TABB FORUM

WHERE CAPITAL MARKETS SPEAK

*April 10, 2019*

*Matt Clarke*
XTX
| Profile | More

Share

**Clearly, certain market participants are willing to go to great lengths and considerable cost to gain the ultimate speed advantage. But there are occasions when latency arbitrage activity has a negative effect on outcomes for long-term investors, and, as such, it may be beneficial to curb this activity through implementation of thoughtfully designed anti-latency arbitrage mechanisms. XTX Markets' Matt Clarke looks at the types of ALA mechanisms in the market today and the arguments for and against them.**


Recently, there has been an increasing preoccupation with "speed traders" and the lengths to which some are going to establish and protect their relative speed advantages. Stories like the below from Bloomberg capture the imagination:

*"The Gazillion-Dollar Standoff Over Two High-Frequency Trading Towers: The hunt for a millionth-of-a-second advantage in the town best known for Wayne's World is getting heated."*

In fact, a major motion picture was recently released that has as its plot a race to shave a couple of milliseconds (thousandths of a second) off the time it takes to send an order message from a fictional exchange data center in Kansas City to the NYSE: *www.thehummingbirdproject.film*.

Clearly, certain market participants are willing to go to great lengths and considerable cost to gain the ultimate speed advantage. How valuable is the advantage to them, and what is the impact on and cost to other market participants?

These stories beg the question:

*Isn't there room in market structure for venues that do not benefit those willing or able to achieve latency advantages? And might marketplaces and market participants be better off with a speedbump or any other mechanism that seeks to reduce latency advantages across participants (an anti-latency arbitrage mechanism)?*

Most market structure arguments – like this one – are generally informed by the author's frame of reference.

We believe the purpose of markets is ultimately to serve end users – that is, long-term investors and their agents. It is from that perspective that we will approach this topic. Does a given market design feature holistically benefit or harm long-term investors?

Market makers and liquidity providers are important but only inasmuch as they provide a service to end users – i.e., providing immediacy of risk transfer and taking on time/fluctuation risk in return for a spread.

Arbitrageurs – who may or may not also make markets – can be helpful, as they perform the unexciting but useful service of aligning multiple venues. However, if they introduce negative externalities for end users while doing so, the needs of arbitrageurs should be considered as subordinate to those of long-term investors.

As we will explore below, there are occasions when latency arbitrage activity has a negative effect on outcomes for end users, and, as such, it may be beneficial to curb this activity through implementation of thoughtfully designed ALA mechanisms.

**The State of Latency in 2019**

There is an enduringly popular legend that the Rothschild family made a fortune by utilizing its private pigeon network to learn the outcome of the Battle of Waterloo and trading on this information. Although financial markets have become much more sophisticated and electronic since then, the competition among participants to receive information first remains just as fierce. There are, however, two important things to realize.

First, as Donald MacKenzie of the University of Edinburgh notes in his excellent TabbFORUM piece, "How Fragile Is Competition in High-Frequency Trading?" the exchange groups' considerable and successful focus on reducing "jitter" (quasi-random fluctuations in processing times) on their exchanges means "even

tiny speed advantages" have become incredibly important. Given that exchange jitter is now measured in single-digit nanoseconds (a nanosecond being a thousand-millionth of a second) it follows that *"in a particular market … one HFT firm – or a small number of firms – may achieve an advantage in speed that's very hard and very costly for their rivals to overcome."*

Second, this is not about open competition. In order to obtain that last half-turn of the screw of latency advantage, the latest technologies being considered include such offerings as low-earth orbit satellite constellations, some of which propose to use free space optics – yes, space lasers – and autonomous solar-powered drones that could repeat signals from one data center to another over water. It is crucial to note that the aim of these efforts to reduce latency is not the promotion of open and free competition but rather to exploit scarcity for relative advantage.
This pattern – latency decreasing, cost increasing, scarcity increasing – is observed repeatedly. When ICE released roof space in its Basildon data center, there were a total of 32 slots – of which a small fraction were optimal for relaying signals to the LSE and trading venues with servers in Slough. Once these advantageous slots have been reserved, the market participants that have access to these resources enjoy an insuperable latency advantage over everyone else. Similarly, there are "optimal" frequencies for microwave channels and, once these are licensed, they are unavailable to other participants. This theme is echoed along the path of the microwave circuits, where access to specific rooftops and frequencies are fiercely protected and exclusively maintained assets.

