Financial Innovation

RESEARCH Open Access

Latency arbitrage and the synchronized placement of orders

Wolfgang Kuhle[1,2,3*]

*Correspondence: wkuhle@gmx.de

[1] Department of Economics, Corvinus University, Budapest, Hungary
[2] MEA, Max Planck Institute for Social Law and Social Policy, Munich, Germany
[3] Zhejiang University, Hangzhou, China

Abstract

We argue that owing to traders' inability to fully express their preferences over the execution times of their orders, contemporary stock market designs are prone to latency arbitrage. In turn, we propose a new order type, which allows traders to specify the time at which their orders are executed after reaching the exchange. Using recent latency data, we demonstrate that the order type proposed here allows traders to synchronize order executions across different exchanges, such that high-frequency traders, even if they operate at the speed of light, can no-longer engage in latency arbitrage.

Keywords: Market design, High-frequency trading, Latency arbitrage

JEL: D47

Introduction

The execution of large stock orders moves prices, which is costly for the investor. To reduce these costs, orders are usually broken-up into several smaller orders, which are then placed on different exchanges. Moreover, to ensure that other market participants cannot engage in arbitrage, these smaller orders are sent to exchanges, such that they are executed simultaneously. One obstacle to such simultaneous order executions are randomly varying latencies.[1]

In what follows, we describe the problem of latency arbitrage for the contemporary stock market design. In a second step, we propose a mechanism that helps investors to avoid the costs of latency arbitrage.

Our model assumes that investors buy and sell one homogenous asset on two geographically distinct exchanges. Trading is complicated by randomly varying latencies,

[1] That is, suppose an investor simultaneously sends two buy orders to two different exchanges. Due to random latencies, one of these orders, e.g. Order 1, reaches Exchange 1 earlier than Order 2 reaches Exchange 2. This scenario allows a high-frequency trader (HFT), who detects a price movement, i.e. the early execution of Order 1 on Exchange 1, to quickly buy on Exchange 2. In turn, the HFT can sell at a profit when the investor's delayed Order 2 reaches Exchange 2. High frequency traders operate dedicated glass-fiber networks for the purpose of latency arbitrage. Such networks allow for (one way) latencies of roughly 4 milliseconds (ms) between Chicago and New York (NY). At the same time, an investor, e.g. from Albany, faces a distribution of latencies: Albany-New York ($\mu = 51\,\text{ms}, \sigma = 28\,\text{ms}$) and Albany-Chicago ($\mu = 103\,\text{ms}, \sigma = 25.7\,\text{ms}$). That is, most orders sent from Albany to New York and Chicago do not arrive within 4 ms of one-another, and are thus subject to latency arbitrage. Using data from the New York Stock Exchange (NYSE) and the Chicago Mercantile Exchange (CME), Budish et al. (2015) show that such arbitrage opportunities are sizable: they are roughly worth 75 million USD annually for the trade in the Standard & Poor's 500 exchange-traded fund (SPY ETF) alone.

and by the presence of high-frequency traders (HFTs), who enjoy lower latencies than all other market participants. This model reveals that contemporary stock market designs, where traders can only choose when to send orders, are prone to latency arbitrage. This observation motivates an alternative market design, which allows traders not only to choose when to send orders *but also to specify the time at which their orders are executed after reaching the exchange*. Recent latency data indicate that this enables traders to synchronize order executions across different exchanges, such that HFTs can no-longer engage in latency arbitrage.

Related literature

Stiglitz (2014), Budish et al. (2015), and Aquilina et al. (2021) review[2] several proposals aimed at reducing latency arbitrage. One line of research recommends to Tobin-tax financial transactions, or to tax high-frequency trading, or to tax low-latency infrastructure. Other arguments, aimed at diluting the speed advantage of HFTs, involve reductions in the speed with which exchanges process orders, or limits to the speed with which market participants can place/cancel orders. Other models suggest that fast traders should compete in a "fast market," while slow traders participate in a "slow market." All these proposals have in common that they place additional restrictions on markets and market participants. The present study offers an alternative perspective: it demonstrates that latency arbitrage can be addressed by removing, rather than adding to, the restrictions that market participants face.

Perhaps closest to the present approach is Budish et al. (2015), who argue that latency arbitrage is "a symptom of a flawed market design." In turn, Budish et al. (2015, p. 1549), propose that exchanges should limit trade to discrete points in time, which makes it harder for HFTs to front-run other traders' orders. Unlike Budish et al. (2015), who propose a "discrete time trading" constraint, the present study argues that traders should be endowed with finer, rather than coarser, instruments, which make it easier for traders to cope with random latencies. Put differently, the new order type proposed here is a finer instrument in the sense that it gives traders additional choice variables, which help them to better express their preferences over the execution times of their trades. The market design proposed by Budish et al. (2015) is coarser in the sense that it constrains all market participants to trade at discrete, prescribed, points in time.

To better understand why our arguments differ from those in Budish et al. (2015), it is useful to note that Budish et al. (2015, p. 1552), view the presence of latency arbitrage opportunities, *as an exogenous, empirical, fact*. That is, they observe "obvious mechanical arbitrage opportunities, available to whoever is fastest. For instance, at 1:51:39.590 PM, after the price of the ES [Chicago] has just jumped roughly 2.5 index points, the arbitrage opportunity is to buy SPY [NYSE] and sell ES [Chicago]."[3] In the present paper, on the contrary, such arbitrage opportunities result endogenously whenever traders

[2] See also Kauffman et al. (2015) and Linton and Mahmoodzadeh (2017) for a broader review on high-frequency trading. See Roth and Xing (1994); Roth and Ockenfels (2002) for a broader market design perspective on the optimal frequency with which markets open and close.

[3] Put differently, Budish et al. (2015) observe that the price of the SPY in Chicago is not perfectly correlated with the price of the SPY on the NYSE. Put yet differently, Budish et al. (2015) show that the law of one price does not hold at very short time horizons. See also Epps (1979), who describes similar breakdowns in short-run correlations in older data sets.

place orders which, due to random latencies, are not executed simultaneously across exchanges. Creating such arbitrage opportunities is (1) costly for traders and (2) can be avoided if traders can specify the time at which orders are executed after reaching the exchange. Put differently, once traders use the order type proposed here, the arbitrage opportunities, upon which Budish et al. (2015) build their argument for slowing markets, are no-longer present.

The present paper shows that there exists a mechanism, which helps traders to avoid the costs of latency arbitrage. In turn, recent results on multi-criteria decision making (MCDM) in the context of financial innovations, as well as on consensus reaching, by Kou et al. (2014, 2021), Chao et al. (2022) and Li et al. (2022), may guide future quantitative work regarding a potential implementation of the present proposal.

Relevance and magnitude of the latency arbitrage problem

Budish et al. (2015) and Aquilina et al. (2021) emphasize the practical significance of latency arbitrage rents. Using 2015 data for the Financial Times Stock Exchange 100 Index (FTSE 100) stocks, Aquilina et al. (2021) reveal that latency arbitrage races account for 20% of all trading volume on the London Stock Exchange. More importantly, they identify that latency arbitrage represents a 0.5 basis points (BP) "tax" on trading. To put this number into perspective, Aquilina et al. (2021, p. 500), argue that empirically observed bid-ask spreads average 3 BP, such that latency arbitrage related costs add roughly 33% to the effective spread. Another way to view the importance of the latency arbitrage problem is to note that large EU- and US-based funds have trading costs of roughly 3 BP. Taking this view, latency arbitrage costs of 0.5 BP increase trading expenses from 3 BP to 3.5 BP, that is, by 17%. Hence, the proposed mechanism would reduce trading costs of large funds by 17%. Such a reduction in trading costs would arguably reduce market frictions and benefit investors. Regarding aggregates, Aquilina et al. (2021) estimate that latency arbitrage rents in global equity markets total roughly $5 billion annually.

