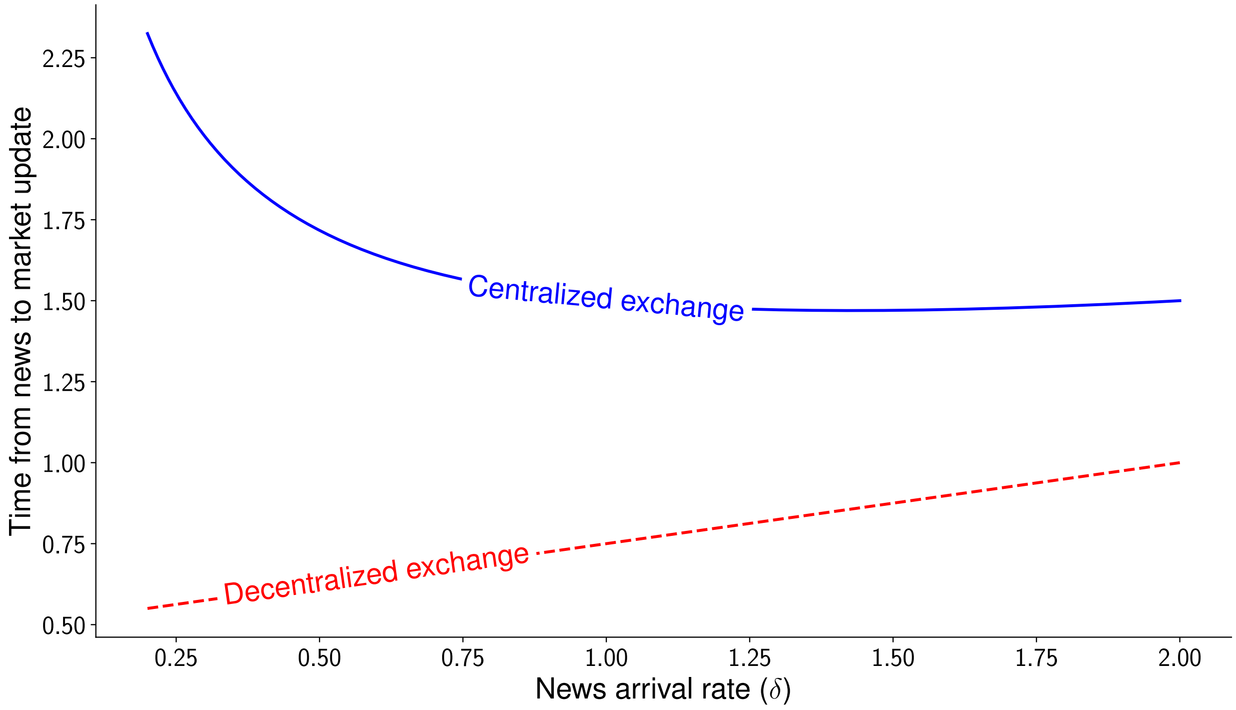
Proof. See Appendix. □

Proposition 3 aligns with intuition that HFTs will commit more resources to speed intensity—and pay a higher average per-unit price to do so—if they know that the speed race will occur. In a centralized market where HFTs must pre-commit to processing capacity knowing that said resources may remain idle, the competition for speed intensity is lower.

Irrespective of the market environment, the outcome of the speed race is identical: stale quotes are removed from the market to make way for quotes that incorporate new information, thereby contributing to the price discovery process. We can thus define a measure of price discovery that describes the expected race-time, and thus measure how quickly, on average, news is impounded into prices. From the time of news arrival, the expected time until the *first* order arrives is given by

Figure 4: **Speed of Price Discovery**

This figure illustrates the expected time elapsed between a news event and the next trade or quote update, in both centralized and decentralized exchanges, as a function of the news arrival rate δ. Parameter values: $\mu = 2$, $\kappa = 0.25$, and $\sigma = 1$.



$\frac{1}{2\lambda^\star}$, which is the inverse of the sum of the equilibrium speed intensity investment by both HFTs. Thus, taking the difference $\frac{1}{2\lambda^\star_{\text{PC}}} - \frac{1}{2\lambda^\star_{\text{OD}}}$, we can infer from Proposition 3 that the greater competition for speed intensity ($\lambda^\star_{\text{OD}} > \lambda^\star_{\text{PC}}$) in the decentralized market leads to faster price discovery relative to the centralized market. We summarize our price discovery result in Corollary 2 below.

Corollary 2 (Price Discovery)**.** *The expected time required for news to impound into prices is shorter in a decentralized market than in a centralized market.*

Figure 4 illustrates the result in Corollary 2: price discovery is faster on decentralized exchanges since HFTs engage in a more intense arms race (albeit for shorter intervals). One caveat remains that since decentralized exchange do not allow for co-location (since the infrastructure is cloud-based), the trader-to-exchange latency is higher than at centralized exchanges. Corollary 2 implies that this effect is at least partly compensated by higher investments in on-demand speed on decentralized exchanges.