The aim is to find, obtain and protect a systemic advantage: to find the rooftop that you can lease so no one else can; the intellectual property that you can exclusively license; to register the frequency that prevents others taking the same path; and so on. The costs are no longer the same as training pigeons.
Furthermore, the cost of this private infrastructure is ultimately paid for by real-world end users.

**What Kinds of ALA Mechanisms Exist Today?**

The purpose of ALA mechanisms is to prevent latency arbitrage by levelling latency across all participants so everyone can trade and compete with equal access to timely market data. There are two kinds of ALA mechanisms: technology-based and policy-based.

***Technology-based ALA mechanisms***

These diverse implementations include "latency floors" or "speed bumps." The precise implementation will differ across venues to reflect differences in products, rule books, regulatory regimes and proxy markets. Implementation details must take all these factors into account and are crucial in ensuring a well-designed ALA mechanism.

One can consider an illustrative symmetrical speed bump implementation. Incoming orders are subject to a speed bump of typically several milliseconds before being eligible for a match. In some designs the length of the speed bump may be randomized.

How does this prevent latency arbitrage?

Imagine that a related instrument jumps in price in a different market center; both the market maker (passive order) and latency arbitrageur (aggressive order) observe this at the same time, but the participant

seeking to latency arbitrage has a two-millisecond speed advantage in sending this information over to the speed-bumped venue due to its private microwave network. Instead of being able to pick off the stale offer immediately, it must traverse the three-millisecond speed bump, which affords liquidity providers a level playing field, as they can incorporate the same information into their pricing and cancel the stale offer before it matches and is picked off. Both passive and aggressive incoming orders are subject to the speed bump and latency has been floored.

Other examples of technology-based ALA mechanisms include those that impose a few millisecond delay on incoming orders to remove liquidity, thereby giving market makers the opportunity to react to new information and cancel stale orders.

***Policy-based ALA mechanisms***

A good example of a policy-based ALA mechanism can be found in Aquis, a pan-European equities exchange. Aquis is publicly traded, and in full disclosure, XTX Markets owns a non-controlling minority stake, an investment that was made because we believed it was a positive example of market structure that is good for end users and would therefore prove popular over time.

As stated on its website, Aquis does not permit "aggressive non-client proprietary trading." Only order flow deriving from natural buy-side exposures is eligible to remove liquidity from the platform. High-frequency trading firms may supply liquidity to the platform, but they cannot take liquidity from this market. As a result, the venue's market makers (and end users leaving pegged orders) may be able to offer tighter spreads and/or larger size to this natural buy-side platform because they know they will not be latency arbitraged by participants with a systematic speed advantage.

If market makers are instead forced to quote blindly into an orderbook whose incoming orders might originate from end users but might equally be latency arbitrage, it follows that market makers will quote wider and in smaller size. After all, they must quote to their average experience on the venue – one participant may be subsidizing the activity of another.

This theory is intuitive; but has it proved successful in practice? Based on analysis by Liquidmetrix, again on its website, Aquis believes its ALA mechanism policy has resulted in "lower toxicity and signalling risk than other trading venues in Europe." It has certainly proved popular with the buy side, as the venue's rapid and sustained market share growth demonstrates. It has grown from 1% to 4.5% of pan-European market share over the past three years.

**What Are the Arguments in Favor of ALA Mechanisms?**

We have now examined the role of latency in modern markets and several methods by which latency can be levelled across participants. Now we must ask: Is this a desirable aim? Is the effect of latency arbitrage positive or negative for long-term investors? In this section we'll examine the core arguments in favor of ALA mechanisms.

1. ***They reduce the indirect operational tax on end users of markets.***

If raw speed is the determining factor, any liquidity provider that is systematically outpaced will consistently get picked off, as the fastest arbitrageurs observe quotes moving on one venue and race to hit quotes on another venue a few milliseconds before the liquidity provider receives the same market data and can react. The end result is that liquidity providers may be forced into an expensive arms race.

This is a classic prisoner's dilemma wherein participants are commercially obliged to participate in a negative-sum activity due to the participation of others. Liquidity providers are not charities and the significant operational expenditure incurred in becoming or remaining low latency – always relative to other participants and therefore relevant even at increasingly diminishing timescales – is ultimately passed on to long-term investors. The transmission mechanism for this is typically as follows:

2. ***They will lead to tighter pricing and deeper books for end users.***

End users are hedging genuine exposures or making long-term investments and not reacting to millisecond-level external events, unlike latency arbitrageurs. Any market maker on an all-to-all exchange has no idea with whom it will trade; it gets a mix of latency arbitrageur flow and regular end-user flow.