Kauffman et al. (2015) focus more broadly on high-frequency trading practices, rather than just on latency arbitrage. For the years 2008–2014, Kauffman et al. (2015, p. 6), document that such trading accounted for 20–40% of EU equity trading and for 35–60% of US equity turnover. Kauffman et al. (2015) also highlight that regulatory constraints, such as stamp duties and bounds on the frequency with which stocks can be bought and sold, have limited the growth of HFT practices in many countries in the Asia-Pacific region. Against this background, the present mechanism would allow to relax regulation, for example, on the frequency with which stocks can be bought and sold, without the costs of increased latency arbitrage activity.

Hendershott et al. (2011) and others document that modern technologies have drastically reduced overall trading costs.[4] The problem of latency arbitrage may thus be viewed as a byproduct of a technological revolution that otherwise brought great benefits to markets. In turn, many researchers have argued that some of this technological

[4] See also Menkveld (2013), Shkilko and Sokolov (2020) for costs and benefits of HFT practices. Kauffman et al. (2015) and MacKenzie (2021) review the historical evolution of financial markets, including the recent HFT revolution. Osipovich (2021) discusses how HFTs have recently started to use fast satellite connections, rather than undersea cables, to exploit price movements on different continents.

advance should be sacrificed to address the problem of latency arbitrage. On the contrary, this study aims to present a mechanism, which is complementary to the growth of modern electronic markets.

Organization

The rest of the paper is organized as follows. We begin with a "Deterministic benchmark" model. In turn, we introduce "Random latencies", and show that contemporary market designs are prone to latency arbitrage, even if traders strategically delay the sending of orders. The section on "Synchronized order placement" proposes an order type, which helps traders to synchronize order executions across exchanges. Using recent latency data, sections "Calibration", "Technical feasibility" and "Costs and benefits of synchronized order placements" illustrate the effectiveness of the proposed market design. The section "Open questions, limitations, and directions for future research" discusses limitations of our study as well as directions for future research. Finally, concluding remarks are provided in the "Conclusion".

Deterministic benchmark

One asset is traded on two exchanges $m = L, S$. Each exchange has its own limit order book/excess demand function for the asset. The number/density of shares $f(P)$, which are on offer at each price P, differs across exchanges. To distinguish between the two exchanges, it is useful to assume that there is a large exchange L, which is more liquid than the smaller exchange S in the sense that $f_L(P) > f_S(P) \forall P$.[5] We also assume that, unless a large order is placed in a manner that brings prices into temporary disequilibrium, the market satisfies the law of one price $P_L = P_S = P_0$.[6]

A trader, who buys all shares offered for prices less or equal P^*_m on exchange m, receives a quantity X_m:

$$X_m := \int_{P_0}^{P^*_m} f_m(P)dP, \quad m = L, S. \tag{1}$$

The cost of buying X_m shares on exchange m is thus:

$$E_m := \int_{P_0}^{P^*_m} Pf_m(P)dP, \quad m = L, S. \tag{2}$$

If traders have access to both exchanges, they can minimize the cost of acquiring a given bundle of shares X by sending separate orders to both exchanges:

$$\min_{P^*_L, P^*_S} \left\{ \int_{P_0}^{P^*_L} Pf_L(P)dP + \int_{P_0}^{P^*_S} Pf_S(P)dP \right\} \quad s.t. \quad X_L + X_S = X, \tag{3}$$

where the first-order conditions to problem (3) imply that:

[5] The CME-Group (2016, p. 3), estimates that its market for the SPY future is 7 times more liquid than the NYSE's market for the SPY ETF. The CME-Group (2016, p. 3), also estimates that buying 100 Million worth of the S &P 500 costs 1.25 basis points (BP) on the CME while the cost is 2 BP if the same amount of the SPY ETF is bought on the NYSE.

[6] "Appendix A" presents such a marketplace, consisting of two local markets/exchanges, each of which with a distinct (excess) demand function/limit-order book.

$$P_L^* = P_S^* = P^*, \tag{4}$$

which yields:

Lemma 1 *Large traders split orders between both exchanges such that the law of one price is not violated.*

Random latencies

Suppose now that traders communicate their orders via a telecommunication network with random latencies. That is, orders to exchanges L and S may be delayed such that one order is executed earlier than the other. A high-frequency trader (HFT) can exploit this. Once the price on, for example, the small exchange increases to P_S^* the HFT knows from Eq. (4) that there is another order X_L^*, P_L^* on its way to exchange L. The HFT thus quickly buys the quantity X_L^* on the large exchange to sell at the higher price P_L^* once the delayed order arrives at exchange L.[7] This yields a rent for the HFT:

$$P_L^* X_L - \int_{P_0}^{P_L^*} P f_L(P) dP > 0, \tag{5}$$

and adds to the cost at which an investor acquires stocks on exchange L.[8]

The three scenarios summarized in Fig. 1 are associated with different costs for the investor. In the case where the large exchange L reveals the trade to the HFT (top left), let E_L denote the investor's total expenditures. These costs of early revelation on the large exchange are given in Eq. (7). Whenever the small exchange S reveals the trade, the investor's total expenditures are denoted by E_S. The costs of early revelation on exchange S are given in (7). The investor's expenditure is denoted by E_{sim}, whenever orders are executed "simultaneously."[9] The costs of simultaneous execution are given in (6).

Our assumption on market liquidity, namely $f_S(P) < f_L(P)$, allows us to rank these outcomes:

Lemma 2 $E_{sim} < E_L < E_S$.

Proof

$$E_{sim} = \int_{P_0}^{P^*} P f_L(P) dP + \int_{P_0}^{P^*} P f_S(P) dP < \int_{P_0}^{P^*} P f_L(P) dP + P^* X_S = E_L \tag{6}$$

[7] If an HFT's order arrives late, the HFT gets no fill and cancels the order. That is, the HFT acts as a pure arbitrageur in our model. The HFT can of course also arbitrage sell orders by first shorting the stock, just to buy it back at a lower price, once the trader's sell order arrives.

[8] Note that even if the trader knew that his order is front-run by the HFT with probability one, he would still buy from the HFT: executing the whole order on just one exchange would increase the (short-run) price on that exchange beyond the price P^*, which he is paying when he buys from the HFT. Note also that a HFT, who acts as a pure arbitrageur, will not carry inventory/buy more than what the investor is "willing" to buy from him.

[9] The term "simultaneously," refers to cases where orders arrive such that the HFT cannot engage in latency arbitrage.