Total Resource Usage and Rents from the Speed Race. Though a decentralized market leads the competition for processing speed to intensify at the point when processing resources are acquired, the length of time for which these resources need to be retained is a key advantage of a decentralized market: an HFT rents resources to process a single order, and upon order completion or cancellation, the resources are released.

Using the equilibrium HFT investment in processing speed intensity $\lambda^{\star}_{\text{PC}}$, we compute total resource usage in a centralized exchange environment. We denote total resource usage as $\Lambda^{\star}_{PC}$, which computes twice the per-HFT processor speed intensity multiplied by the average processor rental time $\mathbb{E}$[processor rental time | $\lambda^{\star}_{\text{PC}}$]:

$$\begin{aligned}\Lambda^{\star}_{PC} &\equiv 2\lambda^{\star}_{\text{PC}} \times \mathbb{E}[\text{processor rental time} \mid \lambda^{\star}_{\text{PC}}],\\ &= 2\lambda^{\star}_{\text{PC}} \times \left(\frac{1}{\delta+\mu} + \frac{\delta}{\delta+\mu}\frac{1}{2\lambda^{\star}_{\text{PC}}}\right),\\ &= \frac{\delta}{\delta+\mu} + \frac{\sqrt{\delta^2 + \frac{3\delta\mu\sigma}{\kappa(\delta+\mu)}} - \delta}{3(\delta+\mu)}. \end{aligned} \tag{20}$$

On-demand total resource usage, denoted by Λ^{*}_{OD}, is given by,

$$\Lambda^{*}_{\text{OD}} = 2\lambda^{\star}_{\text{OD}} \times \left(\frac{\delta}{\delta+\mu}\frac{1}{2\lambda^{\star}_{\text{OD}}}\right) = \frac{\delta}{\delta+\mu}. \tag{21}$$

Because speed intensity and time-to-execution following news arrival are perfectly inversely-related, resource usage in the event of news is identical in both environments, regardless of the investment in speed intensity. On a centralized exchange, however, pre-commitment generates excessive resource consumption: both HFTs rent a mass $\lambda^{\star}_{\text{PC}}$ of processors which are rented-but-idle from $t = 0$ until the a trigger event occurs, a period with an expected length of $\frac{1}{\delta+\mu}$.

Corollary 3 (Resource Usage)**.** *The expected total resource allocation to processing speed is lower in the decentralized market than the centralized market,* $\Lambda^{\star}_{PC} > \Lambda^{\star}_{OD}$.

Figure 5: **Total Infrastructure Committed to Low-Latency Trading**

This figure illustrates the total usage of exchange infrastructure resources, in both centralized and decentralized markets, as a function of the news arrival rate δ. Parameter values: $\mu = 2$, $\kappa = 0.25$, and $\sigma = 1$.

