Who needs a faster exchange? Evidence from agricultural futures electronic trading

Abstract

Like high-frequency traders, exchanges have invested heavily in reducing order-processing latency. Yet little is known about whether lower exchange latency improves market quality or how its benefits and costs are distributed across market participants. Using raw message data from the Chicago Mercantile Exchange (CME), we examine the causal effects of exchange latency on market quality and trading outcomes for liquidity providers and liquidity takers. To address endogeneity, we exploit cross-commodity processing congestion within the exchange as an instrument, whereby message traffic from other agricultural commodities delays the processing of a given order but is otherwise unrelated to its market quality. We find that lower exchange latency primarily benefits liquidity providers by increasing market-making revenues while reducing adverse selection costs, whereas higher exchange latency lowers the immediate trading costs borne by liquidity takers. Because all agricultural futures markets currently share a common order processing system, our findings shed light on a congestion-based externality and suggest that exchanges should account for these trade-offs in market design.

Keywords: Exchange latency; Trading speed; Market takers; Market makers; Market liquidity; Electronic trading
JEL Classification Numbers: G12, G13, G14, Q02, Q13

1 Introduction

Commodity trading has transitioned to electronic platforms over the past two decades, reducing trading latency to milliseconds. While many traders have invested heavily in upgrading their technology, exchanges, which act as intermediaries between market participants, have likewise modernized their trading infrastructure to enable a faster and more efficient order processing system (e.g., Menkveld and Zoican 2017; Pagnotta and Philippon 2018). The speed competition, either among traders or among exchanges, has raised concerns about whether fast trading impairs market liquidity and overall welfare (e.g., Hasbrouck and Saar 2013; Budish, Cramton, and Shim 2015; Menkveld 2016). This paper examines whether order processing time, i.e., exchange latency, affects market quality and trading outcomes between liquidity providers and liquidity takers. All agricultural futures markets currently share a common order processing system within the exchange, implying that message traffic generated in one commodity can delay order processing in others. Possible congestion costs are not borne solely by the traders generating the message flow but may also be imposed on other participants, suggesting a congestion-based externality. We reveal that exchange-level speed, which is independent to traders' own speed choice, matters for market quality.

Electronic traders can act as liquidity providers that continuously engage in market-making through intensive quote management (limit order) (Aït-Sahalia and Sağlam 2024), or liquidity takers that actively remove quoted liquidity through trades (market orders or crossing limit orders) before those orders are updated or canceled (Biais, Foucault, and Moinas 2015).[1] Although traders race to be the first to access exchanges, their individual (trader-specific) speeds are not relevant once the order enters the exchanges' order processing queue (Menkveld and Zoican 2017). Submitted orders need to wait in a queue before being processed by the exchange matching engine. Lower exchange latency can reduce the probability that incoming orders have to queue behind other orders in the processing queue. Furthermore, exchange latency affects both groups of traders. Liquidity providers may benefit from faster exchanges through timely quote updates and avoid being picked off by liquidity takers, reducing adverse selection risk and possibly increasing market-making revenue. However, faster exchanges may also allow liquidity takers to more readily take advantage of stale quotes, which increases the adverse selection risk born by liquidity providers. Menkveld and Zoican (2017) refer to the aforementioned two channels respectively as "payoff-risk" and "dueling" channel, and market quality depends on their net effect.

[1] Because our data do not allow us to identify whether individual traders are high-frequency traders, we do not distinguish between high-frequency and other traders. Instead, we classify all traders according to their liquidity roles as either liquidity providers or liquidity takers, which also facilitates the interpretation of our findings.

We use packet capture (PCAP) data, a raw exchange message dataset covering all agricultural futures traded on the CME in 2024, to assess how exchange latency affects market quality and how costs and benefits of a faster exchange are distributed across different trader groups. The PCAP data records every message (order) communicated between traders and exchanges in all agricultural futures markets.[2] Furthermore, it records timestamps at which messages are received by exchanges as well as those at which messages are sent to the market data feed, allowing us to directly measure the exchange latency. More importantly, it allows us to accurately calculate the queue length within the order processing system to characterize the congestion of the matching engine. We focus on six major agricultural commodities with different market liquidity characteristics: corn, soybeans, Chicago wheat, Kansas wheat, soybean meal, and soybean oil. We calculate the exchange latency for both quote messages at the top-of-the-book and trade messages to explore how messages from liquidity providers and takers affect market quality, respectively.

We define exchange latency as the time elapsed between an order's entry into the exchange and its dissemination through the market data feed. Our results suggest that the hourly median exchange latency of quote messages is about half of a millisecond, while that of trade messages is slightly higher to about 1 millisecond. This pattern is consistent with trade messages requiring additional time to complete matching with resting limit orders, whereas quote messages are inserted into the limit order book without matching.

Following Hendershott, Jones, and Menkveld (2011), we assess market quality using three variables: For liquidity takers, we consider effective spread, which measures the immediate trading costs; For liquidity providers, we calculate realized spread, which captures the gross market-making revenue, and price impact, which measures the adverse selection costs imposed by informed liquidity takers. We aggregate them by calculating hourly volume-weighted averages. Consistent with previous studies (e.g., Hu, Serra, and Garcia 2020), we find the price impact (3.90 bps) account for about 95% of the effective spread (4.09 bps), suggesting that liquidity providers bear a substantial proportion of adverse selection risk.

Exchange latency is possibly endogenous because, for example, message traffic may also condition on market quality, where a reverse causality may occur. To facilitate causal interpretation, we use the number of messages from other commodities positioned ahead of a focal message in the queue upon its arrival at the exchange, named *queue-ahead*, as an instrumental variable for the endogenous exchange latency analysis.

Our identification strategy relies on the CME's unique institutional features. Messages from all agricultural commodities need to be processed by a uniform system, including the same market segment gateway and matching engine. In turn, given a message, a longer queue

[2]We use "order" and "message" interchangeably hereafter.

will delay its processing and increase the exchange latency, affecting market quality. Traders cannot control messages queued ahead of their own. This mechanism can be captured by our *queue-ahead* variable, which characterizes message traffic already in the processing system which has not yet completed processing, allowing us to study the effect of the order queue on market latency and market quality. Furthermore, we follow a "leave-one-out" logic, constructing the *queue-ahead* measure using only messages from commodities other than the commodity associated with the focal message. This design reduces the likelihood that the instrument directly affects market quality in the focal commodity. Our exclusion restriction assumes that messages from other commodities positioned ahead of a focal message affect the focal commodity's market quality only by delaying the processing of that message, after controlling for relevant covariates. Our placebo test suggests that unobservable factors correlated with market quality, such as information flows, are not significantly associated with our instrument.[3] This supports the validity of the exclusion restriction and reduces concerns that our instrumental variable captures preexisting market conditions.

We estimate the effect of exchange latency on market quality using a two-stage least squares instrumental variable (2SLS-IV) regression with an hourly sample. Our results suggest that exchange latency of both trade and quote messages has distinct influence on liquidity providers and takers. We find a one-standard-deviation (0.18 milliseconds) decrease in quote latency (i.e., faster exchange) significantly increases the realized spread by 0.12–0.16 bps, representing 59.19%–83.67% of its sample mean. Similarly, the same magnitude of decrease in quote latency significantly reduces the price impact by 0.12–0.15 bps, but representing merely about 3% of its sample mean. Nevertheless, the effects on realized spread and price impact represent similar variation relative to their respective sample standard deviations, at approximately 6%. We do not find that the quote latency significantly influences the effective spread, the immediate trading costs borne by liquidity takers. In terms of trade latency, an opposite pattern is observed. A one-standard-deviation (1.01 milliseconds) increase in trade latency (i.e., slower exchange) significantly reduce the effective spread by 0.08 bps, representing only 2% of its sample mean while 5.6% of standard deviation. Interestingly, higher trade latency harms liquidity providers, significantly reducing the realized spread by 0.26 bps and increasing the price impact by 0.18 bps. To illustrate this point, we find that liquidity provision increases immediately before the completion of market orders with longer processing times, although such liquidity provision is not frequent. Additional liquidity reduces effective spreads, benefiting liquidity takers, but may increase adverse se-

[3]Since liquidity providers and takers participate continuously in electronic trading, it is unlikely that their quoting or trading activities are strongly conditional on the message traffic. This reasoning further supports the exogeneity of our instrument.

lection costs and reduce market-making revenues, for liquidity providers. Although some of our results are economically small in basis points, they accumulate over time since both liquidity providers and takers continuously participate in electronic trading. Furthermore, we also decompose quote messages into liquidity withdrawal and provision and a consistent explanation is obtained. In sum, our results imply that faster exchanges primarily enhance liquidity by facilitating timely quote updates, rather than enabling liquidity takers to more readily pick off stale quotes. Several robustness checks are conducted, and our main results are validated.

Our heterogeneity analysis reveals that all commodities exhibit a similar pattern. Moreover, the exchange latency of quote messages has stronger effects on highly liquid corn and soybean markets. For example, a one-standard-deviation (0.19 millisecond) decrease in quote latency significantly reduces price impact by 0.32–0.37 bps, approximately three times the magnitude estimated in the main analysis. We also find that faster exchanges have limited ability to reduce the price impact in highly liquid contracts and day trading sessions. Last, we do not find that the USDA public announcements significantly change the effect of the exchange latency on market quality.

Previous studies focus on both the trader-specific speed (e.g., Brogaard et al. 2015; Budish et al. 2015; Baron et al. 2019; Boehmer et al. 2021), and exchange speed (e.g., Menkveld and Zoican 2017; Pagnotta and Philippon 2018; Brolley and Cimon 2020), on market quality in stock markets. In terms of exchange latency, Menkveld and Zoican (2017) investigate the order processing latency within an exchange, whereas Pagnotta and Philippon (2018) and Brolley and Cimon (2020) extend the scope to the competition in speeds among different exchanges. Empirical works primarily focus on competition in exchange latency (e.g., Anderson et al. 2022), and trader's speed or messaging control (e.g., Skjeltorp, Sojli, and Tham 2016; Aldrich and Friedman 2023). However, relatively few studies empirically assess the order processing latency. Riordan and Storkenmaier (2012) examines the upgrade of trading system in Deutsche Boerse and find that market quality improves when a faster matching engine is introduced. Studies that attribute changes in market quality may overlook the role of system-wide message traffic as an explanation, especially multiple securities share the same order processing system, like in our case. We fill this gap by leveraging a uniform order processing system for all agricultural commodities and using a unique dataset that records all timestamps with the processing of each message. In doing so, our work provides empirical evidence on the predictions of the theoretical framework proposed by Menkveld and Zoican (2017). We show that faster exchanges primarily benefit liquidity providers, which is consistent with their "payoff-risk" channel.

We also contribute to literature on electronic trading in agricultural futures markets. Pre-

vious studies extensively document the effect of electronic trading on market liquidity (e.g., Wang, Garcia, and Irwin 2014; Hu, Serra, and Garcia 2020; Peng, Hu, and Robe 2024), information incorporation (e.g., Lehecka et al. 2014; Adjemian and Irwin 2018; Arzandeh and Frank 2019; Ma and Serra 2025), price behavior (e.g., Couleau, Serra, and Garcia 2019, 2020), and market resilience (e.g., He, Serra, and Garcia 2021). However, it remains unclear whether order processing speed by the exchange can help explain market quality and how fast processing benefits and costs are distributed between liquidity providers and takers, especially under the current trading architecture.[4] We differ from the literature by shedding light on the congestion-based externality. Moreover, literature on speed in commodity markets remains limited. In terms of trader's speed, Fishe, Haynes, and Onur (2019) suggest that trader's speed is not related to anticipatory trading in the West Texas Intermediate (WTI) crude oil futures markets. Huang, Serra, and Garcia (2022) show that low-latency traders earn non-negligible profits by adversely selecting stale quotes during the USDA announcements. Coughlan and Orlov (2023) use a regulatory account-level dataset and reveal that the high-frequency trading is strongly associated with an improvement of market quality. We complement them by showing how exchange processing, which is independent to trader's speed choice, affects market quality. When multiple futures contracts share the same order-processing system, as in our setting, deterioration in market quality may arise from message traffic congestion rather than only from traders' speed arms race. Answering this question can better help exchange design for cross-commodity allocation of processing capacity.