The end-user flow is thus subsidizing the latency arbitrageur flow because the spreads charged on a venue are determined by the average quality of flow on the venue. An ALA mechanism normalizes market data transport across all participants: If a market ticks in Chicago and a latency arbitrageur is able to ship that data over to New York (before you can), the speed bump will give a liquidity provider the opportunity to see and incorporate that tick before the latency arbitrageur can pick off its stale price.

ALA mechanisms make it harder for latency arbitrageur taking strategies to perform latency arbitrage on liquidity providers and thus encourages market makers to quote tighter and in larger size to compete for and attract more flow from end users whose orders stem from genuine economic exposures rather than intermarket races.

3. ***They reduce barriers to entry and encourage competition.***

If raw speed is a prerequisite for success in liquidity provision, any participants – including new entrants, which cannot afford such expensive infrastructure – cannot compete and will logically withdraw. We've known this since the 1970s.

This is detrimental, as such liquidity providers may well have risk absorption appetite, as well as unique pricing and time horizons. Removing these resting limit orders from the market entirely (because they systematically get picked off each time a related market moves) reduces valuable liquidity.

Regional banks in FX are a good example: They are an important source of liquidity to the interbank market because they perform hedging on behalf of corporate end users that have natural resting interest at certain price levels.

4. ***They increase diversity and reduce systemic risk.***

Latency-sensitive markets tend to have heavily concentrated market share among a small group of extremely fast participants.This leads to systemic risk, as a small number of HFT firms have limited risk

absorption capabilities in relation to their outsized market share, and the failure or operational interruption, even if brief, of such an entity would have a disproportionate adverse impact on the market and liquidity relative to its size.

Reducing the focus on minor speed advantages encourages more competition and a wider group of participants which will deepen the risk absorption capacity of the overall market.

**What Are the Arguments Against ALA Mechanisms?**

Not everyone is in favor of ALA mechanisms, so we should review the counter-arguments – again, through the prism of their effect on long-term investors. We have tried to compile an exhaustive list, engaging with the strongest form of each argument.

1. ***This is just like the 'last look window' in FX.***

This is an erroneous and disingenuous argument, typically made by certain participants who conflate two entirely different topics.

With last look, the liquidity provider knows about an incoming order, even if it is not ultimately filled, and can themselves choose whether to accept this order. This is highly problematic since it leaks information, and "last look" has a deserved bad reputation. For example, bad actors may engage in the practice of "pre-hedging," which is performed as follows:

With a speed bump, on the other hand, a neutral venue determines the match – the liquidity provider has no choice at all since its quotes are firm – and the liquidity provider of course has no knowledge of any orders that miss. Self-evidently, the information leakage associated with last look does not occur and harmful practices such as pre-hedging would remain impossible.

2. ***Any additional complexity is bad.***

All being equal, simpler is better since end users and their agents tend to react to complexity and change less efficiently than specialized high-frequency traders.

The proliferation of order types in US equities is a good example of complexity harming long-term investors. End users simply cannot devote whole teams to study each order type and are therefore disadvantaged when placing orders, relative to HFTs, some of which may even support the increased complexity as they are able to exploit more edge-case scenarios. It would be ironic for participants that have contributed to the proliferation of US equity orders to object to speed bumps on the grounds of complexity!

As a principle it is therefore entirely reasonable to aim for simplicity, but this must be considered alongside the benefits of innovation to long-term investors. It is worth noting that the existing effort of trying to measure latencies and jitter and optimally splitting orders across multiple venues is far more complex than any proposed ALA mechanism and that long-term investors appear extremely comfortable trading on venues with ALA mechanisms today.

3. ***ALA mechanisms could be used not only by liquidity providers but also by criminals engaged in spoofing.***

Spoofing is illegal and accordingly exchanges have robust methods for detecting and punishing such activity. Such behavior is already subject to criminal sanctions, which acts as a material deterrent.

One doesn't hear the argument that cars should not be available to the public because they may also be used as getaway vehicles by bank robbers: The car is clearly not the problem!

4. ***Taking is a form of liquidity provision and end users' passive orders could miss out on valuable fills from aggressive latency arbitrage orders.***

The only fills end users would miss out on due to an ALA mechanism are fills which instantly move adversely against them because they have just been latency arbitraged.