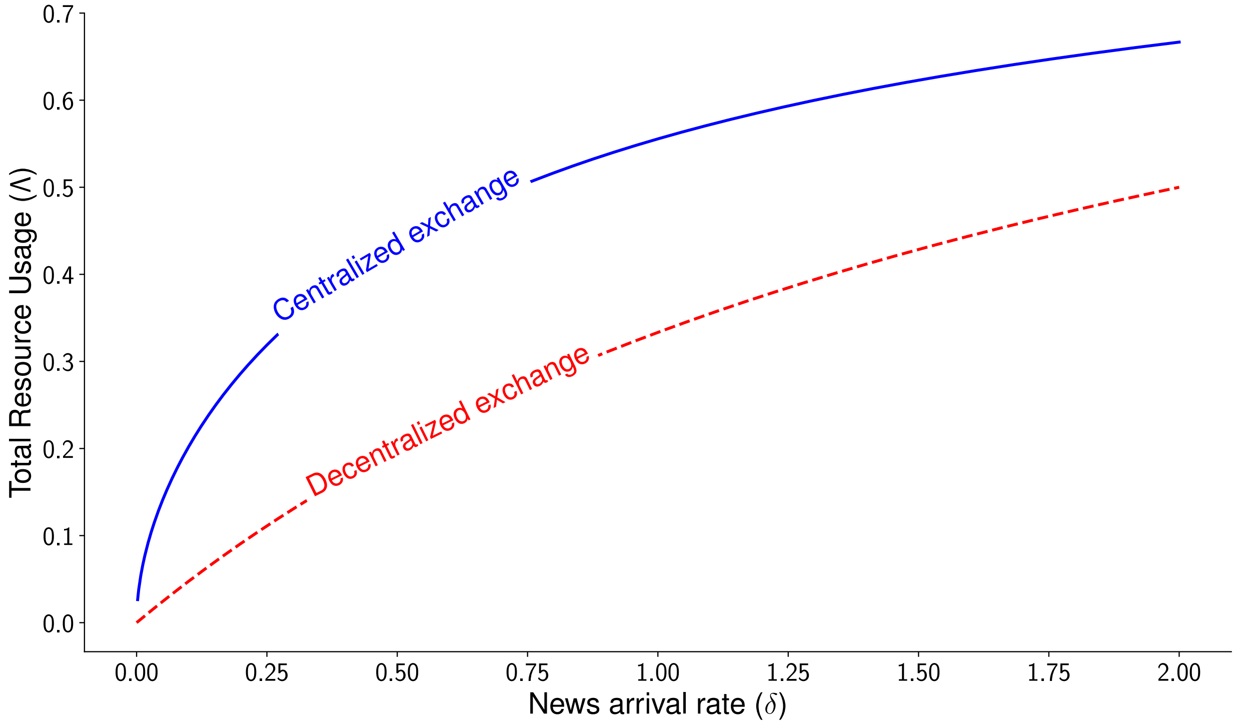


Figure 5 illustrates Corollary 3. As the news arrival rate increases, the expected resource usage increases in both centralized and decentralized markets, as HFT duels become more likely. On the one hand, HFTs acquire more CPU power upon observing news in decentralized markets; however, infrastructure is used more efficiently, with CPU power being rented only when a sniping opportunity emerges. We predict that the second effect dominates, yielding lower resource usage at decentralized exchanges.

Finally, we evaluate how on-demand speed in decentralized markets impacts the rents earned by high frequency traders from the speed race. First, we evaluate the equilibrium HFT profit π_0^i within the centralized market environment, from which we obtain:

$$\pi_0^{\mathrm{i}}(\lambda^\star_{\mathrm{PC}}) = \frac{\lambda^\star_{\mathrm{PC}}}{\lambda^\star_{\mathrm{PC}} + \lambda^\star_{\mathrm{PC}}} \frac{\delta\mu\sigma}{(\delta+\mu)^2} - \left(\frac{1}{\delta+\mu} + \frac{\delta}{(\delta+\mu)} \frac{1}{2\lambda^\star_{\mathrm{PC}}}\right) \times \lambda^\star_{\mathrm{PC}}\kappa\left(\lambda^\star_{\mathrm{PC}} + \lambda^\star_{\mathrm{PC}}\right), \tag{22}$$

$$= \frac{\delta\kappa}{18(\delta+\mu)} \times \left(\delta + \frac{6\mu\sigma}{\kappa(\delta+\mu)} - \sqrt{\delta^2 + \frac{3\mu\delta\sigma}{\kappa(\delta+\mu)}}\right). \tag{23}$$

Similarly, evaluating π_0^i at $\lambda^\star_{\mathrm{OD}}$ and $\mathsf{ask}^\star$, we obtain the equilibrium HFT profit under the decentralized market environment:

$$\pi_0^{\mathrm{i}}(\lambda_{\mathrm{OD}}^{\star}) = \frac{\lambda_{\mathrm{OD}}^{\star}}{\lambda_{\mathrm{OD}}^{\star} + \lambda_{\mathrm{OD}}^{\star}} \frac{\delta\mu\sigma}{(\delta+\mu)^2} - \frac{\delta}{\delta+\mu} \frac{1}{2\lambda_{\mathrm{OD}}^{\star}} \kappa \lambda_{\mathrm{OD}}^{\star} (\lambda_{\mathrm{OD}}^{\star} + \lambda_{\mathrm{OD}}^{\star}), \tag{24}$$

$$= \frac{\delta\mu\sigma}{4(\delta+\mu)^2}. \tag{25}$$

Comparing $\pi_0^{\mathrm{i}}(\lambda_{\mathrm{PC}}^{\star})$ and $\pi_0^{\mathrm{i}}(\lambda_{\mathrm{OD}}^{\star})$ from equations (23) and (25), respectively, we arrive at the following Corollary.