Agricultural shareholders have expressed multiple concerns about electronic trading and corresponding exchange design. Livestock market participants questioned whether high-frequency traders contributed to the elevated volatility and complicated hedging activities in futures markets during 2016–2017. CME then responded with various reforms to enhance market integrity (CME 2016). Although our paper does not focus directly on high-frequency traders, it sheds light on the implications of exchange-level latency, which affects all market participants regardless of their own speed choices. Our findings suggest that regulators should consider the potentially trade-offs resulting from exchange latency between liquidity providers and takers. Specifically, lower exchange latency in agricultural commodity markets appears to primarily benefit liquidity providers by enabling timely quote updates and reducing their exposure to adverse selection, while higher exchange latency helps liquidity takers lower their trading costs.

Our study also has implications for exchange design. Because the current order-processing

[4]Since May 2020, all agricultural commodities have migrated to the same order processing system. Previously, soybeans and soybean complex were processed in a different system than other major agricultural commodities, such as corn.

architecture groups multiple futures markets within shared systems, exchanges should account for cross-commodity congestion when allocating processing capacity and evaluating infrastructure upgrades. More importantly, as the Commodity Futures Trading Commission (CFTC) has recently requested feedback regarding 24/7 trading in several commodity futures markets,[5] our work is timely for a potential expansion of round-the-clock operation. Extending trading hours would require exchanges to maintain reliable order-processing capacity during both peak and off-peak periods. Our robust night-session results, which can represent off-peak expansion periods (e.g., weekends), reveal that exchange-level latency during less active periods can still affect market quality. Exchanges should therefore carefully evaluate the message traffic and order-processing intensity across different periods before launching the 24/7 trading on a large scale.

2 Data and institutional details

2.1 CME message processing

We first illustrate how traders' orders are processed in the CME, which is important for our empirical analyses. We exhibit a simplified electronic trading skeleton in Figure 1.[6] Traders submit orders for different agricultural commodities and these orders need to arrive at the same Market Segment Gateway 70 for all agricultural commodities. All messages are queued based on a first-in-first-out (FIFO) sequencing. Then, a matching engine processes all messages, including both limit and market orders for every commodity. For limit orders, it checks whether they can be directly executed or need to rest in the limit order book (LOB). For market orders, it matches the resting limit orders based on stipulated matching algorithms.[7] The matching engine is operated for each contract and each commodity, after which these orders become effective. Afterwards, all messages are disseminated from the market data feed to clients who reconstruct the LOBs for each commodity. Technically, orders for a specific commodity directly affects the market quality of that commodity rather than others. For example, orders of corn futures do not affect soybean's market quality.

The whole order processing system is like a one-way and single-lane tunnel. Under FIFO sequencing, an order's processing time depends on the number of messages already ahead of it in the queue, thereby generating exchange latency. If the tunnel is congested, the

[5] See https://www.cftc.gov/PressRoom/PressReleases/9259-26.

[6] We show Figure 1 for illustration purposes only and the actual order processing system is more complicated.

[7] Note that the matching algorithms only work within the matching engine, and they may not be fully FIFO-based in agricultural futures markets. However, this does not affect the nature of FIFO-based entry priority within the order processing system.

exchange latency becomes higher. Thus, although traders may race to be first to access the exchange, they cannot control the queue ahead of them in the tunnel. Message processing can be delayed by a congested tunnel and ultimately influence the market quality of the corresponding commodity market.

[Insert Figure 1 here.]

2.2 Data

We use the CME packet capture (PCAP) data obtained from the CME Datamine through January 8, 2024, to December 27, 2024. The packet is a basic unit to transmit raw messages over the network. Our PCAP data cover all agricultural futures and record raw messages for all market events, allowing us accurately calculate the queue length within the order processing system.[8] Previous studies (e.g., Ma and Serra 2025) typically focus on highly-liquid commodities (e.g., corn and soybeans), thus the queue length can hardly be measured as their data cannot cover messages of other commodities that may be positioned ahead.[9] More importantly, the PCAP data record the chronological ordering of all messages, which is independent of commodities, helping us assessing the order processing system as a whole. All messages have two timestamps (Figure 1): receiving timestamp when an order enters the gateway and the other is the sending time when a message leaves the market data feed.[10] All timestamps are in a nanosecond resolution. A unique message sequence number is also assigned to help sort messages.

We focus on six agricultural commodities with various market liquidity: corn, soybeans, Chicago wheat, Kansas wheat, soybean meal, and soybean oil. They share the same trading schedules with two sessions: night session is from 19:00–7:45 and day session is from 8:30–13:20, all U.S. Central Time (see Figure A1 of Appendix A.1). We use both front- and back-month contracts. The front-month contract is typically the most-traded one and we roll to the next most-traded contract when its trading volume exceeds the previous one for two consecutive trading days.[11] The grain futures are quoted in cents/bushel with each

[8] Although livestock futures markets also share the same market segment gateway, our PCAP data do not cover them since their data are recorded in a different market data feed. However, since they have different trading schedules than most agricultural commodities, we exclude them from our sample. Since livestock markets typically are less liquid (e.g., He, Serra, and Garcia 2021), excluding them is unlikely to materially affect our conclusions.

[9] The PCAP data include all data products used in previous studies, including both Market by Price (e.g., Arzandeh and Frank 2019; He, Serra, and Garcia 2021), and Market by Order (e.g., Ma and Serra 2025).

[10] Our data do not record the timestamp when a message leaves the matching engine and the calculated exchange latency might be slightly higher. However, since such latency increase applies to every message, we argue this is not likely to materially affect our conclusions.

[11] As a result, the September contract for corn, August and September contracts for soybeans, and the August, September, and October contracts for soybean meal and soybean oil are not selected.

contract representing 5,000 bushels. Soybean meal (soybean oil) is quoted in dollars/short ton (cents/pound) with contract size 100 short tons (60,000 pounds). Tick sizes (minimum price increments) are 0.25 cents/bushel for grains, 10 cents/short ton for soybean meal, and 0.01 cents/pound for soybean oil.

For exchange latency calculation, we consider quote messages at the best-bid-offer (BBO) as they are the most relevant to our market quality measures (section 3.2). Our PCAP data records all quote messages that are at the order-level. Every limit order is assigned a unique order ID for tracking. We only consider quotes that are posted during the continuous trading sessions, thus excluding good-till-day and good-till-cancel quotes.[12] We follow Ma and Serra (2025) to clean the quote messages to ensure all of them are initiated by traders.[13] We then decompose the quote messages into liquidity provision and liquidity withdraw messages. Liquidity provision refers to quotes that either post better BBO prices or increase the quoted quantity at the BBO, while liquidity withdrawal refers to those that either revise prices off the BBO or reduce the quoted quantity at the BBO.

Our trade messages refer to trader's market orders.[14] For each trade, our data record both the market order message and that of the corresponding matching (execution) limit orders. We do not consider each limit orders' execution messages to avoid double counting. This activity is not initiated by traders either. Our data directly provides information about trade direction to identify whether a trade is buyer-initiated or seller-initiated.

To measure the overall message traffic, we consider all futures contracts that have the same trading schedules as our focal commodities for comparability.[15] Our focal messages come from both the quote and trade messages from the six selected commodities. Moreover, CME also operates an internal computing to generate messages about the status of LOB and implied quotes (Arzandeh and Frank 2019), which also affects the exchange latency. Hence, when we calculate our *queue-ahead* variable (section 3.3.1), we also consider these internal messages. We construct consolidated limit order books based on Arzandeh and Frank (2019). Following Easley, de Prado, and O'Hara (2016), we also preprocess our data to remove potentially erroneous messages.

[12] Quotes submitted outside the continuous trading sessions represent less than 1% of total orders submitted at a trading day.

[13] This avoids double counting when a limit order get either partially or fully executed. Under these scenarios, a revision or deletion message is generated by the exchange to update the limit order status. See more details in Ma and Serra (2025).

[14] Specifically, as mentioned in Ma and Serra (2025), we define trades as trade summary messages that record the execution of the aggressor's market order.

[15] This include all types of futures contracts, typically including calendar spreads, Trade-at-Settlement (TAS), mini-sized, etc. Also, our coverage of commodities is over those six selected.

3 Exchange latency and market quality

3.1 Exchange latency

According to Figure 1, we calculate the exchange latency for message i as follows:

$$\text{Exchange latency}_i = \text{Sending time}_i - \text{Receiving time}_i. \tag{1}$$

Figure 2 displays the hourly median latency for the BBO quote messages and the number of messages for both front- and back-month contracts. Generally, all markets show a similar pattern. We find the quote is around half milliseconds throughout the day in all markets, and the latency for back-month contract is slightly higher than that of the front-month contract. The quote latency is higher during the market open at the night session, which reaches to 0.6 milliseconds, while gradually reduces to around 0.3 milliseconds, representing a typical less active nature of the overnight trading. Consistently, the night session witnesses fewer messages than the day session. Interestingly, we find the quote latency slightly increases after 1 a.m. and stabilizes around 0.4 milliseconds until the night session ends. This might be related to the fact that overseas traders (e.g., Europe) start to trade during this period (Bondarenko and Muravyev 2023), which is also consistent with Ma and Serra (2025). During the day session, massive quote messages are submitted, especially at market open and close, reaching a level about ten times that observed during the night session. However, we do not find a large variation during the day session and the quote latency is all around 0.4 milliseconds.

[Insert Figure 2 here.]

We also assess the trade latency and exhibit the results in Figure 3. Generally, the trade latency is slightly higher than the quote latency. We find that hourly median trade latency ranges from 0.32 to 1.90 milliseconds, with the Kansas wheat market exhibiting the largest variation, especially for the back-month contract. This is consistent with the operation of the matching engine: it needs additional time to match the resting limit orders when market orders come. Similarly, the front-month contract has a slightly lower trade latency than the back-month contract, and most trades occur during the day session. Compared to the quote messages, all markets witness substantially less trade message volume, which is consistent with Ma and Serra (2025) who suggest that quotes dominate in agricultural futures markets.

[Insert Figure 3 here.]

We further decompose the quote messages into liquidity provision and withdrawal, and show their hourly distributions in Figure 4. We pool the front- and back-month contracts.

The overall pattern is similar to that in Figure 2. All markets suggest that the liquidity withdrawal messages have slightly higher latency than liquidity provision, with the largest different observed in the corn market. This indicates that liquidity providers may experience higher processing delay when they reduce their order aggressiveness. Moreover, the two message categories exhibit similar message volume, with liquidity-provision messages being slightly more prevalent.

[Insert Figure 4 here.]

Importantly, across all commodities, exchange latency does not appear to be strongly correlated with message volume, suggesting that even large surges in message traffic do not necessarily translate into longer processing latency. Instead, latency might depend more directly on the number of messages already queued ahead: even during periods of heavy message arrivals, latency can remain low when the existing queue is short.

3.2 Market quality measures

Following Hendershott, Jones, and Menkveld (2011), we use three market quality measures that are widely used in the literature (e.g., Hasbrouck and Saar 2013; Brogaard et al. 2015; Boehmer et al. 2021).[16] These measures can better characterize interactions between liquidity providers and liquidity takers.

We calculate the effective spread, which is defined as the difference between trade execution price and the prevailing midpoint price:

$$\text{Effective spread}_t = 2 \times \frac{q_t \left(p_t - m_t\right)}{m_t}, \tag{2}$$

where q_t denotes the trade direction that equals $+1$ for buyer-initiated trades and -1 for seller-initiated trades. p_t denotes the trade price and m_t is the prevailing midpoint price at the time of trade. As mentioned in section 2.2, our data can directly identify the trade direction. From the liquidity taker's perspective, the effective spread measures the immediate cost of trade execution. In other words, it measures the gross compensation of liquidity providers for supplying immediacy to liquidity takers.