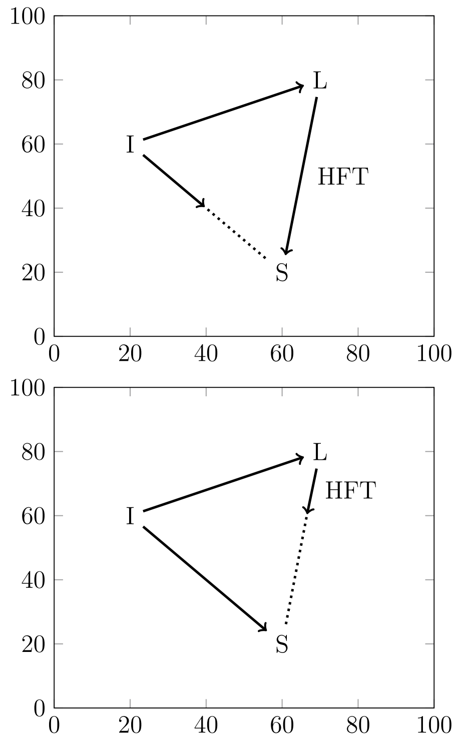

Fig. 1 Latency Arbitrage Triangle. An investor *I*, who sends orders to exchanges *L* and *S*, faces three different outcomes. Top left: the large exchange *L* reveals the trade, and the HFT front-runs the investor's order to the small exchange. Top right: the small exchange *S* reveals the trade, and the HFT front-runs the investor's order to the large exchange. Bottom left: orders are executed "simultaneously"

$$E_L = \int_{P_0}^{P^*} Pf_L(P)dP + P^*X_S < P^*X_L + \int_{P_0}^{P^*} Pf_S(P)dP = E_S \tag{7}$$

□

The ranking of expenditures in (6) and (7) indicates that investors have a preference for simultaneous order executions. Moreover, investors buy more shares on the large/liquid exchange, and thus prefer that trades are first revealed on the large exchange, rather than on the small exchange.

Optimal order delay

Let us now examine why contemporary market designs, where orders are executed as soon as they arrive at the exchange, which forces traders to tradeoff early execution of orders on one exchange against early executions on the other exchange. More precisely, this section demonstrates that it its rational for investors to delay orders such that trades are revealed more often on the larger exchange.[10] In turn, this observation motivates a

[10] As previously discussed, CME-Group (2016, p. 3), argue that the Chicago/CME market for the S &P 500 is significantly more liquid than that of the NYSE. Our model thus predicts that traders delay orders such that early executions are more frequent in Chicago than on the NYSE. This is indeed in-line with the empirical evidence that Budish et al. (2015, p. 1569), present: "[t]he majority (88.56 percent) of arbitrage opportunities in our data set are initiated by a price change in ES [Chicago], with the remaining 11.44 percent initiated by a price change in SPY [NYSE]." Moreover, they remark that this "is consistent with the practitioner perception that the ES [Chicago] market is the center for price discovery in the S &P 500 index."

simple solution to the problem of latency arbitrage, which is presented in the "Synchronized order placement" section.

Let $\delta \in \mathcal{R}$, denote the delay with which traders send orders to the small exchange.[11] Moreover, let H represent the time that a message sent by the HFT needs to travel from one exchange to the other. The investor can now choose the delay δ to minimize expected execution costs:

$$\min_{\delta} \left\{ \pi_{sim}(\delta, H)E_{sim} + \pi_L(\delta, H)E_L + \pi_S(\delta, H)E_S \right\}, \quad \pi_{sim} + \pi_L + \pi_S = 1. \tag{8}$$

Where π_{sim}, π_L, π_S are the respective probabilities with which orders are (i) executed simultaneously, (ii) first revealed on the large exchange, or (iii) first revealed on the small exchange. Regarding these probabilities, it is assumed that $\pi_{L,\delta} := \frac{\partial \pi_L}{\partial \delta} \geq 0$, $\pi_{S,\delta} := \frac{\partial \pi_S}{\partial \delta} \leq 0$.

Using this notation, the first-order condition to problem (8) can be written as:

$$\pi_{sim,\delta}(E_{sim} - E_L) = \pi_{S,\delta}(E_L - E_S), \tag{9}$$

respectively as

$$\pi_{L,\delta}(E_L - E_{sim}) + \pi_{S,\delta}(E_S - E_{sim}) = 0. \tag{10}$$

Using (9) and (10) yields

Lemma 3 *The optimal delay* δ^* *ensures that* $\pi_{sim,\delta} < 0$ *and* $|\pi_{L,\delta}| > |\pi_{S,\delta}|$.

Proof $\pi_{sim,\delta} < 0$ follows from (9) and Lemma 2. $|\pi_{L,\delta}| > |\pi_{S,\delta}|$ follows from (10) and Lemma 2. Finally, "Appendix B" contains an example where, assuming normally distributed latencies, first-order condition (9) can be solved explicitly for δ^*, which establishes that interior solutions to the optimal delay problem exist. □

According to Lemma 3, traders choose the delay δ such that $\pi_{sim,\delta} < 0$. That is, traders do not maximize the probability of simultaneous execution. Instead, they delay orders such that early executions on the large exchange are more frequent than those on the small exchange.

Figure 2 illustrates how a delay δ shifts early executions from the small exchange to the large exchange, which, as Lemma 3 shows, reduces the expected cost of latency arbitrage. The plot in Fig. 2 relies on the normally distributed example from "Appendix B", where latency times to the small exchange are given by $l_S = \mu_S + \delta + \sigma_S \xi, \xi \sim \mathcal{N}(0, 1)$ and those to the large exchange by $l_L = \mu_L + \sigma_L \varepsilon, \varepsilon \sim \mathcal{N}(0, 1)$.

Synchronized order placement

We now consider an alternative market design, where orders are accompanied by a time identifier, which specifies the exact time T at which the order is executed, respectively, added to the exchanges' order-book. That is, orders are sent out to the exchanges at time

[11] Using this notation, a delay $\delta = 10\,\text{ms}$ means that the order to the small exchange S is sent 10ms after the order to exchange L was sent. Likewise, a negative delay $\delta = -20\,\text{ms}$ means that the order to the small exchange is sent 20 ms earlier than the order to the large exchange.

$t = 0$, with an identifier T, indicating when orders are to be executed.[12] Thus, the market processes orders as follows:

1. Orders are sent to exchanges. In addition to price and quantity, these also specify the exact time of execution/addition to the limit order book.
2. Once orders arrive at the respective exchanges, they are not executed/placed until the specified placement time is reached. Exchanges are not allowed to publish the receipt of these orders until the placement time has been reached.
3. If an order arrives at the exchange after the desired placement time, it is placed immediately.

Lemma 4 *Under the market design described in* 1.–3. *traders can ensure simultaneous order executions.*

Proof Note that:

$$\begin{aligned}\pi_{sim} \geq & P(l_S \leq T+H)P(l_L \leq T+H) + P(|l_S - l_L| \leq H)(1 - P(l_S \leq T+H)P(l_L \leq T+H)) \\ \geq & P(l_S \leq T+H)P(l_L \leq T+H)\end{aligned} \tag{11}$$

$$\pi_{sim} = 1 - \pi_L - \pi_S \tag{12}$$

Equations (11) and (12), and the fact that $\pi_L, \pi_S \geq 0$ and $\lim_{T\to\infty} P(l_S \leq T+H)$ $P(l_L \leq T+H) = 1$, imply: $\lim_{T\to\infty} \pi_{sim} = 1$, $\lim_{T\to\infty} \pi_L = 0$, and $\lim_{T\to\infty} \pi_S = 0$. □

According to Lemma 4, traders can use the time identifier T, to ensure simultaneous order executions across exchanges. Put differently, increases in placement time T simultaneously reduce the probabilities π_L and π_S, and thus increase the probability of simultaneous executions π_{sim}.[13] This is an improvement over the contemporary market design detailed in the section on "Optimal order delay", where traders could only tradeoff early execution on one exchange against early execution on the other.