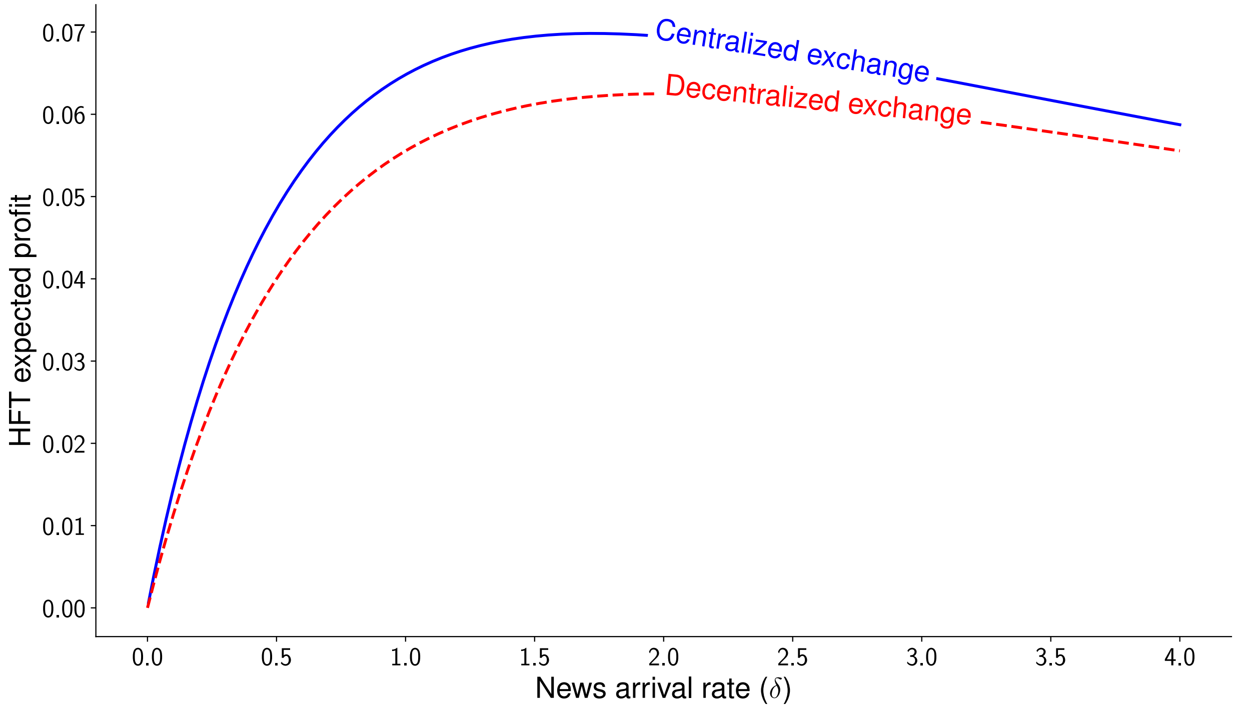
Corollary 4 (HFT Rents from the Speed Race)**.** *The expected rent earned from the speed race by an HFT is greater in the centralized market than in the decentralized market.*

Proof. See Appendix. □

On decentralized exchanges, HFTs earn lower rents from speed. Figure 6 presents this result graphically. At a first glance, the outcome is surprising since, from Proposition 3, HFTs employ fewer infrastructure resources. Moreover, equations (23) and (25) underscores that HFTs do not extract extra rents from liquidity traders, as the first terms of both equations are equal. These effects, however, are dominated by the "surge-pricing" feature of decentralized markets, as HFTs start acquiring speed following news arrival–during a "micro-burst"–where competition for speed intensifies beyond that achieved in a centralized market (Proposition 3). Therefore, the on-demand speed race of decentralized markets allows for a transfer of value from HFTs to suppliers of processing speed infrastructure.

Figure 6: **HFT Speed Race Rents**

This figure illustrates the equilibrium HFT profits in both centralized and decentralized exchanges, net of technology cost, as a function of the news arrival rate δ. Parameter values: $\mu = 2$, $\kappa = 0.25$, and $\sigma = 1$.



6 On-demand speed in practice: the distributed exchange

At exchanges operated as an electronic centralized limit order book, such as NYSE and Nasdaq, it is difficult to envision how one might dynamically price and deliver access to speed on-demand, where physical colocation to the matching engine is not priced on a continuous basis. In recent years, however, distributed ledger technology has given rise to a new form of electronic exchange that may allow for the dynamic pricing of speed–the so-called "decentralized exchange" (DEX).

By their nature, DEX provide a method by which the fill of orders is facilitated by market participants who will perform this service for a fee. Moreover, the heterogeneity in the speeds at which these facilitators can complete the order leads to a competition in fees offered to attract the fastest facilitators; this is particularly important for high-frequency arbitrageurs that seek to snipe stale quotes. In Section B, we provide some additional discussion on the technical aspects of the DEX architecture, whereas our discussion here focuses on the key mechanisms and intuition.

As of 2019, several decentralized exchanges (DEX) emerged in the realm of cryptosecurities, notably Binance DEX, IDEX, or Ethex. How does a decentralized exchange (DEX) work? We begin with a trader who wishes to execute an order. The trader must first broadcast the order message through the DEX. The trader may then (voluntarily) append a *gas* fee to the message to compensate peer-to-peer (P2P) platform operators for facilitating the order (see, e.g., Easley, O'Hara, and Basu, 2019), and finally signs the order with her private key. Cryptographic signatures prevent miners from tampering with, or front-running, the order.

Following the order broadcast, one or more of these facilitators (or "relayers" in the language of Warren and Bandeali, 2017) observe the pending order, pick it up, and rush to "write it" to the smart contract that encapsulates the matching engine. These contracts serve as the limit order book of the decentralized exchange. If the smart contract contains an order from another trader that yields a match, then the trade is completed; otherwise, her order "rests" on the smart contract, awaiting a counterparty. Once the order is written to the smart contract, the writer collects the gas fee. Because the act of writing an order to the smart contract is a first-past-the-post race, only the winner of this race collects the gas fee. Hence, a higher gas fee is likely to attract more miners and therefore reduce the order processing delay. This process reflects our model of speed on-demand, where the higher the investment in speed by one high-frequency sniper relative to another (i.e., higher gas fee), the greater likelihood that this sniper successfully completes their order. More specifically, the gas fee reflects our model's investment in speed variable, λ_i. In the context of our model, a trader may wish to increase the gas fee if they possess valuable fundamental information, as demand for the quickest facilitators at this time may be high, presumably from other traders possessing the same level of information. In times of no news, however, traders (such as our liquidity investor) would have the incentive to append only a nominal gas fee, coinciding with times of low demand for speed.