When liquidity providers trade with their counterparties, they subsequently need to manage or unwind the resulting inventory to realize market-making revenues. However, when they trade with liquidity takers who have information and/or speed advantages, their quotes

[16] We do not consider other market quality measures, such as market depth, as quote messages might have direct effects on market depths. For example, liquidity withdrawal messages have direct reductions in depths, and this might not strongly related to exchange latency.

may become stale and be picked off, exposing them to adverse selection losses because they lack comparable advantages (Kyle 1985; Glosten and Milgrom 1985). We calculate the realized spread to measure the gross revenue of market making, which is defined as the difference between the trade price and the midpoint price 5 seconds later:

$$\text{Realized spread}_t = 2 \times \frac{q_t \left(p_t - m_{t+5s}\right)}{m_t}, \tag{3}$$

where m_{t+5s} denotes the midpoint price 5 seconds after the time of trade execution. O'Hara (2015) suggest that under the current trading environment, liquidity providers typically unwind their inventory in a short horizon. We select a 5-second horizon in our analysis.[17] Intuitively, if the midpoint price moves against the position direction, liquidity providers are likely to earn positive revenues from supplying liquidity. To measure the adverse selection costs, we calculate the price impact, which is a function of the change in the midpoint price from the time of trade execution to 5 seconds afterwards:

$$\text{Price impact}_t = 2 \times \frac{q_t \left(m_{t+5s} - m_t\right)}{m_t}. \tag{4}$$

Thus, if the midpoint price moves towards the inventory direction, liquidity providers cannot unwind their inventory with profits, and incur adverse selection losses. The three measures have the following arithmetic relationship:

$$\text{Effective spread}_t = \text{Realized spread}_t + \text{Price impact}_t, \tag{5}$$

which suggests that the liquidity takers' immediate trading costs are decomposed into market-making revenue and adverse selection costs for liquidity providers. We calculate the three measures for each trade and aggregate them using trading-volume-weighted average for each hour. Table 1 summarizes our definitions.

[Insert Table 1 here.]

Table 2 reports the summary statistics of our hourly market quality measures, pooling all commodities. We find the average effective spread is 4.10 bps, while the largest constituent is price impact (3.90 bps), accounting for about 95%. However, the average realized spread is only 0.20 bps. Our results are consistent with Hu, Serra, and Garcia (2020). The possible reason for such distribution is large price impacts are typically resulted from large trades that substantially consume the BBO liquidity. When we calculate the volume-weighted averages,

[17] Table E4 of Appendix E.4 confirms that our results are qualitatively similar using 1-second, 10-second, and 30-second windows.

such large trades receive higher weights, and in turn increase the average values.

3.3 Endogeneity and identification strategy

When assessing the relationship between market quality and exchange latency, a noteworthy concern is the endogenous exchange latency. First, since traders may have their own quoting or reference preference, the exchange latency is not likely to be exogenous. Thus, a standard OLS regression may produce biased estimates. Second, exchange latency is likely to condition on market quality. For example, worse market quality may also affect trader's messaging strategy of inventory management, possibly influencing exchange latency. Hence, a reverse causality may occur. To facilitate causal interpretation, we use the instrumental variable method to address the endogeneity.

3.3.1 Instrumental variable

As discussed in section 3.1, the exchange latency is likely to be more related to the existing queue rather than the message inflow. Following this logic, we use the number of messages from other commodities positioned ahead a focal message in the queue upon its arrival at the exchange as our instrument variable. In short, we name it as *queue-ahead*. Specifically, the queue ahead of message i is constructed as follows:

$$\begin{aligned} Queueahead = &\#\text{messages}\{j : \text{Receiving time}_j < \text{Receiving time}_i, \\ &\text{Sending time}_j > \text{Receiving time}_i, \text{Commodity}_i \neq \text{Commodity}_j\}. \end{aligned} \tag{6}$$

Three criteria are considered here. First, message j needs to arrive at the gateway prior to message i (i.e., Receiving time$_j$ < Receiving time$_i$). Second, the message j should leave the market data feed after message i has arrived at the gateway (i.e., Sending time$_j$ > Receiving time$_i$). Hence, our *queue-ahead* variable measures the number of messages that have already been queued yet has not completed processing upon a focal message enters the processing time. More importantly, we apply the "leave-one-out" method, which means message j is not associated with the same commodity as message i. For example, when an incoming message is associated with corn, we only count ahead messages from other commodities. This restriction helps reduce the likelihood that our instrument has direct effect on market quality. As mentioned in section 2.2, we consider both quote messages from traders and internal messages from the exchange. If no messages positioned ahead, the instrument is assigned a zero value. We calculate the *queue-ahead* variable for each quote and trade message and aggregate it using the hourly median value.

3.3.2 Instrument validity support

A valid instrument needs to be plausibly exogenous. We argue that our *queue-ahead* instrument satisfies the endogeneity in our setting. First, liquidity providers and takers continuously participate in electronic trading (O'Hara 2015), especially designated market makers.[18] Hence, it is unlikely that their quoting or trading strategies are highly conditional on the message traffic in the "tunnel." Considering profitable trading signals are typically short-lived in current trading environment (O'Hara 2015), traders are not likely to delay their trading schedules based on the message traffic. In other words, traders do not intend to quote more when the message traffic is light. Conversions with market participants also confirm this.[19] Second, once an order is submitted, it can affect only subsequent messages and is unrelated to the messages already queued ahead of it. Thus, our *queue-ahead* instrument should not be correlated to trader's own messaging preference.[20] Together, these two arguments support the plausibility of our instrument's exogeneity.

Figure 5 depicts the relationships between the exchange latency (quotes and trades) and our instrument, *queue-ahead*, at each hourly interval. Obviously, we find for both quotes and trades, higher exchange latency is associated with more messages positioned ahead. Such positive relationships are observed across all hourly intervals. Figure 5 also help supports the exogeneity of our instrument. We do not observe a dense clustering of data points around the origin, suggesting that strategic order messaging based on the tunnel's message traffic is unlikely. This also supports the relevance of our instrument.

As discussed in section 3.3.1, we only count messages queued ahead from other commodities.[21] Hence, our exclusion restriction assumes that messages queued ahead from other commodities affect the market quality of the focal commodity only by delaying the processing of its messages, while considering relevant controls. The reasoning is that orders are specific to commodities. This restriction helps us exclude possible direct relationships between the instrument and outcome variable. We acknowledge that our exclusion restriction is intuitively compelling but empirically untestable.

A noteworthy concern is unobservable factors that are associated with the market quality

[18]Sophisticated market-making firms may enter a designated market-maker arrangements with the CME, under which they receive incentives in exchange for satisfying obligated continuous liquidity provision. See more details: `https://www.cmegroup.com/markets/market-maker-directories.html`.

[19]CME has implemented a messaging efficiency program, and firms are strictly prohibited from intentionally creating systematic processing delays through excessive messaging practice.

[20]Conventional queuing theory typically views the order arrival as random and follows a specific statistical distribution (e.g., Poisson). Recent theoretical studies also typically assume order arrival is random. See Li, Wang, and Ye (2021) and Yueshen (2026).

[21]Specifically, we exclude all kinds of futures contracts related to the same commodity, such as mini-sized, micro-sized, calendar spreads, and inter-commodity spreads, thereby producing a conservative measure.

are likely to be also related to our instrument (Danieli et al. 2026). Unobservable factors, such as information flows, competition among traders, and broader market conditions, are likely to affect market quality (e.g., Adjemian and Irwin 2018; Brogaard and Garriott 2019) by shaping traders' responses. For example, frequent information arrivals may prompt traders to adjust their quotes more intensively, thereby influencing market quality. We concern these unobservable variables are likely to be correlated with our instrument, thus violating the exclusion restriction and the exogeneity. If so, our instrument may primarily capture the preexisting condition message traffic, rather than a causal effect of exchange latency.

Following Danieli et al. (2026), we conduct a placebo test to examine whether our instrument is associated with the lagged market quality measures. Table B1 of Appendix B.1 suggests that no significant relationship is found between our instrument and the lagged market quality measures. This reveals that unobservable factors associated with outcomes appear not to be correlated with our instrument, thus validating our identification strategy.

Moreover, our instrument construction also sheds light on the congestion-based externality within the shared order processing system. Messages generated from other commodities can delay the order processing of a focal message, which is strongly correlated with the exchange latency.

3.4 Two-stage least squared regression

3.4.1 Regression specifications

To implement our instrument, we first regress our *queue-ahead* variable on the market quality measures in the first stage to obtain the predicted value of exchange latency. Specifically, our first-stage regression is

$$\begin{aligned} Exchange\ Latency_{i,m,h,t} =& \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \\ & \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t}. \end{aligned} \tag{7}$$

where *Queueahead* is our "leave-one-out" instrument. We report our first-stage regression results in Table C1 of Appendix C.1. Consistent with Figure 5, we find our instruments are significantly related to the exchnage latency. In the second stage, we estimate the following regression:

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t}, \tag{8}$$

where $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency of individual futures contract i of commodity m at hour h on day t.[22] MQ refers to market quality measures, including effective spread (ES), realized spread (RS), and price impact (PI). Our panel data pool all commodities together. We include individual futures contract fixed effects λ_i and commodity fixed effects γ_m to capture time-invariant contract and commodity attributes. We also include hour $\times$ day interacted fixed effects to capture the contract and commodity invariant attributes, such as common fundamental information. Standard errors are double clustered by individual futures contract and day. We winsorize all variables at the 1st and 99th percentiles by contract and hour to mitigate outlier influence.

We select our controls according to the literature (e.g., Hendershott et al. 2011; Brogaard et al. 2015; Shkilko and Sokolov 2020; Anderson et al. 2022). We include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. Log trade size captures variation in average market order size that may independently affect market quality, while inverse price accounts for commodity price-level differences that may influence the construction of proportional market quality measures. We also account for realized volatility and control the general messaging intensity. Detailed variable descriptions about our regression are reported in Table 1 and summary statistics are reported in Table 2. We obtain 52,124 observations in total.

[Insert Table 2 here.]

3.4.2 Baseline results

We report our baseline results in Table 3. For each market quality measure, we estimate a regression with only quote latency and another one with both quote latency and trade latency. The results of all market quality measures are organized in columns. We calculate the Kleibergen and Paap (2006)'s F-statistics and all regressions show that the F-statistics are greater than 10, indicating our instrument is not weak.

Generally, we find the exchange latency has distinct influence on different trader groups. For liquidity providers, we find the quote latency does not have a statistically significant effect on effective spread in both specifications (1) and (2). This suggests that quote processing latency does not affect the liquidity taker's immediate trading costs. However, we find the quote latency has a negative and significant effect on the realized spread, indicating lower quote latency (i.e., faster exchange) improves the gross market-making revenue. Moreover, the magnitude is economically meaningful. A one-standard-deviation (0.181 milliseconds, see

[22] Since we have two instruments for two endogenous variables, our regression is exactly identified.

Table 2) decrease in the quote latency increases the realized spread by 0.116–0.163 bps, representing 59.18%–83.16% of its sample mean (0.196 bps). Similarly, the lower quote latency also significantly reduces the price impact, suggesting that faster quote processing allows liquidity providers to bear lower adverse selection costs. A one-standard-deviation decrease in the quote latency reduces the price impact by 0.12–0.15 bps, representing merely 3.08%–3.85% of its sample mean (3.896 bps). Although the relative changes differ substantially, we find their effects represent similar variation relative to their respective sample standard deviations (about 6%).

For liquidity takers, we find the trade latency has a significant and positive effect on the effective spread, suggesting slower exchange benefits them. A one-standard-deviation increase in the trade latency (1.014 milliseconds) reduces the effective spread by 0.08 bps, representing 1.96% of its sample mean while 5.60% of its standard deviation. Interestingly, we find that liquidity providers still benefit from the lower trade latency, which is seemingly counterintuitive. Specifically, one-standard-deviation decrease in the trade latency improves the realized spread by 0.26 bps and 0.18 bps. A possible explanation is that higher trade latency may allow liquidity takers match with better quotes submitted before trades finish their processing, thereby reducing the effective spread. However, such better quotes may be more likely to be exposed by adverse selection risk, ultimately reducing realized spread and increasing price impact.

To evaluate this point, we follow the construction of our *queue-ahead* variable to count liquidity provision messages (section 2.2) that are queued ahead of a focal trade message and associated with the same futures contract. Figure C1 of Appendix C.2 depicts the relationship between the number of liquidity provision messages and the trade latency by buyer- and seller-initiated trades. Consistent with our explanation, higher buyer (seller)-initiated trade latency is associated with more liquidity provision messages at the best ask (bid) side. However, such liquidity provision is not frequent. On average, only 1–2 liquidity provision messages queued ahead of a trade. This may also explain why the reduction in the effective spread is limited in magnitude.