Figure 3 (left) graphs the probability of simultaneous executions, for a particular realization $\tilde{l}_S$, given that a trader specified execution time T. The smaller shaded area in Fig. 3 (right) indicates that the probability of simultaneous executions would be much lower under the contemporary market design, where traders cannot specify execution time T.

[12] Technically, instead of choosing the same execution time T for both orders, such that $T_S = T_L = T$, traders could of course choose different placement times $T_m, m = L, S$ for the two different exchanges. It is easy to see, however, that choosing $|T_L - T_S| > H$ reduces the probability of simultaneous executions, i.e. makes it easier for the HFT to engage in latency arbitrage.

[13] To illustrate this argument, consider, e.g. latency data from 27.05.2021 at 9:22 GMT. Let us start with a particularly bad connection: the *mean* latencies from Kampala to Manhattan (latencies to servers in New Jersey are similar) and Chicago were both roughly 440 ms. The *highest* observed latencies from Kampala to Manhattan and Chicago were 640 ms and 671 ms, respectively. Hence, a trader from Kampala could have ensured simultaneous executions by setting $T = 671$ ms. Within the US, Knoxville, Tennessee, did put up the highest latency numbers (the maximum observed latencies to Manhattan and Chicago were 70 ms and 80 ms respectively). Accordingly, Knoxville based traders could have ensured simultaneous executions by setting T = 0.08 s. All other US based traders, who enjoyed better connections, could have ensured simultaneous executions by choosing even lower values for T. Likewise, traders from London and Frankfurt could have chosen $T = 80$ ms and $T = 95$ ms, respectively to ensure simultaneous order executions in New York and Chicago. The maximum latency times discussed in this footnote are of course random variables, which are, ex-ante, unknown to investors. The "Calibration" section computes ex-ante probabilities for simultaneous executions.

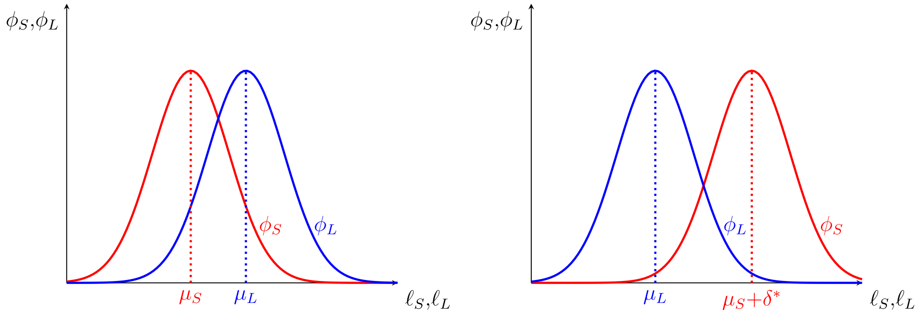

Fig. 2 Optimal Order Delay δ^*. (left) plots the density functions ϕ_S, ϕ_L for latencies l_S, l_L, to exchanges L, S, given that orders are sent without delay. In turn, (right) indicates how delaying orders to the small exchange shifts early executions to the large exchange

Calibration

To complement Fig. 3, it is useful to consider a numeric example, which demonstrates how the present proposal allows traders to synchronize order executions: An Albany (New York State) based investor, who trades the S &P 500 in New York and Chicago, faces the following model coefficients:

1. The HFT's one-way Chicago to New York time is roughly H=4ms.
2. Latencies[14]: Albany-NY $(\mu_S = 51ms, \sigma_S = 28ms)$, Albany-Chicago $(\mu_L = 103ms, \sigma_L = 25.7ms)$.
3. Relative cost of early executions $\frac{(E_S - E_{sim})}{(E_L - E_{sim})}$: CME-Group (2016) estimate the price impact of placing a 100 million Dollar SPY order in Chicago as 1,25 BP. Placing the same order on the NYSE has an estimated price impact of 2 BP. Using the linear model of "Appendix A" thus yields $\frac{E_S - E_{sim}}{E_L - Esim} = \frac{2}{1.25} = 1.6$.

Taken together, these coefficients allow us to compute execution probabilities π_S, π_L, π_{sim} for the three different model scenarios.

Contemporary market design without strategic order delay: HFTs can send messages from New York to Chicago in roughly $H = 4ms$, and an Albany based trader faces latencies Albany-NY $(\mu_S = 51ms, \sigma_S = 28ms)$, Albany-Chicago $(\mu_L = 103ms, \sigma_L = 25.7ms)$. Accordingly[15] 96% of all trades are subject to latency arbitrage under the contemporary market design, provided that the trader does not delay orders strategically. Put differently, 96% of all orders arrive at the two exchanges with a time gap of over 4 ms, which allows the HFT to engage in latency arbitrage. Only the remaining 4% of all orders arrive at the two exchanges within 4 ms of each other, and are thus executed "simultaneously." This low probability of simultaneous executions corresponds to the narrow, shaded band in Fig. 3 (right).

Contemporary market design with strategic order delay: The section "Optimal order delay" presented a model of optimal order delay. Given the distribution of latencies,

[14] Latency data is taken from wondernetwork (2021), 27.05.2021, 9:22 GMT.

[15] Calculations are given in "Appendix C".

and the different costs of early executions in both market places, this model predicts[16] that it is optimal to delay orders such that roughly 98% of all trades are first revealed in Chicago. Only 1% of executions are first revealed in New York.[17] Finally, only 1% of all executions are "simultaneous." Comparing this low percentage of simultaneous executions to the previous scenario without strategic order delay, where 4% of all orders were executed simultaneously, reveals that the incentive to skew early executions towards the large exchange is quite strong.

Synchronized order placement: In the foregoing two scenarios, orders were only executed simultaneously if they arrived within a narrow time window, as graphed in Fig. 3 (right), at the two exchanges. The size of this time window depended on the short timespan H, which the HFT needs to send a message from one exchange to the other. Given the level of noise in the latencies to New York and Chicago, simultaneous executions were thus very unlikely, and over 95% of trades were subject to latency arbitrage.

On the contrary, Fig. 3 (left) shows that the probability of simultaneous order executions increases drastically once traders can choose the time T, at which their orders are executed after they reach the respective exchange. Put differently, under the market design proposed in the section on "Synchronized order placement", simultaneous executions only require that both latency times are less than the execution time T. That is, via the choice of the execution time T, traders can directly choose the probability with which orders are executed simultaneously.

To quantify the increase in simultaneous executions, let us consider the case where traders set an execution time $T = 150\,\text{ms}$. Given that an Albany based trader faces latencies Albany-NY ($\mu_S = 51\,\text{ms}, \sigma_\text{S} = 28\,\text{ms}$), Albany-Chicago ($\mu_L = 103\,\text{ms}, \sigma_\text{L} = 25.7\,\text{ms}$), this ensures that roughly 97% of the time,[18] both orders arrive within 150ms at the exchanges, in which case trades are executed simultaneously at $T = 150\,\text{ms}$. A further increase to $T = 200$ brings the probability of simultaneous executions to around 99%. These probabilities are much higher than the 4%, respectively 1%, of simultaneous executions that obtained in the other two scenarios, in which traders had to rely on the contemporary market design.