Our model argues that the "speed on-demand" mechanism employed by decentralized exchanges can improve, or at least not harm, traditional measures of market quality, while reducing resource usage and rents of speed races. We also find relatively quicker price discovery at decentralized exchanges at no cost to liquidity. That price discovery improves is subject to the strong caveat that cloud-based exchanges are slower at other stages of the trading process when compared to centralized markets (i.e., from the trader computer to the exchange front end). Therefore, we caution that our model does not insist that decentralized exchanges are the fastest trading mechanism from order submission to settlement, nor do they necessarily yield quicker price discovery than centralized exchanges *overall*. Instead, our intention is to showcase decentralized exchanges as an example of speed on-demand currently in practice. Moreover, because decentralized exchanges are relatively new, various stages of trade at DEX may see improvements in speed in the near future. [4]

[4] In February 2019, Binance (one of the largest cryptoasset exchange) launched a decentralized exchange allowing for one-second settlement times. See: Binance releases a first version of its decentralized crypto exchange.

7 Concluding Remarks

In this paper, we argue that distributed exchanges can mitigate the negative consequences of the high-frequency trading arms race. On decentralized exchanges, HFTs acquire speed in real time, on an as-needed basis, whereas centralized exchanges provision excess capacity to these accommodate "trading micro-bursts," following trading signals. By its nature, an on-demand speed environment (e.g., at distributed exchanges) eliminates the negative externality associated with maintaining idle capacity that occurs, for example, with co-location at centralized exchanges. We find that decentralized exchanges encourage short-lived, though intense, HFT races.

On decentralized exchanges, intense HFT competition for speed during microbursts triggers a surge in the price of computer power. As a result, HFTs earn lower rents from low-latency trading and, at the same time, the overall resource consumption is lower. Moreover, decentralized exchanges can improve, or at least not harm, traditional measures of market quality. We find relatively quicker price discovery at decentralized exchanges at no cost to liquidity.

Our result on price discovery is subject to the caveat that cloud-based exchanges exhibit slower stages of the trading process than centralized markets (i.e., from the trader computer to the exchange front end). Therefore, we caution that our model does not insist that decentralized exchanges are the fastest trading mechanism from order submission to settlement, nor do they necessarily yield quicker price discovery than centralized exchanges *overall*. We contend, however, that these stages may see improvements in speed in the near future. This would highlight decentralized exchanges as a viable solution toward reducing the social cost associated with the low-latency arms race, without harming market quality as measured, for example, by liquidity and price discovery.[5]

References

Aoyagi, Jun, 2019, Speed Choice by High-Frequency Traders with Speed Bumps, *Manuscript*.

Aune, Rune Tevasvold, Adam Krellenstein, Maureen O'Hara, and Ouziel Slama, 2017, Footprints on a Blockchain: Trading and Information Leakage in Distributed Ledgers, *The Journal of Trading* 12, 5–13.

Aurora Labs, 2017, IDEX : A Real-Time and High-Throughput Ethereum Smart Contract Exchange, pp. 1–11.

Azevedo, Eduardo M, and E Glen Weyl, 2016, Matching markets in the digital age, *Science* 352, 1056 LP – 1057.

[5] In February 2019, Binance (one of the largest cryptoasset exchange) launched a decentralized exchange allowing for one-second settlement times. See: Binance releases a first version of its decentralized crypto exchange.

Baldauf, Markus, and Joshua Mollner, 2018, Fast Trading on Fake News: The Role of Speed in Liquidity Provision, *Manuscript*.

———, 2019, High-frequency trading and market performance, *Manuscript*.

Baron, Matthew, Jonathan Brogaard, Björn Hagströmer, and Andrei Kirilenko, 2019, Risk and Return in High-Frequency Trading, *Journal of Financial and Quantitative Analysis* 54, 993–1024.

Basu, Soumya, David Easley, Maureen O'Hara, and Emin Sirer, 2019, Towards a Functional Fee Market for Cryptocurrencies, *Manuscript*.

Bhutoria, Mrinalini, 2018, Decentralized Exchanges (DEX), *Circle Research Whitepaper*.

Biais, Bruno, Christophe Bisière, Matthieu Bouvard, and Catherine Casamatta, 2019, The Blockchain Folk Theorem, *The Review of Financial Studies* 32, 1662–1715.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2015, Equilibrium Fast Trading, *Journal of Financial Economics* 116, 292–313.

Boulatov, Alex, and Thomas J. George, 2013, Hidden and displayed liquidity in securities markets with informed liquidity providers, *Review of Financial Studies* 26, 2096–2137.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2014, High frequency trading and price discovery, *Review of Financial Studies* 27, 2267–2306.