We alternatively assess the economic meaning of our results through a dollar term. Since liquidity providers and takers continuously participate in trading, the estimated effects are accumulated over time. Considering the daily average total dollar trading volume in the six commodity markets is about \$12 billion, hence, a one-standard-deviation decrease in the quote latency improves daily realized spread by about \$143.18–\$201.19 thousand, which translates to \$0.39–\$0.59 per contract.[23] Similarly, the same magnitude of the quote latency

[23]For example, the per-contract amount is: \$12,343,057,355 $\times$ 0.116 bps $\times$ 0.0001 $\div$ 366,697 contracts $\approx$ \$0.39.

reduction reduces the daily price impact by \$0.40–\$0.50 per contract. Although the daily dollar effect is economically small compared to the total trading volume, its cumulative effect is non-negligible. Annually, it is equivalent to \$36–\$50 million. For liquidity takers, we find a one-standard-deviation increase in the trade latency reduces the immediate trading costs by \$98 thousand, or \$0.27 per contract.

The control variables generally behave consistently with economic intuition. For example, large trade typically bears higher effective spread, reducing the realized spread while increasing price impact. Similar effects are also observed for volatility.

In sum, our findings imply that faster exchanges help liquidity providers facilitate timely and efficient quote updates, rather than allowing readily picking off stale quotes. Liquidity providers may benefit from better liquidity provision when their market orders can be slowly processed. We provide empirical evidence for Menkveld and Zoican (2017) and support their "payoff-risk" channel.

[Insert Table 3 here.]

3.4.3 Liquidity provision and liquidity withdrawal

Our baseline results in Table 3 show the overall effect of the quote latency on the market quality. However, different quote messages may have heterogeneous effects on the market quality. For example, the reduction in price impact and the increase in realized spread are more likely to be driven by liquidity withdrawal messages. Hence, we reexamine the effects of quote latency on market quality by assessing whether they differ between liquidity provision and liquidity withdrawal messages.

Specifically, we estimate the following regression:

$$\begin{aligned} MQ_{i,m,h,t} =& \beta_1 \times \widehat{\text{Withdrawal latency}}_{i,m,h,t} + \beta_2 \times \widehat{\text{Provision latency}}_{i,m,h,t} + \\ & \beta_3 \times \widehat{\text{Trade latency}}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t}, \end{aligned} \tag{9}$$

where $\widehat{\text{Withdrawal latency}}$ and $\widehat{\text{Provision latency}}$ denote the fitted values of liquidity-withdrawal latency and liquidity-provision latency from the first-stage regressions, respectively. We calculate the *queue-ahead* variables for each type of quote messages separately and aggregate them by hourly median values. We include the same set of controls and fixed effects as our main specification (Equation 8).

Table 4 reports the results. We find faster processing of liquidity withdrawal messages increases the effective spread though the effect is weakly significant at the 10% level. Moreover, faster processing of liquidity provision messages reduces the effective spread, although the effect is statistically significant only in specification (2) at the 5% level. This pattern

is consistent with our explanation in section 3.4.2: When liquidity provision messages are processed before a trade, liquidity takers can benefit from the these quotes, thereby reducing the effective spread. The opposite is true for liquidity withdrawal messages and these are all confirmed in Table 4. Consistent with our expectation, we find our main results are primarily driven by the liquidity withdrawal messages with larger effects. A one-standard-deviation (0.208 milliseconds) decrease in their latency is expected to improve the realized spread by 0.39–0.42 bps and reduce the price impact by 0.32–0.34 bps. Consistently, opposite effects are found for the liquidity withdrawal messages. Their slow processing benefit liquidity providers through increasing the realized spread by 0.17–0.19 bps and reducing the price impact by about 0.12 bps, following a one-standard-deviation change (0.175 milliseconds). Last, trade latency remains similar effects as our baseline results.

[Insert Table 4 here.]

3.4.4 Announcement analysis

Public announcements may change how the exchange latency affects the market quality. To assess this, we estimate our regression during 9:30–12:30 using a 5-minute sample. Following Huang, Serra, and Garcia (2022), we consider the World Agricultural Supply and Demand Estimates (WASDE), Grain Stock, Prospective Plantings, Acreage, and Crop Production announcements, yielding 15 announcement days in total. All announcements are released at 11:00, U.S. Central Time. We exclude soybean meal and soybean oil markets because all announcements are pertinent to grains. We follow Fleming, Mizrach, and Nguyen (2018) to select five days proceeding and following each announcement day as our non-announcement counterparts. Our regression specification is

$$\begin{aligned} MQ_{i,m,h,t} =& \beta_1 \times \widehat{Exchange\ Latency}_{i,m,h,t} \times Ann_{h,t} + \beta_2 \times \widehat{Exchange\ Latency}_{i,m,h,t} + \\ & \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_t + \varepsilon_{i,m,h,t}, \end{aligned} \tag{10}$$

where Ann is a dummy that equals 1 after 11:00 on announcement days, and 0 otherwise.[24] Thus, our parameter of interest β_1 captures how the exchange latency behaves after the news arrival, compared to pre-release periods and non-announcement counterparts. Our control variables also include the Ann dummy itself.

We report the results in Table D1 of Appendix D. We find the effect of the quote latency still remain for the realized spread and and price impact. However, different from the baseline results, lower quote latency significantly increase the effective spread at the 5% level. Unchanged effects are also found for the trade latency. However, we find that the

[24] Our instrument is also interacted with the announcement dummy when estimating the regression.

interaction term involving quote latency has only weak effects on market quality, with estimates either insignificant or significant only at the 10% level. This indicates that the news arrival does not significantly change the documented relationship between the quote latency and market quality. Furthermore, slower processing of trade after news arrival only significantly reduces the immediate trading costs by 0.27 bps, compared to pre-release periods and non-announcement counterparts.

In sum, our results do not suggest that the exchange latency has a heterogeneous effect on the market quality during announcement days. A possible explanation is that severe processing congestion following report releases is typically short-lived and therefore does not substantially affect the median queue length.

3.4.5 Robustness analysis

We first assess whether our results our baseline results are driven by announcements. We examine this by excluding the announcement days in our sample. Table E1 of Appendix E.1 reports the results. We find the documented effects still remain with comparable mastitides. Thus, our findings are not only pertinent to public announcement days when considerable fundamental news arrives at the markets.

Market open and close typically attract intensive trading and quoting activities, when massive messages are observed, as shown in Figures 1–3 as well as in the literature (e.g., Lehecka, Wang, and Garcia 2014). We examine whether our results are still robust when excluding the market open and close periods. Specifically, we reestimate our baseline regression by excluding three hourly intervals: 19:00–20:00, 8:30–9:30, and 12:30–13:20. The results are reported in Table E2 of Appendix E.2. Our findings remain robust after excluding these periods. This is also consistent with the fact that the market quality is strongly related to the queue length ahead of incoming messages, rather than the number of incoming messages.

Another concern is that our hourly sample might still be too coarse to fully capture the effect of the exchange latency on market quality. We also assess whether our findings are robust to a 30-minute sample and the results are reported in Table E3 of Appendix E.3. We find the documented relationships are qualitatively similar to our baseline results with slightly larger effects. For example, a one-standard-deviation (0.425 milliseconds) decrease in quote latency is expected to reduce the price impact by 0.20–0.29 bps.

We finally check whether our main results are sensitive to the time horizons of our market quality measures. We consider realized spreads and price impacts at 1 second, 10 seconds, and 30 seconds after the time of trade execution. The results are reported in Table E4 of Appendix E.4. We find that our main findings are still robust under different time horizons. Moreover, we find that the magnitude of the effect of quote latency increases incrementally

over longer time horizons. However, an opposite pattern is found for the trade latency. In sum, our main findings are not sensitive to the time horizons.

3.4.6 Heterogeneity analysis

Our main results are based on pooled panel data, which may obscure potential heterogeneity across commodities, contract months, and trading sessions. Hence, we conduct heterogeneity analyses to assess whether our documented relationships vary across these dimensions.

We first estimate separate regressions by commodities. Specifically, we group corn and soybeans, Chicago wheat and Kansas wheat, and soybean meal and soybean oil into three separate commodity groups. The results are reported in Tables F1–F3 of Appendices F.1–F.3, respectively. Overall, the effects of the exchange latency on market quality remain qualitatively similar across all commodity groups, though with some exceptions. We find that the corn and soybean markets, the most liquid commodities, witness the largest effects on market quality, for both types of latency. For example, a one-standard-deviation (0.194 milliseconds) decrease in the quote latency is expected to increase the realized spread by 0.28–0.34 bps while reducing the price impact by 0.32–0.37 bps. These effects are about as three times as those reported in the baseline results. However, the trade latency does not significantly affect the effective spread in corn and soybeans. The remaining commodity groups exhibit coefficients generally consistent with our main results. Overall, these findings suggest that the influence of exchange latency is more pronounced in the most liquid commodity markets.

We also estimate separate regression by different contract months, polling all commodities, and the results are reported in Table F4 of Appendix F.4. We find the effects are larger for the back-month contracts that are typically less liquid than the front-month contracts, especially for the quote latency. Moreover, we find that the trade latency does not significantly affect the price impact for the front-month contracts where information is more exposed. These findings suggest that less liquid contract months are more likely to need exchange speed.

We last assess how exchange latency affects the market quality at different trading sessions with the results reported in Table F5 of Appendix F.5. We find that the effects are relatively more pronounced during the day session, with coefficients slightly larger than those estimated for the night session. However, both trade and quote latency have only weak effects on price impact during the day session. This finding is consistent with markets incorporating information more actively during the day session. Taken together, our results suggest that faster exchanges have limited ability to reduce price impact in highly liquid contracts and trading sessions.

4 Conclusions

Exchanges, which serve as intermediaries among market participants, have also invested considerably in upgrading their own trading infrastructure. Currently, all agricultural futures markets share a common order processing system within the CME. This paper examines whether the exchange latency, defined as the order processing time, affects the market quality in six CME agricultural commodities: corn, soybeans, Chicago wheat, Kansas what, soybean meal, and soybean oil. Using the raw-message PCAP data in 2024, we find the exchange latency, which is independent to traders' own speed choice, affects both liquidity providers and liquidity takers with distinct effects.

We find that the BBO quote message latency is about half millisecond while the trade message latency is around one millisecond, across all commodities. This pattern is consistent with the operation of matching engine: trades need additional time to match rest limit orders.

Our identification strategy relies on the institutional features of CME agricultural markets. We use an instrumental variable, named *queue-ahead* to address the endogeneity. Specifically, we calculate the number of messages from other commodities queued ahead of a focal message upon its entry to the market segment gateway. The construction of our instrument follows the "leave-one-out" logic, reducing the likelihood that it can directly affect market quality. Hence, our exclusion restriction assumes that messages from other commodities positioned ahead a focal message only affects the focal commodity's market quality by delaying the processing of that message. We find our *queue-ahead* variable is strongly associated with the endogenous exchange latency and our regression does not suffer from weak instrument issue.

Following Hendershott, Jones, and Menkveld (2011), we consider three market quality measures: effective spread, realized spread, and price impact to assess trading outcomes for both liquidity providers and takers. Our 2SLS-IV regressions suggest that a one-standard-deviation decrease in quote latency improves the realized spread and reduce the price impact by about 0.12 bps ($0.40 per contract). However, it does not significantly affect the effective spread, which measures the immediate trading costs borne by liquidity takers. For trade latency, a one-standard-deviation increase in it reduces the effective spread by 0.08 bps ($0.27 per contract). Our results show that liquidity takers can benefit additional liquidity provision prior to their market orders get executed. Consistent with the quote latency, faster trade latency also benefits liquidity providers. Although the magnitudes of the effects appear to be economically small in basis points, they accumulate over time since both liquidity providers and takers participate electronic trading continuously. In sum, our findings suggest that faster exchanges primarily benefit liquidity providers, whereas slower exchanges benefit

liquidity takers.