Technical feasibility

To implement the current proposal, exchanges have to use sufficiently precise clocks. Recent markets in financial instruments directive (MIFID) II regulation, European-Commission (2016), requires that all market transactions within the EU, which are related to high-frequency trading, are recorded with a precision of at least 100 microseconds, that is, 0.1 ms. This provides an upper bound of 0.2 ms for the time span within which orders would be placed on two separate exchanges once the market design proposed in the

[16] Calculations are given in "Appendix C".

[17] As mentioned earlier, this finding is in line with the empirical observations in Budish et al. (2015, p. 1569), who note that "[t]he majority (88.56 percent) of arbitrage opportunities in our data set are initiated by a price change in ES [Chicago], with the remaining 11.44 percent initiated by a price change in SPY [NYSE]." That is, the present model of the contemporary stock market design, where traders can only delay orders, generates a distribution of arbitrage opportunities, which is in line with the empirical data that Budish et al. (2015) present.

[18] See "Appendix C" for the calculation. Alternatively, for the Albany-Chicago connection, note that an execution time $T = 150\,\text{ms}$ is roughly two standard deviations $\sigma_L = 25.7\,\text{ms}$ larger than the mean latency $\mu_L = 103\,\text{ms}$. That is, over 97% of orders sent from Albany to Chicago arrive before the placement time T is reached, and are thus executed at

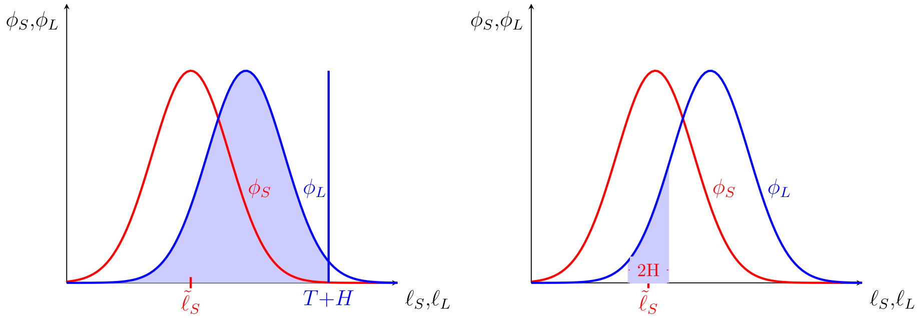

Fig. 3 Choosing an execution time *T* (left) allows traders to ensure a high probability of simultaneous order executions. The probability of simultaneous execution is low for contemporary market design (right), where both orders have to reach the exchange within a time frame of length 2 H

section on "Synchronized order placement" is adopted. This error is an order of magnitude smaller than the time frames that any HFT[19] could exploit.

Costs and benefits of synchronized order placements

The proposed order type allows traders to synchronize order executions across different exchanges. This synchronization comes at the cost of a delay in order executions, which depends on the distribution of latencies that traders face.

The following paragraphs compare costs and benefits of the present mechanism. Moreover, given that traders' preferences over these costs and benefits are not directly observable, we discuss indirect evidence on agent's preferences for synchronized order placements. Finally, the last paragraph stresses that the present order type allows traders to freely choose by how much they would like to delay orders to avoid the costs of latency arbitrage. On the contrary, earlier proposals in the literature, which rely on more drastic interventions, such as slowing the whole market to discrete time as in Budish et al. (2015), do not give traders such a choice.

Delay times Figure 3 and the numeric examples in the "Calibration" section illustrate that the delays, which, for example, ensure that over 97% of trades are executed simultaneously, depend on the distribution of latencies, and thus on the geographic location of the trader. Good connections, such as Frankfurt–New York, Frankfurt–Chicago, or London–New York and London–Chicago, require mean delays of 10–20ms.[20] Bad connections, such as Albany-NY and Albany-Chicago, examined in detail in the "Calibration"

Footnote 18 (continued)

time T. The time window for Albany-NY latencies ($\mu_S = 51$ ms, $\sigma_S = 28$ ms), is even larger at roughly $4\sigma_S$. That is, over 99% of orders arrive in NY before time T, and are thus executed at time T. Taken together, this implies that over 97% of orders are executed simultaneously at time T.

[19] Sending messages from Chicago to New York takes 3.66 ms, if it travels at light speed in a direct line. Time frames of 0.2 ms are still sufficient in the context of European markets, where it takes roughly 1 ms for light to travel from London to Frankfurt. Aquilina et al. (2021) note that the actual accuracy, with which the London Stock Exchange records the receipt of orders, is around 1 microsecond, which is much more accurate than the 100 microseconds that the European-Commission (2016) regulation requires.

[20] In order to establish upper bounds for the delay times, it is helpful to consider periods where latencies are more volatile than usual: For example on 29.01.2023, 11 am, latency times provided by wondernetwork (2021) are mean: 90ms minimum: 88ms maximum: 111 ms for the connection Frankfurt–New York, and mean: 98 ms minimum: 97 ms, maximum: 103 ms for Frankfurt–Chicago. Hence setting $T = 111$ms would have ensured simultaneous executions for a Frankfurt based trader at the cost of a 21 ms mean delay. Likewise on 28.01.2023, 6pm, latencies for London–New York are mean 75.025 ms minimum 72.746 ms maximum 83.812 ms. Latencies for London–Chicago are mean: 84.035 ms minimum: 83.689 ms maximum 87.911 ms. Hence, setting $T = 87$ ms would have ensured simultaneous executions at the cost of a 14 ms mean delay. These mean delay times of 21 ms and 14 ms would have ensured simultaneous execu-

section, require delays of 50 ms and 100 ms respectively to increase the percentage of simultaneous order executions from roughly 4% to over 97%.

To put the length of these delays into context, it is helpful to note that a blink of an eye takes roughly between 100 and 400 ms. Moreover, most humans can perform somewhere between 3 and 5 mouse clicks per second, such that a mouse click takes somewhere between 200 and 300 ms. Traders, who use keyboard and mouse, thus need several seconds to complete an order-form. Delays of 10 and 100 ms are thus imperceptible in the context of a traditional investor's order process. Moreover, traders who do employ trading strategies, which do not allow for delays, can choose an execution time $T = 0$ such that orders are executed as soon as they arrive at the exchange, just like under the contemporary market design.

Returns on the time of delay Using data from the London Stock Exchange, Aquilina et al. (2021) demonstrate that the average costs of latency arbitrage amount to roughly 0.5 BP per trade. Hence, if an investor can use a 50 ms delay to avoid latency arbitrage, for example, they save on average 0.5 BP on the purchase price of the stock. A return of 0.5 BP in 50 ms corresponds to a 10.5% return per second, which is orders of magnitude larger than the roughly 3 BP average daily return of the overall stock market.[21] Taking this view, the return on a 50 ms delay would be attractive, as long as the trader does not expect an adverse price change of more than 10% within the next second.[22]

Preferences over order delays Traders' preferences over delays are not directly observable. However, our model in the "Optimal order delay" section, and the numerical examples in the "Calibration" section, predict that traders in the SPY ETF should delay orders to the less liquid New York marketplace, such that trades are revealed more often on the liquid Chicago exchange. This theoretical prediction is in line with the empirical evidence presented by Budish et al. (2015, p. 1569), who find that 88.56% of all latency arbitrage opportunities are first revealed by price movements on the Chicago exchange; the remaining 11.44% were revealed by those on the NYSE. These data thus indicate that traders are indeed willing to delay orders to reduce the costs of latency arbitrage under the contemporary market design. Traders' willingness to delay orders to avoid only a fraction of the costs of latency arbitrage, indicates that a mechanism, which allows them to synchronize order placements such that latency arbitrage is no-longer possible, is likely useful.