Brolley, Michael, and David Cimon, 2019, Order Flow Segmentation, Liquidity and Price Discovery: The Role of Latency Delays, *Manuscript*.

Budish, Eric, Peter Cramton, and John Shim, 2015, The high-frequency trading arms race: Frequent batch auctions as a market design response, *Quarterly Journal of Economics*.

Budish, Eric, Robin Lee, and John Shim, 2019, Will the Market Fix the Market? A Theory of Stock Exchange Competition and Innovation, *NBER Working Paper No. 25855*.

Chao, Yong, Chen Yao, and Mao Ye, 2017, Discrete Pricing and Market Fragmentation: A Tale of Two-Sided Markets, *American Economic Review* 107, 196–199.

Chiu, Jonathan, and Thorsten V Koeppl, 2019, Blockchain-Based Settlement for Asset Trading, *The Review of Financial Studies* 32, 1716–1753.

Cong, Lin William, and Zhiguo He, 2019, Blockchain Disruption and Smart Contracts, *The Review of Financial Studies* 32, 1754–1797.

———, and Jiasun Li, 2019, Decentralized Mining in Centralized Pools, *NBER Working Paper No. 25592*.

Cramer, Judd, and Alan B Krueger, 2016, Disruptive Change in the Taxi Business: The Case of Uber, *American Economic Review* 106, 177–182.

Daian, Philip, Steven Goldfeder, Tyler Kell, Yunqi Li, Xueyuan Zhao, Iddo Bentov, Lorenz Breidenbach, and Ari Juels, 2019, Flash Boys 2.0: Frontrunning, Transaction Reordering, and Consensus Instability in Decentralized Exchanges, *CoRR* abs/1904.0.

Easley, David, Maureen O'Hara, and Soumya Basu, 2019, From Mining to Markets: The Evolution of Bitcoin Transaction Fees, *Journal of Financial Economics*.

Foucault, Thierry, Johan Hombert, and Ioanid Rosu, 2016, News Trading and Speed, *Journal of Finance* 71, 335–382.

Foucault, Thierry, Roman Kozhan, and Wing Wah Tham, 2016, Toxic Arbitrage, *The Review of Financial Studies* 30, 1053–1094.

Foucault, Thierry, and Sophie Moinas, 2019, Is Trading Fast Dangerous?, in *Global Algorithmic Capital Markets* pp. 9–27.

Glosten, Lawrence R., and Paul R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71–100.

Joskow, Paul L, and Catherine D Wolfram, 2012, Dynamic Pricing of Electricity, *American Economic Review* 102, 381–385.

Khapko, Mariana, and Marius Zoican, 2019, How fast should trades settle?, *Management Science* Forthcom.

Kyle, Albert (Pete) S, and Jeongmin Lee, 2017, Toward a Fully Continuous Exchange, *Oxford Review of Economic Policy* 33, 650–675.

Malinova, Katya, and Andreas Park, 2017, Market Design for Trading with Blockchain Technology, *Working paper*.

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

Menkveld, Albert J, 2018, High-Frequency Trading as Viewed through an Electron Microscope, *Financial Analysts Journal* 74, 24–31.

Pagnotta, Emiliano S, and Thomas Philippon, 2018, Competing on Speed, *Econometrica* 86, 1067–1115.

Shkilko, Andriy, and Konstantin Sokolov, 2019, Every Cloud Has a Silver Lining: Fast Trading, Microwave Connectivity and Trading Costs, *Manuscript*.

Warren, Will, and Amir Bandeali, 2017, 0x: An open protocol for decentralized exchange on the Ethereum blockchain, *Whitepaper*.

Yesalavich, Donna Kardos, 2010, Microbursts Present Big Issues For High-Speed Traders, .

A Notation summary

Variable Subscripts

Subscript	Definition
M	pertaining to HFT market-maker role
B	pertaining to HFT “bandit” (i.e., sniper) role
PC	“pre-commitment,” or centralized market
OD	“on demand,”, or decentralized market

Exogenous Parameters

Parameters	Definition
v	asset value at $t = 0$, normalized to zero.
δ	Poisson arrival rate of news, i.e., common value innovations.
μ	Poisson arrival rate of liquidity traders (LI).
κ	elasticity of CPU power supply function.
σ	absolute size of common value innovations, if there is news.

Endogenous Quantities

Variable	Definition
ask	The price at which the market-maker is willing to sell one unit of the asset at $t = 1$.
λ_i	HFT speed, i.e., the Poisson intensity of the HFT market arrival process.
Λ	Total expenditure with market CPU infrastructure.
$p(\cdot)$	Market price of CPU infrastructure, $p = \kappa \sum_i \lambda_i$.
$\pi^0(\cdot)$	HFT profit from trading, net of speed cost.

B Architecture of a Decentralized Exchange

Decentralized exchanges (DEX) such as Binance DEX, IDEX, or Ethex are built on top of distributed application (dApp) platforms, such as Ethereum, and use tamper-proof and automatically-executed smart contracts to match trades (Bhutoria, 2018). Following the definition of Cong and He (2019), smart contracts are digital contracts enforced through network consensus. On a fully decentralized exchange such as Ethex, the computer code required to match orders and to maintain the state of the order book at any given time runs on a peer-to-peer (P2P) network of computers, rather than on a centralized exchange server.