Our study provides empirical evidence on a theoretical work by Menkveld and Zoican (2017). Furthermore, we shed light on a congestion-based externality in the current agricultural commodity trading. We characterize that message traffic generated from one commodity can delay order messaging in others, potentially imposing congestion costs on other market participants. This implication is helpful for the exchange design. Exchanges should account for cross-commodity congestion when implementing market reforms. More importantly, as the CFTC recently solicited public feedback on 24/7 trading, our findings can inform the design of order-processing systems for both peak and off-peak trading periods, thereby facilitating a smooth transition to round-the-clock trading.

Future research could examine how exchange latency interacts with algorithmic trading and how the two jointly affect market quality. Such analysis would connect trader speed with exchange speed, providing a more complete picture of fast electronic trading.

References

Adjemian, M. K. and S. H. Irwin (2018). USDA announcement effects in real-time. *American Journal of Agricultural Economics 100*, 1151–1171.

Aldrich, E. M. and D. Friedman (2023). Order Protection Through Delayed Messaging. *Management Science 69*, 774–790.

Anderson, L., E. Andrews, B. Devani, M. Mueller, and A. Walton (2022). Speed segmentation on exchanges: Competition for slow flow. *Journal of Financial Markets 58*, 100632.

Arzandeh, M. and J. Frank (2019). Price Discovery in Agricultural Futures Markets: Should We Look beyond the Best Bid-Ask Spread? *American Journal of Agricultural Economics 101*, 1482–1498.

Aït-Sahalia, Y. and M. Sağlam (2024). High frequency market making: The role of speed. *Journal of Econometrics 239*, 105421.

Baron, M., J. Brogaard, B. Hagströmer, and A. Kirilenko (2019). Risk and return in high-frequency trading. *Journal of Financial and Quantitative Analysis 54*, 993–1024.

Biais, B., T. Foucault, and S. Moinas (2015). Equilibrium fast trading. *Journal of Financial Economics 116*, 292–313.

Boehmer, E., K. Fong, and J. J. Wu (2021). Algorithmic Trading and Market Quality: International Evidence. *Journal of Financial and Quantitative Analysis 56*, 2659–2688.

Bondarenko, O. and D. Muravyev (2023). Market Return Around the Clock: A Puzzle. *Journal of Financial and Quantitative Analysis 58*, 939–967.

Brogaard, J. and C. Garriott (2019). High-Frequency Trading Competition. *Journal of Financial and Quantitative Analysis 54*, 1469–1497.

Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan (2015). Trading Fast and Slow: Colocation and Liquidity. *Review of Financial Studies 28*, 3407–3443.

Brolley, M. and D. A. Cimon (2020). Order-Flow Segmentation, Liquidity, and Price Discovery: The Role of Latency Delays. *Journal of Financial and Quantitative Analysis 55*, 2555–2587.

Budish, E., P. Cramton, and J. Shim (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics 130*, 1547–1621.

CME (2016). CME cattle market volatility. Available at: https://www.cmegroup.com/trading/agricultural/files/cattle-volatility-faq.pdf.

Coughlan, J. and A. G. Orlov (2023). High-frequency trading and market quality: Evidence from account-level futures data. *Journal of Futures Markets 43*, 1126–1160.

Couleau, A., T. Serra, and P. Garcia (2019). Microstructure Noise and Realized Variance in the Live Cattle Futures Market. *American Journal of Agricultural Economics 101*, 563–578.

Couleau, A., T. Serra, and P. Garcia (2020). Are Corn Futures Prices Getting "Jumpy"? *American Journal of Agricultural Economics 102*, 569–588.

Danieli, O., D. Nevo, I. Walk, B. Weinstein, and D. Zeltzer (2026). Negative Control Falsification Tests for Instrumental Variable Designs. *American Economic Review 116*, 1380–1414.

Easley, D., M. L. de Prado, and M. O'Hara (2016). Discerning information from trade data. *Journal of Financial Economics 120*(2), 269–285.

Fishe, R. P., R. Haynes, and E. Onur (2019). Anticipatory Traders and Trading Speed. *Journal of Financial and Quantitative Analysis 54*, 729–758.

Fleming, M., B. Mizrach, and G. Nguyen (2018). The microstructure of a U.S. Treasury ECN: The BrokerTec platform. *Journal of Financial Markets 40*, 2–22.

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

Hasbrouck, J. and G. Saar (2013). Low-latency trading. *Journal of Financial Markets 16*, 646–679.

He, X., T. Serra, and P. Garcia (2021). Resilience in "Flash Events" in the Corn and Lean Hog Futures Markets. *American Journal of Agricultural Economics 103*, 743–764.

Hendershott, T., C. M. Jones, and A. J. Menkveld (2011). Does algorithmic trading improve liquidity? *Journal of Finance 66*, 1–33.

Hu, Z., T. Serra, and P. Garcia (2020). Algorithmic quoting, trading, and market quality in agricultural commodity futures markets. *Applied Economics 52*, 6277–6291.

Huang, J., T. Serra, and P. Garcia (2022). "Fast money" around Federal Statistics Releases. *American Journal of Agricultural Economics 105*, 1248–1266.

Kleibergen, F. and R. Paap (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics 133*, 97–126.

Kyle, A. S. (1985). Continuous Auctions and Insider Trading. *Econometrica 53*, 1315–1335.

Lehecka, G. V., X. Wang, and P. Garcia (2014). Gone in ten minutes: Intraday evidence of announcement effects in the electronic corn futures market. *Applied Economic Perspectives and Policy 36*, 504–526.

Li, S., X. Wang, and M. Ye (2021). Who provides liquidity, and when? *Journal of Financial Economics 141*, 968–980.

Ma, R. R. and T. Serra (2025). Is liquidity provision informative? Evidence from agricultural futures markets. *American Journal of Agricultural Economics 107*, 125–151.

Menkveld, A. J. (2016). The Economics of High-Frequency Trading: Taking Stock. *Annual Review of Financial Economics 8*, 1–24.

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

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics 116*, 257–270.

Pagnotta, E. S. and T. Philippon (2018). Competing on Speed. *Econometrica 86*, 1067–1115.

Peng, K., Z. Hu, and M. A. Robe (2024). Maximum order size and market quality: Evidence from a natural experiment in commodity futures markets. *Journal of Futures Markets 44*, 803–825.

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

Shkilko, A. and K. Sokolov (2020). Every Cloud Has a Silver Lining: Fast Trading, Microwave Connectivity, and Trading Costs. *Journal of Finance 75*(6), 2899–2927.

Skjeltorp, J. A., E. Sojli, and W. W. Tham (2016). Flashes of Trading Intent at NASDAQ. *Journal of Financial and Quantitative Analysis 51*, 165–196.

Wang, X., P. Garcia, and S. H. Irwin (2014). The behavior of bid-ask spreads in the electronically-traded corn futures market. *American Journal of Agricultural Economics 96*, 557–577.

Yueshen, B. Z. (2026). Queuing Uncertainty of Limit Orders. *Management Science 72*, 4760–4779.

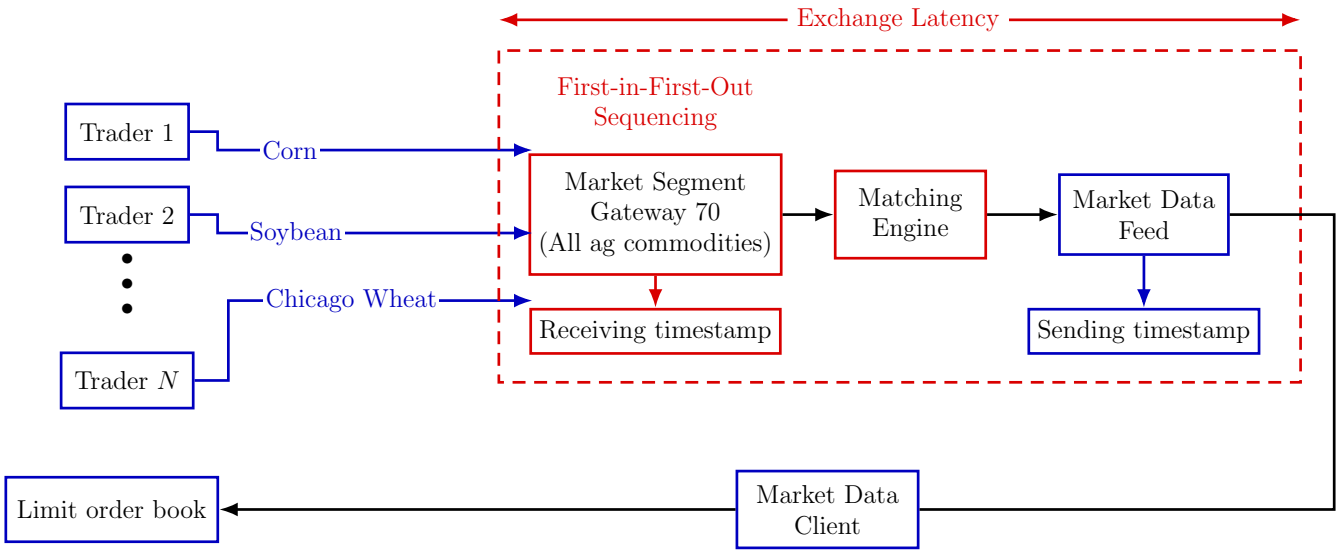


Figure 1: CME message processing architecture.

This figure displays the CME message processing architecture. All agricultural commodity messages need to arrive at a uniform Market Segment Gateway 70 where these messages are processed based on the first-in-first-out sequencing. The matching engine then processes all messages, including both limit orders and market orders, at which point they become effective. The market data feed disseminates them to the clients, allowing them to reconstruct limit order books. The exchange latency refers to the time eclipsed between market segment gateway and market data platform. We define the message latency as the time difference between receiving timestamp and sending timestamp.

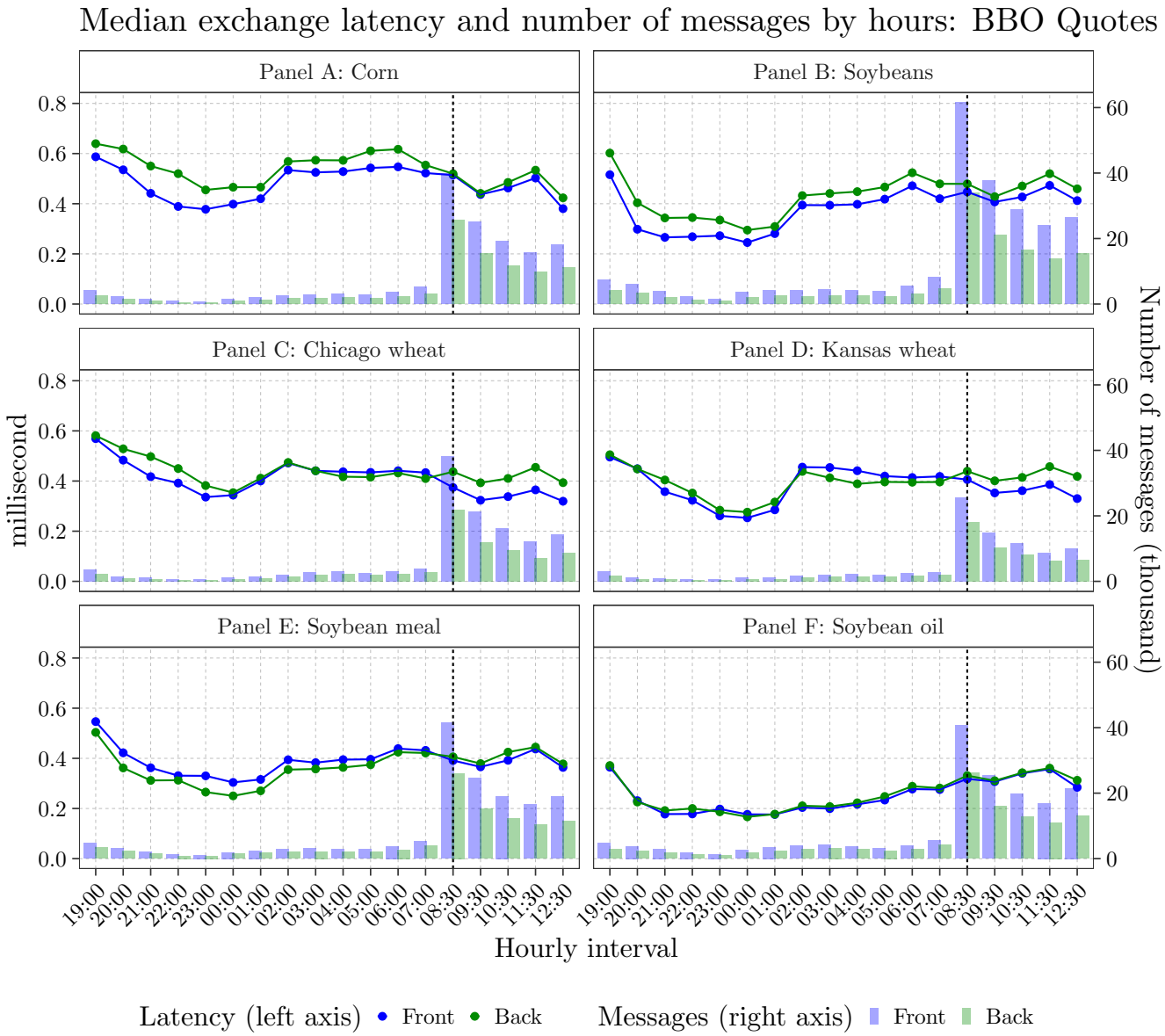


Figure 2: Hourly BBO quote message latency.