The present mechanism thus helps slow traders to compete in markets where others' have access to fast low-latency networks. Moreover, given that traders differ regarding their size, trading strategies, or geographic location, their preferences over order delays are heterogeneous. The present mechanism is thus designed such that traders are free to

Footnote 20 (continued)

tions even for the outliers in the observed distribution of latency times. The delay times needed to ensure that only 97% of trades are executed simultaneously would have been even lower.

[21] That is, with 1000 ms in each second, the per second return would be $1.005^{20} - 1 \approx 0.105$. Assuming an average annual stock market return of 8%, which is spread over 252 trading days, the average daily return is roughly 3 BP. Returns of 10% *per second* are also sizable when compared to the three biggest historic one day drops in the S &P 500 index, which stand at 20.47% on 1987.10.19, 12.3% on 1929.10.28 and 12% on 2020.03.16.

[22] Similar arguments apply to aspects such as slippage, partial fills, and missed fills. That is, if an investor expects e.g. that slippage will not change by much in the next few milliseconds, he can afford a small delay. If an investor expects that slippage will increase drastically over the next few milliseconds, he may choose to not trade at all, since, irrespective of whether he uses the present mechanism or not, orders need a random amount of time to reach the exchange.

individually choose by how much they would like to delay the execution of their orders. Earlier proposals in the literature do not give traders such a choice, and require that all trade is taxed, or constrained to discrete time, for example.

Open questions, limitations, and directions for future research

Our highly stylized model, and our narrow focus on the problem of latency arbitrage, leave several questions for future research.

Potential drawbacks of the mechanism A mechanism that helps agents to synchronize order executions across exchanges could be used to manipulate markets. That is, the present order type could help a group of traders to sell/buy a large number of shares across different exchanges at one particular point in time. This type of coordination may amplify phenomena such as the 2010 flash crash, or the 2021 spikes in the Gamestop stock, or, more generally, short squeezes. That is, the present order type, which helps to avoid the costs of latency arbitrage, may potentially open the door for other forms of rent extraction.

In an alternative interpretation, the present order type may help regulators to monitor collusion in markets more effectively. That is, ex-post, collusion and market manipulations are arguably easier to detect when order data contain detailed requests regarding execution times. Moreover, exchanges could use the time window between the receipt of an order and the specified execution time of that same order to screen for trades that have the potential to destabilize the market.

Another aspect that may deserve closer examination concerns traders' incentive to cancel orders that have been sent to exchanges, but have not yet reached their execution time T. Put differently, one may ask whether traders should be allowed to use the time window, during which an order is held at the exchange, strategically? This question may be of particular interest in the context of heightened market volatility, where rapid price changes may provide incentives to cancel buy and sell orders. Such strategic behavior may either amplify or dampen volatility.[23]

Is it necessary to enforce the mechanism by law? The current mechanism allows traders to avoid the costs of latency arbitrage. However, this does not mean that exchanges have an incentive to offer this order type to traders. In particular, the current mechanism would likely reduce the revenues from selling high-frequency price data to HFT firms.[24] Accordingly, exchanges may choose to not adopt the present order type. One way to approach this question would be via a model where exchanges, which compete for trading volume, can choose whether to introduce the present mechanism. Budish et al. (2020) present one such model with the aim of examining whether exchanges have an incentive to adopt frequent batch auctions.

[23] A buyer, who interprets a price decline as a signal for further declines, may cancel his buy order, which would increase the price decline. A seller, who views a price decline as too steep, may cancel his sell order, which would dampen the price decline. Finally, the possibility to cancel orders may make traders, ex-ante, more willing to send orders during periods of elevated volatility, which might improve liquidity.

[24] Exchanges, which move first in allowing traders to specify the execution times of their orders, will likely have less early executions. At the same time, data on early executions is what HFTs pay for. Moreover, if all exchanges offered the order type proposed here, the sum of all latency arbitrage revenues, which would either accrue to HFTs or to the exchanges that sell data to HFTs, would either fall or disappear. One may thus conjecture that large exchanges, which have considerable market power, may be reluctant to offer the present order type. In turn, it would be necessary to enforce the mechanism proposed here by law.

One influential group, which may demand that exchanges introduce the present order type, are the firms, which pay for the listing of their stock on different exchanges. Given that Aquilina et al. (2021) find that latency arbitrage represents a 0.5 basis points (BP) "tax" on trading, reductions in latency arbitrage would increase the firms' market capitalization. That is, traders will pay less for a stock if they know that they will be taxed by the HFT when they buy. Moreover, traders anticipate (1) that they will be taxed again when they sell in the future, and that (2) the new buyer will also account for the increased cost of buying and selling in the presence of HFT activity. Taking this view, firms' present value/market capitalization would be higher if the stock was not taxed by HFTs. Firms may thus push exchanges to ensure that trade in their stock is not subject to latency arbitrage. Taken together, the above arguments suggest that exchanges will likely disagree with listed firms and traders regarding the implementation of the mechanism proposed here. Recent results on MCDM, and consensus reaching by Kou et al. (2014, 2021), Chao et al. (2022) and Li et al. (2022) may provide a basis to approach this aspect.

Further questions The introduction of the present market design may also impact the equilibrium behavior of market makers, equilibrium spreads, or the informational content of prices. For example, regarding the informational content of prices, the present mechanism would basically equalize prices across exchanges. This would contribute toward an equalization of the informational content of these prices.[25] This, in turn, would be a change from the status quo, where, as Budish et al. (2015, p. 1569), discuss, and as our theoretical model predicts for the contemporary market design, "the ES [Chicago] market is the center for price discovery," while the less liquid "New York market is lagging."

One implication of an equalization of the informational content of prices would be that traders may buy price data from only one, rather than from several, exchanges. This would, once again, likely reduce the data revenue of exchanges, and may incentivise exchanges to not introduce the present order type.

Conclusion

This study proposes a market design, which allows traders to specify the time at which their orders are executed after reaching the exchange. In turn, we demonstrate that this market design enables traders to synchronize order executions across different exchanges such that HFTs can no-longer engage in latency arbitrage. In an alternative interpretation, the proposed market design ensures that the law of one price holds, even in the very short-run.

Earlier proposals in the literature, aimed at reducing latency arbitrage, require taxes on financial transactions, or place restrictions on the speed with which trades are executed, orders placed, or prices quoted. The present study thus offers an alternative perspective, where a relaxation, rather than a tightening, of market constraints helps to avoid the costs of latency arbitrage.

[25] The models of Green (1975), Grossman and Stiglitz (1980), Hellwig (1980) or Admati (1985), which focus on the informational content of price systems, could be used to study this conjecture. The model of Admati (1985) allows for multiple assets with correlated fundamental values, which would open the door towards a model where HFTs exploit price movements of assets, like Ford and General Motors, which are similar, rather than identical.