A *blockchain*, which is a decentralized distributed ledger technology where “blocks” of transactions are cryptographically “chained” together, is a representative example of such a peer-to-peer network.[6] A distributed exchange can be implemented on a blockchain platform associated with a Turing-complete programming language, that is, a platform allowing for sophisticated state-contingent smart contracts.[7] The Ethereum blockchain (paired with the "Solidity" computer language) is an example of such a platform; the Bitcoin blockchain, on the other hand, is not. All real-world examples of decentralized exchanges referenced in this Section are implemented on the Ethereum blockchain.

How does a fully decentralized exchange (DEX) work? A trader first needs to broadcast her intention to buy or to sell a given quantity for a certain price. To this end, she accesses the DEX front-end through an application programming interface (API) or through a web-based platform. The trader voluntarily appends a *gas* fee to the message to compensate P2P platform operators for running the exchange (see, e.g., Easley, O'Hara, and Basu, 2019), and finally signs the order with her private key. Cryptographic signatures prevent miners from tampering with, or front-running, the order: In fact, Aune, Krellenstein, O'Hara, and Slama (2017) propose that the orders themselves be encrypted.

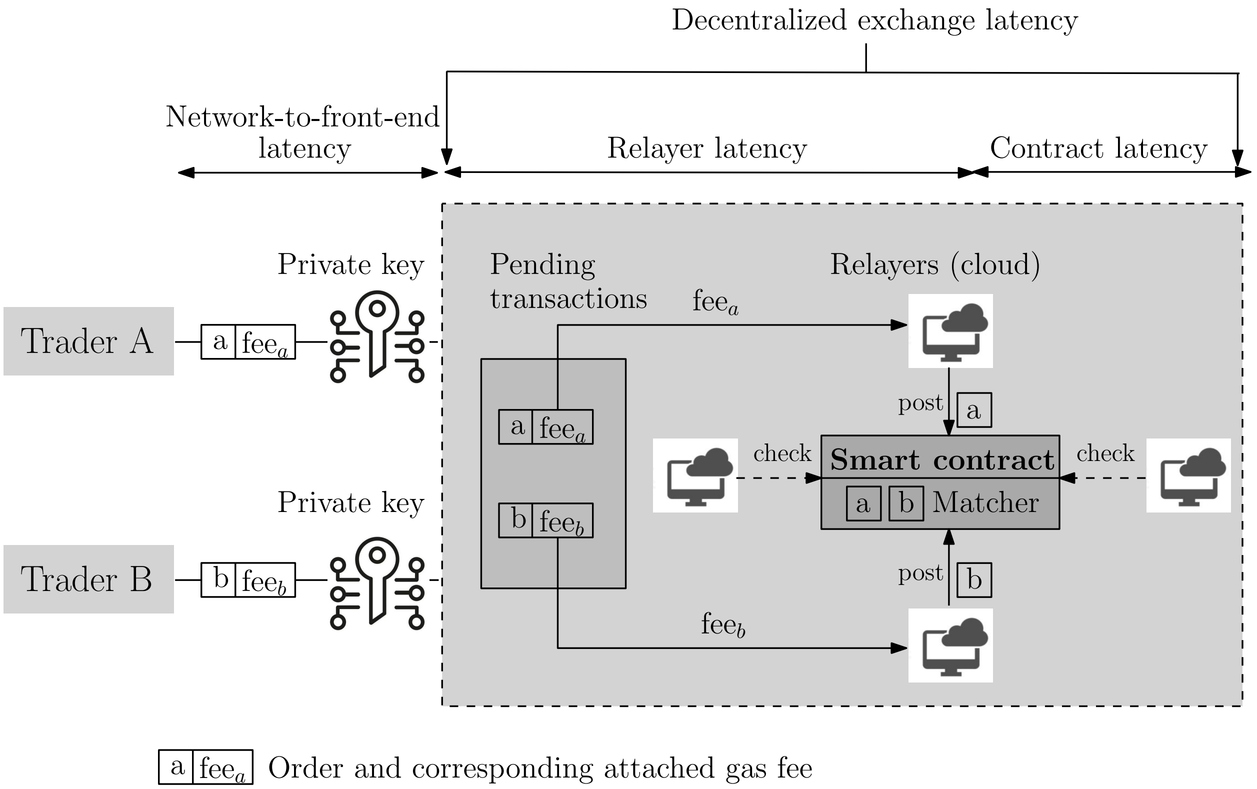
Following the order broadcast, one or more peer-to-peer computers (e.g., "relayers" in the language of Warren and Bandeali, 2017) observe the pending order, pick it up, and "write it" to the smart contract that encapsulates the matching engine. The first relayer to attach the order to the smart contract collects the gas fee. A higher gas fee is likely to attract more miners and therefore reduce the order processing delay. Finally, the distributed matching engine automatically performs one of two actions: either (partially) executes the new order against a resting order if the market is crossed, or adds the order to the limit book otherwise.