This figure displays the hourly median BBO quote message latency and the number of quote messages at the best-bid-offer (BBO) for both front-month and back-month contracts in the CME corn, soybean, Chicago wheat, Kansas wheat, soybean meal, and soybean oil markets. Message latency is defined as the time difference between receiving timestamp and sending timestamp, measured in milliseconds. The BBO quotes include limit order submissions, revisions, and deletions at the BBO. Our sample period spans from January 8, 2024 to December 27, 2024.



Figure 3: Hourly trade message latency.

This figure displays the hourly median trade message latency and the number of trade messages for both front-month and back-month contracts in the CME corn, soybean, Chicago wheat, Kansas wheat, soybean meal, and soybean oil markets. Message latency is defined as the time difference between receiving timestamp and sending timestamp, measured in milliseconds. Our sample period spans from January 8, 2024 to December 27, 2024.

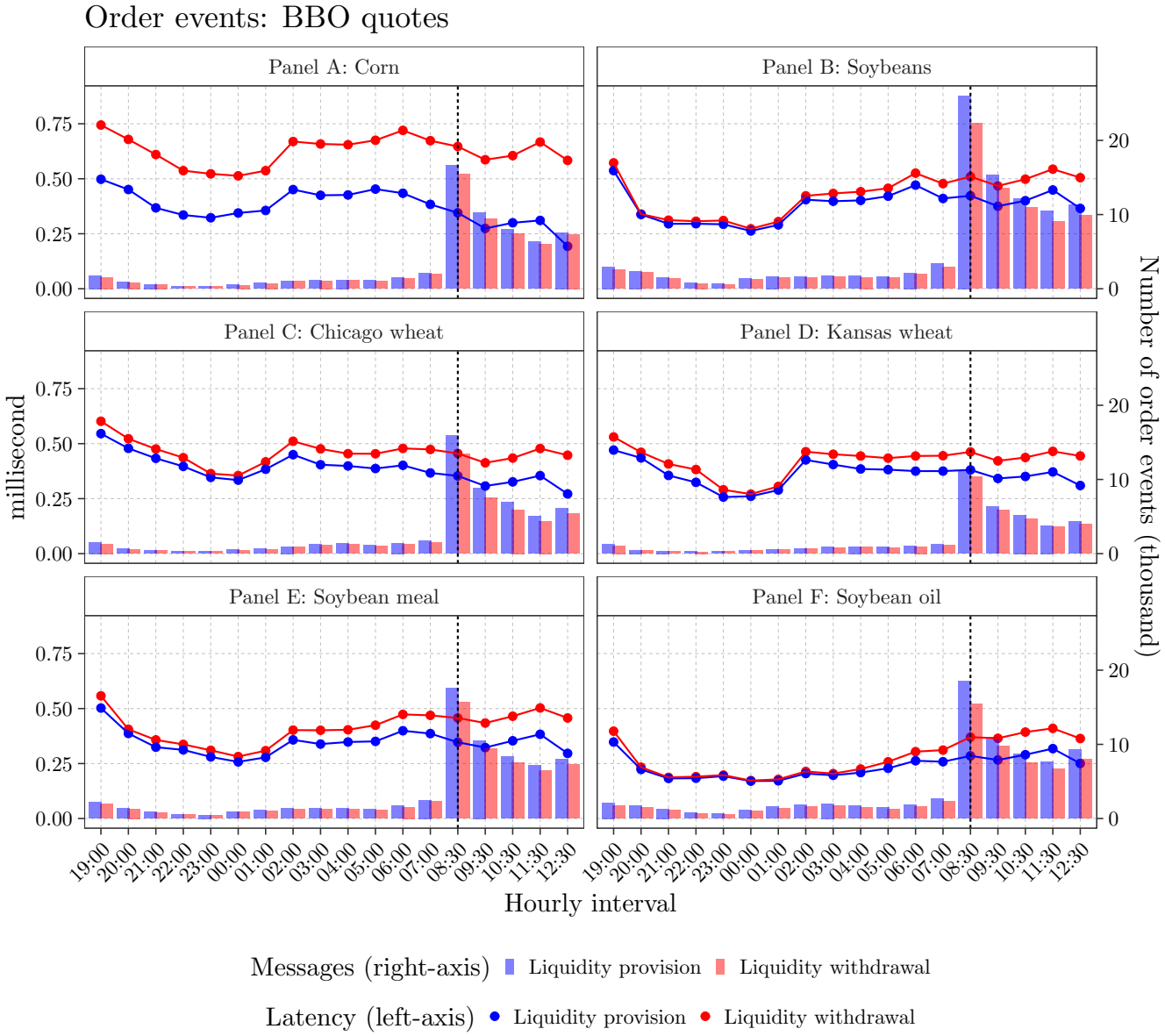


Figure 4: Hourly BBO quote message latency: Liquidity provision and liquidity withdrawal. This figure displays the hourly median BBO message latency and the number of BBO quote messages by two order events, liquidity provision and liquidity withdrawal for both front-month and back-month contracts in the CME corn, soybean, Chicago wheat, Kansas wheat, soybean meal, and soybean oil markets. Message latency is defined as the time difference between receiving timestamp and sending timestamp, measured in milliseconds. The liquidity provision messages refer to submissions and aggressive revisions (quoted quantity increase at the BBO and better quoted prices to the BBO), while liquidity withdrawal messages refer to deletions and less aggressive revisions (quoted quantity decrease at the BBO and worse quoted prices from the BBO). Our sample period spans from January 8, 2024 to December 27, 2024.

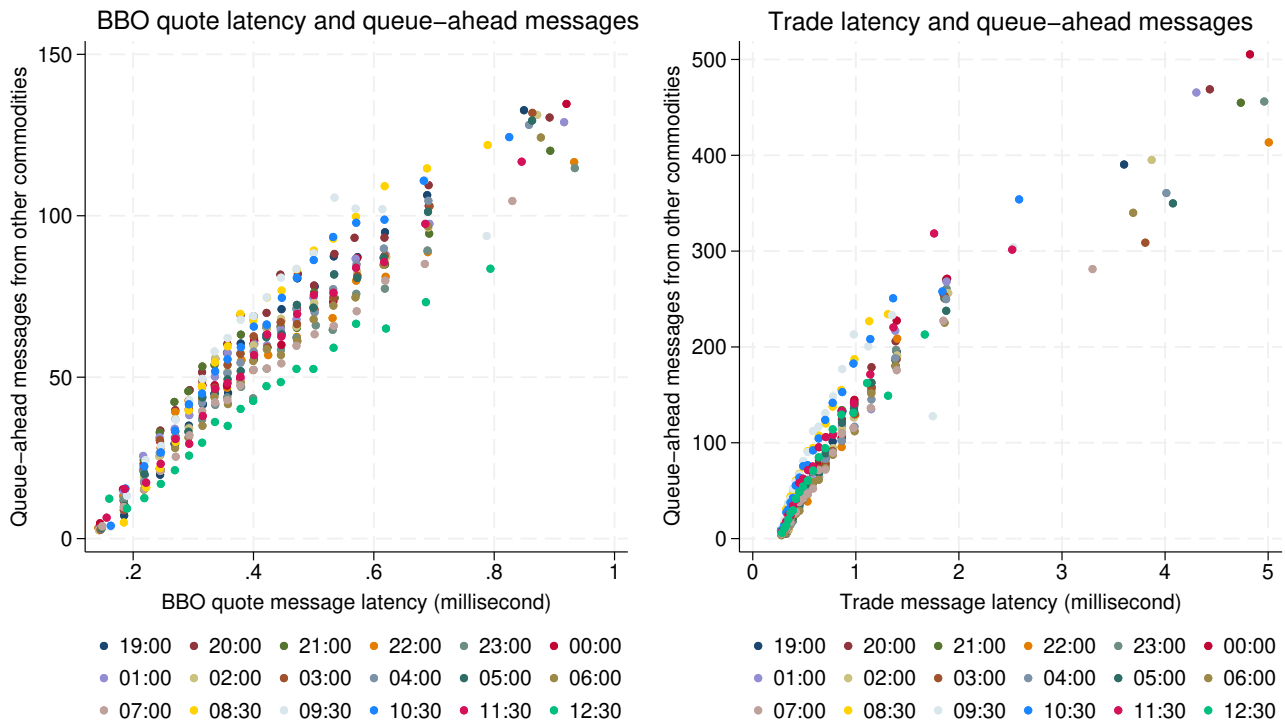


Figure 5: Exchange latency and messages ahead: BBO quotes and trades.

This figure displays the binscatter plots between the median exchange latency and the number of queue-ahead messages from other commodities for BBO quotes messages (left panel) and trade messages (right panel) of each commodity by hourly intervals. The queue-ahead messages refer to those positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform. Our sample period spans from January 8, 2024 to December 27, 2024.

Table 1: Variable descriptions.

Variable	Description
Effective spread (bps)	Trading-volume-weighted effective spread, calculated for an hour and expressed in basis points. It measures the immediate trading cost faced by market takers and is defined as $\frac{2q_t(p_t-m_t)}{m_t}$, where p_t is the trade price, m_t is the prevailing midpoint price at the time of the trade, and q_t equals $+1$ (-1) for buyer- (seller-)initiated trades.
Realized spread (bps)	Trading-volume-weighted realized spread, calculated for an hour and expressed in basis points. It measures gross market-making revenue and is defined as $\frac{2q_t(p_t-m_{t+5s})}{m_t}$, where p_t is the trade price, m_{t+5s} is the midpoint price 5 seconds after the trade, and q_t equals $+1$ (-1) for buyer- (seller-)initiated trades.
Price impact (bps)	Trading-volume-weighted price impact, calculated for an hour and expressed in basis points. It captures adverse selection costs and is defined as $\frac{2q_t(m_{t+5s}-m_t)}{m_t}$, where m_t is the prevailing midpoint price at the time of the trade, m_{t+5s} is the midpoint price 5 seconds after the trade, and q_t equals $+1$ (-1) for buyer- (seller-)initiated trades.
BBO latency (milliseconds)	Hourly median quote message latency at the best-bid-offer (BBO), expressed in milliseconds.
Trade latency (milliseconds)	Hourly median trade message latency, expressed in milliseconds.
Volatility	Realized volatility based on 1-second log midpoint returns for each hour, calculated as $\sqrt{\sum_{t=1}^{N}(\log m_t - \log m_{t-1})^2}$, and annualized by factor $\sqrt{252}$.
log(trade size)	Hourly average dollar trade size, expressed in logs.
Inverse price	The reciprocal of last trade price of each hour.
log(BBO msgs)	Hourly total number of quote messages at the BBO, expressed in logs.
log(trade msgs)	Hourly total number of trade messaged, expressed in logs.

Table 2: Summary statistics.