Appendix A: Linear model

Equations (13) and (14) represent linear asset demands on the two exchanges S and L[26]:

$$X_S = \frac{a}{b} - \frac{1}{b}P_S \quad \Leftrightarrow \quad P_S = a - bX_S, \quad a, b > 0 \tag{13}$$

$$X_L = \frac{c}{d} - \frac{1}{d}P_L \quad \Leftrightarrow \quad P_L = c - dX_L, \quad c, d > 0 \tag{14}$$

Summing over the demands X_S and X_L yields aggregate demand X^D, which meets a fixed aggregate supply $\bar{X}$:

$$X_S + X_L = X^D, \quad X^D = \bar{X}. \tag{15}$$

In a long-run arbitrage free equilibrium, prices satisfy the law of one price:

$$P_S = P_L = P_0, \tag{16}$$

where P_0 is the equilibrium price. Solving (13)–(16) yields equilibrium quantities:

$$\begin{aligned} X_S &= \frac{a + d\bar{X} - c}{b + d} \quad X_L = \frac{-a + b\bar{X} + c}{b + d} \\ P_S &= a - b\frac{a + d\bar{X} - c}{b + d} \\ P_L &= c - d\frac{-a + b\bar{X} + c}{b + d}. \end{aligned}$$

If a trader buys a large number of shares $\tilde{X}$, the new (long-run) prices and quantities are

$$\begin{aligned} X_S &= \frac{a + d(\bar{X} - \tilde{X}) - c}{b + d} \quad X_L = \frac{-a + b(\bar{X} - \tilde{X}) + c}{b + d} \\ P_S &= a - b\frac{a + d(\bar{X} - \tilde{X}) - c}{b + d} \\ P_L &= c - d\frac{-a + b(\bar{X} - \tilde{X}) + c}{b + d}. \end{aligned}$$

Strategy of the HFT and the Limit Order Book Taking derivatives in (13) and (14) yields:

$$\frac{dX_S}{dP_S} = -\frac{1}{b}, \quad \frac{dX_L}{dP_L} = -\frac{1}{d},$$

such that

$$\tilde{X} = -\int_{P_0}^{P^*} \frac{dX_S}{dP_S} dP_S - \int_{P_0}^{P^*} \frac{dX_L}{dP_L} dP_L = \left(P^* - P_0\right)\left(\frac{1}{b} + \frac{1}{d}\right).$$

Expenditures in the case of simultaneous execution are now:

[26] Demand for the stock satisfies the "law of demand," in the sense that it is downward sloping in price. Under the alternative assumption, where demand is upward sloping in price, equilibrium would not necessarily be stable.

$$\tilde{E}_{sim} = -\int_{P_0}^{P^*} P_S \frac{dX_S}{dP_S} dP_S - \int_{P_0}^{P^*} P_L \frac{dX_L}{dP_L} dP_L = \frac{1}{2}\left(P^{*2} - P_0^2\right)\left(\frac{1}{b} + \frac{1}{d}\right)$$

Expenditures in the case where the trade is revealed on exchange L are:

$$\tilde{E}_L = -P^* \int_{P_0}^{P^*} \frac{dX_S}{dP_S} dP_S - \int_{P_0}^{P^*} P_L \frac{dX_L}{dP_L} dP_L = P^{*2}\frac{1}{b} - P^* P_0 \frac{1}{b} + \frac{1}{2}\left(P^{*2} - P_0^2\right)\frac{1}{d}$$

Expenditures in the case where the trade is revealed on exchange L are:

$$\tilde{E}_S = -\int_{P_0}^{P^*} \frac{dX_S}{dP_S} dP_S - P^* \int_{P_0}^{P^*} \frac{dX_L}{dP_L} dP_L = P^{*2}\frac{1}{d} - P^* P_0 \frac{1}{d} + \frac{1}{2}\left(P^{*2} - P_0^2\right)\frac{1}{b}.$$

It follows that

$$\begin{aligned} E_L - E_{sim} &= \frac{1}{2b}\left(P^* - P_0\right)^2 > 0 \\ E_S - E_{sim} &= \frac{1}{2d}\left(P^* - P_0\right)^2 > 0 \\ \frac{E_S - E_{sim}}{E_L - E_{sim}} &= \frac{b}{d} \end{aligned}$$

CME-Group (2016, p. 3), estimate that a purchase worth 100 Million Dollar in the SPY increases prices by $\frac{\Delta P_L}{P_0} = 1.25$ BP on the more liquid Chicago exchange and by $\frac{\Delta P_S}{P_0} = 2$ BP on the less liquid NYSE. Inserting these numbers into demand functions (13) and (14), yields $d = \frac{\Delta P_L}{\Delta X_L} = 1.25\frac{P_0}{\Delta X_L}$ and $b = \frac{\Delta P_S}{\Delta X_S} = 2\frac{P_0}{\Delta X_S}$. Recalling (17), and noting that $\Delta X_S = \Delta X_L$ if 100 million Dollar worth of stock is bought on both exchanges, the relative excess cost of early execution is $\frac{E_S - E_{sim}}{E_L - E_{sim}} = \frac{b}{d} \approx \frac{2}{1.25} = 1.6$.

Appendix B: Optimal delay with normally distributed latency

Latencies to the small and the large exchange are denoted by l_S and l_L, respectively. These latencies are assumed to be normally distributed, such that

$$\begin{aligned} l_S &= \mu_S + \delta + \sigma_S \xi \quad \xi \sim \mathcal{N}(0,1) \\ l_L &= \mu_L + \sigma_L \varepsilon \quad \varepsilon \sim \mathcal{N}(0,1) \end{aligned} \tag{17}$$

The HFT's latency time is $H > 0$. It is useful to define the difference in time, with which messages arrive at the exchanges, as $x := l_S - l_L$. Moreover, note that $\gamma := E[l_S - l_L] = \mu_S - \mu_L + \delta$. In addition, latencies are assumed to be uncorrelated such that $Var(l_S - l_L) = \sigma_S^2 + \sigma_L^2$, and let $\alpha := \frac{1}{Var(l_S - l_L)} = \frac{1}{\sigma_S^2 + \sigma_L^2}$. Finally, the function $\Phi()$ stands for the cumulative standard normal distribution function and $\phi()$ denotes its density. The probabilities of early revelation on exchanges L, S as well as the probability of simultaneous execution are then:

$$\pi_L = P(l_S - l_L > H) = 1 - \Phi(\sqrt{\alpha}(-\gamma + H)) \tag{18}$$

$$\pi_S = P(l_S - l_L < -H) = \Phi(\sqrt{\alpha}(-\gamma - H)) \tag{19}$$

$$\pi_{sim} = P(|l_S - l_L| < H) = 1 - \pi_L - \pi_S \tag{20}$$

Given (18) and (19), the first order condition for optimal delay (10) rewrites:

$$\phi\big(\sqrt{\alpha}\big(-\gamma + H\big)\big)\big(E_L - E_{sim}\big) = \phi\big(\sqrt{\alpha}\big(-\gamma - H\big)\big)\big(E_S - E_{sim}\big). \tag{21}$$

Recalling $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$, (21) yields:

Lemma 5 *Orders to the small exchange are delayed such that* $\delta^* = \mu_L - \mu_S + \frac{\sigma_S^2+\sigma_L^2}{2H} ln(\frac{E_S-E_{sim}}{E_L-E_{sim}})$, *and* $\gamma^*(\delta^*) = \frac{\sigma_S^2+\sigma_L^2}{2H} ln\left(\frac{E_S-E_{sim}}{E_L-E_{sim}}\right) > 0$, *and* $\pi_L(\delta^*) > \pi_S(\delta^*)$.