Figure 7 illustrates a fully decentralized exchange architecture, following Warren and Bandeali (2017) and Aurora Labs (2017). There are three components to the latency of an order: First, there is a trader-specific network latency, that is the required time to broadcast an order. Second, relayer latency corresponds to the delay between order broadcast and the moment when a relayer injects the order into the smart contract. Finally, contract latency measures the time needed for network consensus and the smart contract code execution. The illustration closely mirrors Figure 1 in Menkveld and Zoican (2017, p. 1194) for a centralized exchange. We follow Menkveld and Zoican (2017) and focus on latency due to exchange infrastructure as opposed to network latency to reach the exchange front end.

[6] We note, however, that P2P networks do not need to rely on a blockchain: BitTorrent, a file transfer software, is a peer-to-peer platform that does not use blockchain.

[7] A Turing-complete language can theoretically implement any algorithm or task achievable by a computer.

Figure 7: **Decentralized exchange architecture**



Two design features of real-world decentralized exchanges are noteworthy. First, crypto-asset trading platforms such as 0x, EtherDelta, or IDEX, use a hybrid "semi-centralized" structure featuring off-chain order relay with on-chain settlement. Limit orders are cryptographically signed and broadcast off of the blockchain with a empty counterparty field. Interested traders fill in their own crypto wallet address and inject the order into the smart contract on the blockchain. The hybrid process is faster since a trade settles on the blockchain only after counterparties are matched. Other exchanges, such as Ethex, are fully decentralized and store limit orders on the blockchain. Second, on-chain settlement is typical: in addition to the order matching smart contract, decentralized exchanges may use a second smart contract to modify traders' holdings of digital assets. In the context of cryptoassets, where exchange hacks are common, on-chain settlement augments security as it allows for direct transfers between traders' accounts (Aurora Labs, 2017). However, a decentralized exchange implemented on a regulated, transparent platform, does not necessarily need to integrate trading and settlement.

C Proofs

Proposition 3

Proof. To show that $\lambda^{\star}_{\text{OD}} > \lambda^{\star}_{\text{PC}}$, we compute the difference in speed intensity across the two market settings $\lambda^{\star}_{\text{OD}} - \lambda^{\star}_{\text{PC}}$,

$$\lambda^{\star}_{\text{OD}} - \lambda^{\star}_{\text{PC}} = \frac{\delta + \frac{3\mu\sigma}{2\kappa(\delta+\mu)} - \sqrt{\delta^2 + \frac{3\mu\sigma}{\kappa(\delta+\mu)}}}{6} \tag{C.1}$$

It is enough to show that the numerator is positive, which we obtain by showing the following,

$$\begin{aligned} \delta + \frac{3\mu\sigma}{2\kappa(\delta+\mu)} &> \sqrt{\delta^2 + \frac{3\mu\sigma}{\kappa(\delta+\mu)}}, \qquad \text{(C.2)} \\ \iff \left(\delta + \frac{3\mu\sigma}{2\kappa(\delta+\mu)}\right)^2 &> \left(\sqrt{\delta^2 + \frac{3\mu\sigma}{\kappa(\delta+\mu)}}\right)^2 \\ \iff \delta^2 + \frac{3\mu\sigma}{\kappa(\delta+\mu)} + \frac{9\mu\sigma}{4\kappa(\delta+\mu)} &> \delta^2 + \frac{3\mu\sigma}{\kappa(\delta+\mu)} \\ \iff \frac{9\mu\sigma}{4\kappa(\delta+\mu)} &> 0 \end{aligned}$$

which concludes the proof. □

Corollary 4

Proof. We compute the difference between the HFT expected profit in the two market settings from equations (23) and (25), that is

$$\pi^{\text{i}}_0(\lambda^{\star}_{\text{PC}}) - \pi^{\text{i}}_0(\lambda^{\star}_{\text{OD}}) = \frac{\delta}{36\,(\delta+\mu)^2}\left[2(\delta+\mu)\left(\delta\kappa - \sqrt{\delta\kappa\left(\delta\kappa + \frac{3\mu\sigma}{\delta+\mu}\right)}\right) + 3\mu\sigma\right]. \tag{C.3}$$

It is enough to show that

$$f\,(\sigma) \equiv 2(\delta+\mu)\left(\delta\kappa - \sqrt{\delta\kappa\left(\delta\kappa + \frac{3\mu\sigma}{\delta+\mu}\right)}\right) + 3\mu\sigma > 0. \tag{C.4}$$

We note that $f(0) = 0$ and further that f increases in σ since:

$$\frac{\partial f}{\partial \sigma} = 3\mu \left(1 - \underbrace{\frac{\delta\kappa}{\sqrt{\delta\kappa\left(\delta\kappa + \frac{3\mu\sigma}{\delta+\mu}\right)}}}_{<1} \right) > 0, \tag{C.5}$$

which concludes the proof. □