This table reports summary statistics for the pooled regression samples in agricultural futures markets. The dependent variables are effective spread, realized spread, and price impact. The main independent variable is BBO latency and trade latency, measured in milliseconds. Control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. Variable definitions are provided in Table 1. To mitigate the influence of outliers, we winsorize all variables at the 1st and 99th percentiles by individual contracts and hourly interval. Our sample period spans January 8, 2024, to December 27, 2024.

	Mean	Std	P25	Median	P75	N
Dependent variables						
Effective spread (bps)	4.092	1.429	2.909	4.057	4.955	52,124
Realized spread (bps)	0.196	1.801	−0.486	0.198	0.934	52,124
Price impact (bps)	3.896	2.095	2.564	3.537	4.868	52,124
Independent variables						
BBO latency	0.414	0.181	0.281	0.389	0.516	52,124
Trade latency	0.862	1.014	0.378	0.558	0.917	52,124
Control variables						
log(trade size)	11.132	0.581	10.685	11.003	11.557	52,124
Volatility	0.056	0.032	0.034	0.049	0.071	52,124
Inverse price	0.005	0.008	0.002	0.002	0.003	52,124
log(BBO msgs)	8.204	1.201	7.372	8.044	9.099	52,124
log(trade msgs)	4.951	1.716	3.784	5.004	6.146	52,124

Table 3: Market quality and exchange latency: 2SLS–IV regression.

This table reports the results of the 2SLS–IV regression of market liquidity on exchange latency. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression

$$Exchange\ Latency_{i,m,h,t} = \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where the instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
$\widehat{\text{BBO latency}}$	−0.049	0.039	−0.906***	−0.641***	0.850***	0.672***
	(0.109)	(0.108)	(0.144)	(0.128)	(0.187)	(0.171)
$\widehat{\text{Trade latency}}$		−0.076***		−0.257***		0.181***
		(0.016)		(0.052)		(0.046)
log(trade size)	0.542***	0.626***	−0.685***	−0.653***	1.233***	1.284***
	(0.083)	(0.080)	(0.122)	(0.129)	(0.124)	(0.143)
Volatility	8.302***	8.825***	−22.545***	−22.266***	30.764***	31.002***
	(1.461)	(1.467)	(2.087)	(2.101)	(3.091)	(3.113)
Inverse price	86.284*	100.108**	32.307	28.794	54.088	71.225
	(50.031)	(37.799)	(57.395)	(67.281)	(53.422)	(63.510)
log(BBO msgs)	−0.240***	−0.036	−0.035	−0.149***	−0.200**	0.113*
	(0.049)	(0.046)	(0.071)	(0.056)	(0.090)	(0.070)
log(trade msgs)		−0.214***		0.075		−0.286***
		(0.032)		(0.055)		(0.061)
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	779.80	640.60	779.80	640.60	779.80	640.60
N	52124	52124	52124	52124	52124	52124
Adj. R^2	0.046	0.062	0.030	0.038	0.062	0.077

Table 4: Market quality and exchange latency: Order events.

This table reports the results of the 2SLS–IV regression of market liquidity on exchange latency. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We further classify BBO quote message latency into BBO liquidity withdrawal latency and BBO liquidity provision latency. The liquidity provision messages refer to submissions and aggressive revisions (quoted quantity increase at the BBO and better quoted prices to the BBO), while liquidity withdrawal messages refer to deletions and less aggressive revisions (quoted quantity decrease at the BBO and worse quoted prices from the BBO). The regression specification is

$$\begin{aligned} MQ_{i,m,h,t} =& \beta_1 \times \widehat{\text{Withdrawal latency}}_{i,m,h,t} + \beta_2 \times \widehat{\text{Provision latency}}_{i,m,h,t} + \\ & \beta_3 \times \widehat{\text{Trade latency}}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t}, \end{aligned}$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. Our control variables include log trade size, inverse price, realized volatility, log BBO liquidity provision messages, log BBO liquidity withdrawal messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour $\times$ date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
$\widehat{\text{Withdrawal latency}}$	−0.383*	−0.367*	−2.017***	−1.877***	1.649***	1.525***
	(0.203)	(0.187)	(0.388)	(0.372)	(0.381)	(0.343)
$\widehat{\text{Provision latency}}$	0.311	0.436**	0.998***	1.101***	−0.706**	−0.686**
	(0.192)	(0.164)	(0.331)	(0.321)	(0.349)	(0.314)
$\widehat{\text{Trade latency}}$		−0.084***		−0.257***		0.173***
		(0.015)		(0.054)		(0.048)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	318.91	235.95	318.91	235.95	318.91	235.95
N	52124	52124	52124	52124	52124	52124
Adj. R^2	0.045	0.066	0.019	0.027	0.054	0.071

Online Appendices

A CME institutional details

A.1 CME Globex sessions and trading hours

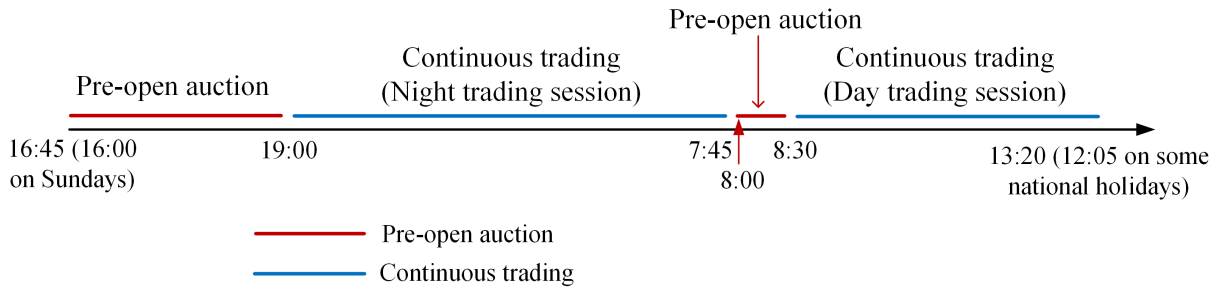


Figure A1: CME Globex sessions and trading hours.

This figure displays the CME Globex sessions and hours over a trading day in U.S. Central Time (CT). The pre-open auction starts at 16:00 (16:45) on Sundays (weekdays). The day trading session is from 8:30 to 13:20 CT and the night session from 19:00 to 7:45 CT. Generally, in our sample markets, CME replaces the continuous trading sessions with extended pre-open auctions on national holidays and may also shorten the continuous trading hours on some specific national holidays. Details on the CME holidays calendar can be found at https://www.cmegroup.com/tools-information/holiday-calendar.html.

B Placebo tests

B.1 Lagged market liquidity and instrumental variables

Table B1: Market quality and exchange latency: First-stage regression.
This table reports the OLS placebo regression of lagged market liquidity on the instrumental variable (IV). The regression specification is

$$MQ_{i,m,h-1,t} = \rho \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where $MQ_{i,m,h-1,t}$ is the first-lag market quality. Our instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. We calculate the IVs of BBO quote message latency and trade message latency separately. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	lagged ES	lagged ES	lagged RS	lagged RS	lagged PI	lagged PI
Queueahead BBO	−0.0004	−0.0004	−0.0000	0.0000	−0.0004	−0.0005
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Queueahead Trade		0.0000		−0.0001		0.0001
		(0.000)		(0.000)		(0.000)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓

C 2SLS-IV regression

C.1 First-stage regression results

Table C1: Market quality and exchange latency: First-stage regression.
This table reports the results of the first-stage regression of exchange latency on the instrumental variable (IV). The regression specification is

$$Exchange\ Latency_{i,m,h,t} = \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where the instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. We calculate the IVs of BBO quote message latency and trade message latency separately. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)
	BBO latency	BBO latency	Trade latency
Queueahead BBO	0.0035***	0.0035***	0.0003
	(0.000)	(0.000)	(0.000)
Queueahead Trade		−0.0000**	0.0033***
		(0.000)	(0.000)
Controls	✓	✓	✓
Contract FE	✓	✓	✓
Commodity FE	✓	✓	✓
Hour × Date FE	✓	✓	✓
N	52124	52124	52124
Adj. R^2	0.822	0.825	0.462

C.2 Liquidity provision prior to trade execution

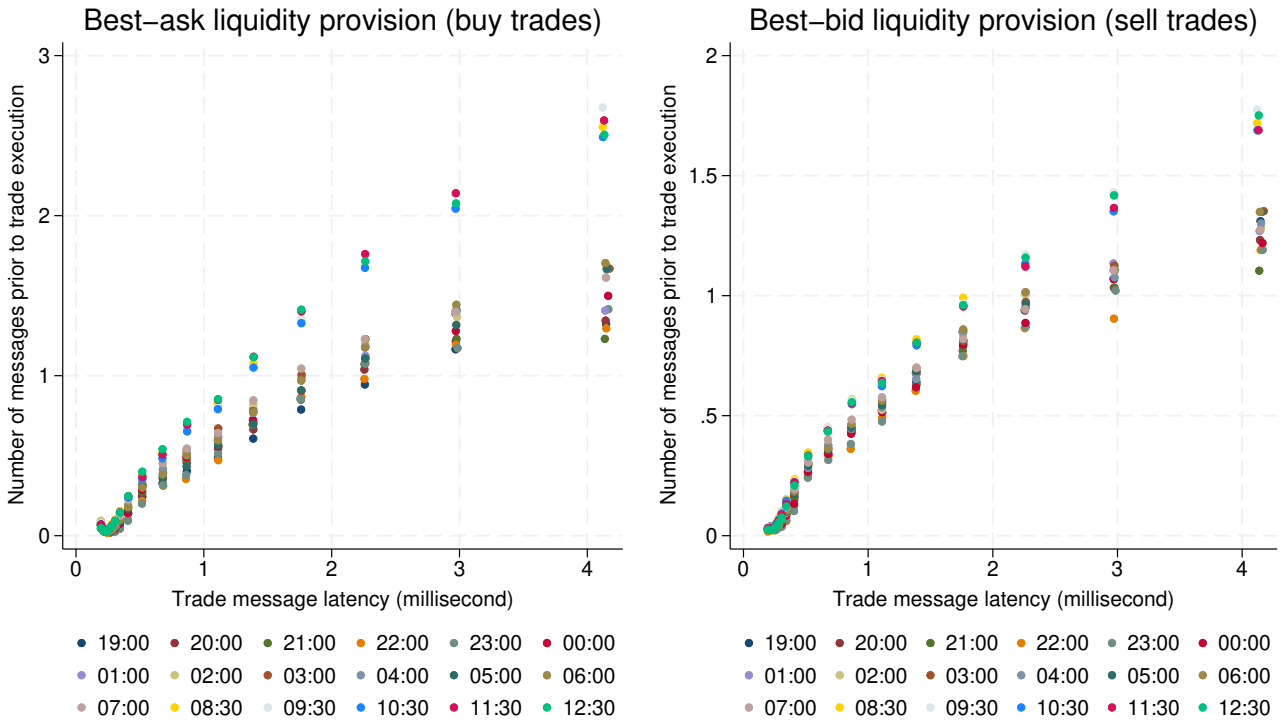


Figure C1: Liquidity provision prior to trade execution.

This figure displays the binscatter plots between trade latency and the number of BBO liquidity provision messages prior to trade execution for all buyer-initiated trades (left panel) and seller-initiated trades (right panel) in each futures market. Specifically, we calculate for each trade, the median number of BBO liquidity provision messages positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform. Our sample spans from January 8, 2024, to December 27, 2024.

D Announcement analysis

Table D1: Market liquidity and exchange latency: Announcements.