Appendix C: Early execution probabilities

The execution probabilities π_S, π_L, π_{sim} depend on the HFT's one-way Chicago to New York latency, which is roughly H = 4 ms. Moreover, latencies are Albany-NY ($\mu_S = 51\,\text{ms}, \sigma_S = 28\,\text{ms}$), and Albany-Chicago ($\mu_L = 103\,\text{ms}, \sigma_L = 25.7\,\text{ms}$). Finally, CME-Group (2016) estimate the price impact of placing a 100 million Dollar SPY order as 1,25 BP. Placing the same order on the NYSE has an estimated price impact of 2 BP. Using the linear model of "Appendix A", yields $\frac{E_S-E_{sim}}{E_L-Esim} = \frac{2}{1.25} = 1.6$.

Orders without strategic delay Setting the delay time $\delta = 0$ in "Appendix B" is the easiest way to compute execution probabilities for the model without delay. To do so, note that $\gamma = \mu_S - \mu_L + \delta = -52$, and $\alpha = \frac{1}{Var(l_S - l_L)} = \frac{1}{\sigma_S^2+\sigma_L^2}$, such that $\sqrt{\alpha} = \sqrt{\frac{1}{28^2+25.7^2}} \approx \frac{1}{38}$. Finally, substituting into (18)–(20), yields $\pi_S \approx \Phi(\frac{48}{38}) \approx 0.89, \pi_L = 1 - \Phi(-\frac{56}{38}) \approx 0.07$ and $\pi_{sim} \approx 0.04$.

Orders with strategic delay Recalling "Appendix B", the optimal delay is $\delta^* = \mu_L - \mu_S + \frac{\sigma_S^2+\sigma_L^2}{2H} ln(\frac{E_S-E_{sim}}{E_L-E_{sim}})$, and $\gamma^*(\delta^*) = \frac{\sigma_S^2+\sigma_L^2}{2H} ln(\frac{E_S-E_{sim}}{E_L-E_{sim}}) \approx \frac{1444}{8} ln(1.6) \approx 84.8$. Moreover, $\alpha = \frac{1}{Var(l_S-l_L)} = \frac{1}{\sigma_S^2+\sigma_L^2}$, such that $\sqrt{\alpha} = \sqrt{\frac{1}{28^2+25.7^2}} \approx \frac{1}{38}$. Substitution into (18)–(20) yields $\pi_S \approx \Phi(\frac{-88.8}{38}) \approx 0.01, \pi_L \approx 1 - \Phi(\frac{-80.8}{38}) \approx 0.98$ and thus $\pi_{sim} \approx 0.01$.

Orders with specified placement time Recall that $\pi_{sim} \geq P(l_S \leq T + H)P(l_L \leq T + H) + P(|l_S - l_L| \leq H)(1 - P(l_S \leq T + H)P(l_L \leq T + H)) \geq \Phi(\frac{103}{28})\Phi(\frac{51}{25.7}) \approx 0.97$. Hence choosing $T = 150\,\text{ms}$, yields $\pi_{sim} \geq 0.97$, i.e. ensures that 97% of orders are executed simultaneously.

Abbreviations

BP	Basis points
CME	Chicago Mercantile Exchange
FTSE 100	Financial Times Stock Exchange 100 Index
HFT	High-frequency trader
HFTs	High-frequency traders
MCDM	Multi-criteria decision making
MIFID	Markets in financial instruments directive
ms	Milliseconds
NY	New York
NYSE	New York Stock Exchange
SPY ETF	Standard & Poor's 500 exchange-traded fund

Acknowledgements
I thank Eric Budish and Al Roth for bringing the topics of market design and high-frequency trading to my attention. I also thank five anonymous referees for comments that helped me to improve this paper substantially. Large parts of this manuscript were written during my time at Zhejiang University, and I thank my colleagues and students for their friendship during this time. First draft August 2019. First arxiv version 26.12.2021.

Author contributions
Prof. Dr. Wolfgang Kuhle is the sole author of this paper. The author read and approved the final manuscript.

Funding
None.

Availability of data and materials
All data used are publicly available.

Declarations

Competing interests
The author declares no competing interests.

Received: 4 April 2022 Accepted: 13 April 2023
Published online: 20 June 2023

References

Admati A (1985) A noisy rational expectations equilibrium for multi-asset securities markets. Econometrica 53:629–657
Aquilina M, Budish E, ONeill P (2021) Quantifying the high-frequency trading arms race. Q J Econ 137(1):493–564
Budish E, Crampton P, Shim J (2015) The high-frequency trading arms race: frequent batch auctions as a market design response. Quart J Econ 130(4):1547–1621
Chao X, Dong Y, Kou G, Peng Y (2022) How to determine the consensus threshold in group decision making: a method based on efficiency benchmark using benefit and cost insight. Ann Oper Res 316:143–177
Epps T (1979) Comovements in stock prices in the very short run. J Am Stat Assoc 74(366):291–298
Grossman S, Stiglitz J (1980) On the impossibility of informationally efficient markets. Am Econ Rev 70(3):393–408
Hellwig M (1980) On the aggregation of information in competitive markets. J Econ Theory 22:477–498
Hendershott T, Jones C, Menkveld A (2011) Does algorithmic trading improve liquidity. J Finance 66(1):1–33
Kauffman RJ, Hu Y, Ma D (2015) Will high-frequency trading practices transform the financial markets in the Asia Pacific region? Financ Innov 1(4):1–27
Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Inf Sci 275:1–12
Kou G, Akdeniz OO, Dincer H, Yuksel S (2021) Fintech investments in European banks: a hybrid IT2 fuzzy multidimensional decision-making approach. Financ Innov 7(39):1–28
Li Y, Kou G, Li G, Peng Y (2022) Consensus reaching process in large-scale group decision making based on bounded confidence and social network. Eur J Oper Res 303:790–802
Linton O, Mahmoodzadeh S (2017) Implications of high-frequency trading for security markets. Annu Rev Econ 10:237–259
MacKenzie D (2021) Trading at the speed of light: how ultrafast algorithms are transforming financial markets. Princeton University Press, Princeton
Menkveld A (2013) High frequency trading and the new market makers. J Financ Mark 16(4):712–740
Roth AE, Ockenfels A (2002) Last-minute bidding and the rules for ending second-price auctions: evidence from ebay and amazon auctions on the internet. Am Econ Rev 92:1093–1103
Roth AE, Xing X (1994) Jumping the gun: imperfections and institutions related to the timing of market transactions. Am Econ Rev 84:992–1044
Shkilko A, Sokolov K (2020) Every cloud has a silver lining: fast trading, microwave connectivity, and trading costs. J Finance 75(6):2899–2927
Budish E, Lee R, Shim J (2020) A theory of stock exchange competition and innovation: will the market fix the market? In: NBER working paper, p 25855
CME-Group (2016) The big picture: a cost comparison of futures and ETFs. https://www.cmegroup.com/trading/equity-index/files/a-cost-comparison-of-futures-and-etfs.pdf, 2. Edition:1–16
European-Commission (2016) Supplementing directive 2014/65/eu of the European parliament and of the council with regard to regulatory technical standards for the level of accuracy of business clocks. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0574
Green JR (1975) Information efficiency and equilibrium. In: HIER working paper (284), pp 1–29
Osipovich A (2021) High-frequency traders eye satellites for ultimate speed boost. Wall Street J
Stiglitz JE (2014) Tapping the brakes: are less active markets safer and better for the economy. Working paper, pp 1–19
wondernetwork (2021) Global ping statistics. https://wondernetwork.com/pings/

Publisher's Note