This table reports the results of the 2SLS–IV regression of market quality on exchange latency during USDA announcements using a 5-minute sample between 9:30 and 12:30 in corn, soybean, Chicago wheat, and Kansas wheat markets. We consider WASDE, Grain Stock, Prospective Plantings, Acreage, Grain Stocks, Crop Production announcements, yielding 15 announcement days in total. We select 5 days proceeding and following each announcement as our non-announcement counterparts. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta_1 \times \widehat{Exchange\ Latency}_{i,m,h,t} \times Ann_{h,t} + \beta_2 \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_t + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at 5-minute interval h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression. $Ann_{h,t}$ is an announcement dummy that equals 1 after 11:00 on announcement days, and 0 otherwise. Our control variables include the announcement dummy, log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_t denote contract, commodity, and date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
$\widehat{\text{BBO latency}}$	−0.107**		−0.772***		0.671***	
	(0.047)		(0.145)		(0.143)	
$\widehat{\text{BBO latency}} \times$ Ann	−0.156		2.155*		−2.316*	
	(0.335)		(1.081)		(1.129)	
$\widehat{\text{Trade latency}}$		−0.061***		−0.195***		0.136***
		(0.011)		(0.041)		(0.046)
$\widehat{\text{Trade latency}} \times$ Ann		−0.269***		−0.134		−0.134
		(0.060)		(0.278)		(0.273)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	424.42	554.25	424.42	554.25	424.42	554.25
N	45206	45206	45206	45206	45206	45206
Adj. R^2	0.189	0.177	0.069	0.077	0.151	0.158

E Robustness analysis

E.1 Robustness to non-announcement days

Table E1: Market quality and exchange latency: Robustness to non-announcement days. This table reports the results of the 2SLS–IV regression of market quality on exchange latency during non-announcement days. We consider WASDE, Prospective Plantings, Acreage, Grain Stocks, and Crop Production announcements, yielding 15 announcement days in total. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression

$$Exchange\ Latency_{i,m,h,t} = \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where the instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
$\widehat{\text{BBO latency}}$	−0.046	0.036	−0.892***	−0.624***	0.844***	0.658***
	(0.106)	(0.107)	(0.150)	(0.133)	(0.189)	(0.174)
$\widehat{\text{Trade latency}}$		−0.074***		−0.260***		0.186***
		(0.017)		(0.054)		(0.048)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	791.49	598.21	791.49	598.21	791.49	598.21
N	48922	48922	48922	48922	48922	48922
Adj. R^2	0.045	0.061	0.030	0.038	0.061	0.076

E.2 Robustness to excluding market open and close

Table E2: Market quality and exchange latency: Robustness to non-announcement days. This table reports the results of the 2SLS–IV regression of market quality on exchange latency excluding market open and close. We exclude three hourly intervals: 19:00–20:00, 8:30–9:30, and 12:30–13:20. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression

$$Exchange\ Latency_{i,m,h,t} = \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where the instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
$\widehat{\text{BBO latency}}$	−0.023	0.057	−0.882***	−0.609***	0.851***	0.657***
	(0.113)	(0.113)	(0.146)	(0.135)	(0.198)	(0.188)
$\widehat{\text{Trade latency}}$		−0.071***		−0.253***		0.183***
		(0.016)		(0.054)		(0.047)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	793.36	574.10	793.36	574.10	793.36	574.10
N	43404	43404	43404	43404	43404	43404
Adj. R^2	0.048	0.066	0.034	0.044	0.071	0.088

E.3 Robustness to 30-minute interval

Table E3: Market quality and exchange latency: Robustness to 30-minute interval. This table reports the results of the 2SLS–IV regression of market quality on exchange latency using a sample with 30-minute intervals. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at 30-minute interval h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression

$$Exchange\ Latency_{i,m,h,t} = \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where the instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and 30-minute interval. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and 30-minute × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
$\widehat{\text{BBO latency}}$	−0.110	−0.037	−0.804***	−0.525***	0.686***	0.480***
	(0.078)	(0.080)	(0.115)	(0.113)	(0.140)	(0.129)
$\widehat{\text{Trade latency}}$		−0.061***		−0.245***		0.184***
		(0.013)		(0.034)		(0.031)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
30-minute × Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	942.20	988.291	942.55	964.53	942.55	964.53
N	100021	100021	100005	100005	100005	100005
Adj. R^2	0.038	0.054	0.036	0.044	0.069	0.083

E.4 Robustness to different time horizons

Table E4: Market quality and exchange latency: Robustness to time horizons.
This table reports the results of the 2SLS–IV regression of market quality on exchange latency for different trade horizons when calculating realized spread and price impact. We consider realized spreads and price impacts at 1 second, 10 seconds, and 30 seconds after the time of trade execution. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression

$$Exchange\ Latency_{i,m,h,t} = \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where the instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	RS $(t+1s)$	RS $(t+10s)$	RS $(t+30s)$	PI $(t+1s)$	PI $(t+5s)$	RS $(t+30s)$
$\widehat{\text{BBO latency}}$	−0.387**	−0.680***	−0.695***	0.442**	0.735***	0.742***
	(0.154)	(0.137)	(0.168)	(0.182)	(0.182)	(0.229)
$\widehat{\text{Trade latency}}$	−0.314***	−0.272***	−0.187***	0.234***	0.191***	0.108**
	(0.052)	(0.053)	(0.055)	(0.046)	(0.047)	(0.049)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	640.60	637.36	641.47	640.60	637.36	641.47
N	52124	52123	52120	52124	52123	52120
Adj. R^2	0.028	0.037	0.028	0.073	0.074	0.055

F Heterogeneity analysis

F.1 Corn and soybean markets

Table F1: Market quality and exchange latency: Corn and soybean markets.
This table reports the results of the 2SLS–IV regression of market quality on exchange latency in corn and soybean markets. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression

$$Exchange\ Latency_{i,m,h,t} = \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where the instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
$\widehat{\text{BBO latency}}$	0.187**	0.237**	−1.743***	−1.419***	1.915***	1.641***
	(0.081)	(0.080)	(0.250)	(0.233)	(0.244)	(0.225)
$\widehat{\text{Trade latency}}$		−0.047		−0.308***		0.262**
		(0.026)		(0.087)		(0.095)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	978.73	149.45	978.73	149.45	978.73	149.45
N	17483	17483	17483	17483	17483	17483
Adj. R^2	0.142	0.142	0.040	0.042	0.108	0.108

F.2 Wheat markets

Table F2: Market quality and exchange latency: Wheat markets.
This table reports the results of the 2SLS–IV regression of market quality on exchange latency in Chicago wheat and Kansas wheat markets. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression

$$Exchange\ Latency_{i,m,h,t} = \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where the instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
$\widehat{\text{BBO latency}}$	−0.288	−0.320	−1.187***	−0.887**	0.896**	0.561
	(0.251)	(0.260)	(0.345)	(0.377)	(0.354)	(0.416)
$\widehat{\text{Trade latency}}$		−0.107***		−0.278***		0.172**
		(0.021)		(0.076)		(0.065)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	1692.79	229.24	1692.79	229.24	1692.79	229.24
N	17322	17322	17322	17322	17322	17322
Adj. R^2	0.030	0.053	0.018	0.031	0.032	0.053

F.3 Soybean meal and soybean oil markets

Table F3: Market quality and exchange latency: Soybean meal and soybean oil markets. This table reports the results of the 2SLS–IV regression of market quality on exchange latency in soybean meal and soybean oil markets. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression

$$Exchange\ Latency_{i,m,h,t} = \alpha \times Queueahead_{i,-m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + e_{i,m,h,t},$$

where the instrumental variable is the median number of messages from other commodities positioned ahead of a given message in the queue after it enters the market segment gateway and before it exits the market data platform, denoted as $Queueahead_{i,-m,h,t}$. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
$\widehat{\text{BBO latency}}$	0.340*	0.365**	−0.592***	−0.425**	0.920***	0.777**
	(0.180)	(0.155)	(0.181)	(0.184)	(0.279)	(0.268)
$\widehat{\text{Trade latency}}$		−0.084***		−0.143***		0.062
		(0.025)		(0.046)		(0.054)
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓
K-P F-stat.	662.40	484.45	662.40	484.45	662.40	484.45
N	17319	17319	17319	17319	17319	17319
Adj. R^2	0.052	0.087	0.066	0.067	0.095	0.106

F.4 Front-month and back-month contracts

Table F4: Market quality and exchange latency: Front-month and back-month contracts. This table reports the results of the 2SLS–IV regression of market quality on exchange latency for front-month (Panel A) and back-month (Panel B) contracts. Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression. Our control variables include log trade size, inverse price, realized volatility, log BBO messages, and log trade messages. We winsorize all variables at the 1st and 99th percentiles by contract and hour. Detailed variable definitions are shown in Table 1. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour × date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)
	ES	ES	RS	RS	PI	PI
Panel A: Front-month contract.						
$\widehat{\text{BBO latency}}$	0.044	0.128	−0.950***	−0.472**	0.985***	0.602**
	(0.139)	(0.144)	(0.166)	(0.177)	(0.250)	(0.244)
$\widehat{\text{Trade latency}}$		−0.149***		−0.257***		0.101
		(0.030)		(0.069)		(0.065)
K-P F-stat.	476.80	150.24	476.80	150.24	476.80	150.24
N	26237	26237	26237	26237	26237	26237
Adj. R^2	0.064	0.071	0.052	0.065	0.106	0.125
Panel B: Back-month contract.						
$\widehat{\text{BBO latency}}$	0.117	0.231**	−1.108***	−0.781***	1.230***	1.017***
	(0.097)	(0.095)	(0.216)	(0.224)	(0.205)	(0.207)
$\widehat{\text{Trade latency}}$		−0.061***		−0.267***		0.206***
		(0.016)		(0.062)		(0.054)
K-P F-stat.	941.60	499.12	941.60	499.12	941.60	499.12
N	25887	25887	25887	25887	25887	25887
Adj. R^2	0.036	0.060	0.020	0.028	0.039	0.053
Controls	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓

F.5 Day and night sessions

Table F5: Market quality and exchange latency: Day and night sessions.

This table reports the results of the 2SLS–IV regression of market quality on exchange latency for the day session (Panel A) and the night session (Panel B). Regression results of effective spread (ES), realized spread (RS), and price impact (PI) are organized in columns. We consider both the BBO quote message latency and the trade message latency as our variables of interests. The regression specification is

$$MQ_{i,m,h,t} = \beta \times \widehat{Exchange\ Latency}_{i,m,h,t} + \boldsymbol{Controls}_{i,m,h,t} + \lambda_i + \gamma_m + \theta_{ht} + \varepsilon_{i,m,h,t},$$

where $MQ_{i,m,h,t}$ refer to the ES, RS, and PI of individual contract i in market m at hour h on day t. $\widehat{Exchange\ Latency}_{i,m,h,t}$ denotes the fitted value of either BBO quote message latency or trade message latency from the first-stage regression. λ_i, γ_m, and θ_{ht} denote contract, commodity, and hour $\times$ date fixed effects, respectively. Standard errors are double clustered by contract and date and are reported in parentheses. We also report the Kleibergen-Paap F statistic for the weak instrumental variable. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Our sample spans from January 8, 2024, to December 27, 2024.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	ES	ES	ES	RS	RS	RS	PI	PI	PI
Panel A: Day session (8:30-13:20)									
$\widehat{\text{BBO latency}}$	−0.202		0.086	−0.923***		−0.241	0.724*		0.264
	(0.209)		(0.194)	(0.287)		(0.255)	(0.399)		(0.346)
$\widehat{\text{Trade latency}}$		−0.240***	−0.253***		−0.371***	−0.328***		0.132	0.141
		(0.058)	(0.051)		(0.082)	(0.081)		(0.099)	(0.088)
K-P F-stat.	238.45	290.47	131.58	238.45	290.47	131.58	238.45	290.47	131.58
N	14543	14543	14543	14543	14543	14543	14543	14543	14543
Adj. R^2	0.091	0.086	0.086	0.070	0.090	0.089	0.119	0.134	0.134
Panel B: Night session (19:00-7:45)									
$\widehat{\text{BBO latency}}$	−0.153		0.026	−0.809***		−0.517***	0.653***		0.449**
	(0.107)		(0.106)	(0.156)		(0.166)	(0.199)		(0.218)
$\widehat{\text{Trade latency}}$		−0.085***	−0.085***		−0.265***	−0.251***		0.180***	0.166***
		(0.016)	(0.015)		(0.053)	(0.054)		(0.047)	(0.047)
K-P F-stat.	864.14	1114.68	520.08	864.14	1114.68	520.08	864.14	1114.68	520.08
N	37581	37581	37581	37581	37581	37581	37581	37581	37581
Adj. R^2	0.050	0.064	0.064	0.033	0.042	0.040	0.071	0.084	0.083
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Contract FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Commodity FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Hour × Date FE	✓	✓	✓	✓	✓	✓	✓	✓	✓