One can imagine a resting bid in a 10 / 12 market being filled in response to a related market crumbling to 6 / 8. Immediately post-fill, the end user's order looks to be off-market, having bought at 10 while the prevailing price is now 6 / 8. Had this end user 'missed out' on this fill due to an ALA mechanism, it would be better off as it can now buy immediately at 8.

Incidentally, this particular form of "liquidity provision" is very common: Multiple arbitrageurs will compete to pick off these orders at the same time.

5. ***Any delay whatsoever increases uncertainty and risk.***

Some participants argue that speed bumping the matching process on venues (even by a handful of milliseconds) is bad for the market as it hampers risk management; but this misses the point. That is absolutely true at extremes – imagine a market updating once an hour versus once per second – but current market structure has gone far, far beyond the point of diminishing returns.
If a market marker is concerned about an increase in risk holding times of milliseconds, it ultimately is acting as an arbitrageur rather than a liquidity provider that absorbs risk for a meaningful period by using its risk capital. Whenever an arbitrageur disappears, experience shows another will immediately pop up and perform the task – maybe a few microseconds later.

David Olsen, president of leading HFT firm Jump Trading, offered the following observation on the diminishing returns of trading speed in a recent interview:

*"Going from 80% efficient to 90% was pretty cheap and a fairly meaningful payoff, but going beyond 99.9% is incredibly expensive."*

In the same article, Robert Walker, CTO of another HFT firm, CMT Trading, expanded on the same point by highlighting the dilemma that such firms face due to today's market structure: invest heavily in nanosecond-level latency reductions or risk not being able to compete.

*"A lot of the tech I've been building in the past five years has been about saving half a microsecond, equivalent to 500 nanoseconds … That edge can be the difference between making money or trading everyone else's exhaust fumes. It's a winner-takes-all scenario."*

6. ***It slows down price discovery and/or creates an illusion of liquidity, which might lead to a lack of confidence in the accuracy and transparency of market prices.***

There will be no illusion of liquidity for end users: What they see is what they will continue to get. ALA mechanisms specifically target latency arbitrage, which no end user engages in when performing natural trading or hedging activity.

Price discovery would indeed be slowed down by several milliseconds. This would have no material effect on end users of the market, however, who tend to have long-run economic exposures in the order of days, weeks and months and whose trading or hedging activity is not motivated by market developments at the millisecond timescale.

Recall that we are talking about a quantum that is significantly less even than the time taken for light (and thus pricing data) to travel from, for example, a futures market in Chicago to an asset manager sitting at her desk in London.

7. ***Wouldn't the fill rate go down for buy-side clients, too?***

Natural liquidity consumers should expect to continue to experience high fill rates on markets with ALA mechanisms because their consumption of liquidity is not driven by millisecond-level external events, unlike latency arbitrageurs. Furthermore, they should expect tighter and deeper pricing.
Deeper pricing is extremely important because market structure is not static. In many markets such as equities, the buy side will outsource the routing of orders to broker Smart Order Routing systems (SORs). Because the displayed size is often very small on lit equities venues, the SORs are forced to send multiple orders to multiple venues simultaneously.

An ALA mechanism on a single venue may instead solve the underlying issue: The market makers may quote in sufficient size so that the SOR can fill its interest with one order on a single venue – preventing a latency arbitrageur observing one order and using its private microwave networks to rush to other venues and trade ahead of the others before they arrive.

8. ***ALA mechanisms are discriminatory.***

Some venues' rulebooks are indeed intentionally discriminatory: Think of buy side-to-buy side venues where HFTs cannot trade. There is a place for these business models and commercial demand will determine their success.

Certain exchanges, on the other hand, have obligations to ensure impartial and non-discriminatory access. This is entirely compatible with technology-based ALA mechanisms, which may be designed to ensure they operate impartially and without undue discrimination. On this topic it is worth noting two further points.

First, latency arbitrage is a behavior and not a type of participant. Certain participants may conduct more or less latency arbitrage – thus acting at times as latency arbitrageurs – but these participants are themselves diverse and cannot be defined or grouped by one aspect of their overall trading activity; indeed they do not even appear to self-define themselves as latency arbitrageurs and will typically flex their businesses and activities to accommodate the specific market structure of each product and market. Venues may

determine for themselves the value of certain forms of behavior within their market ecology and should be free to innovate to encourage more or less of it.

Second, there are several genuinely discriminatory practices in existence on markets today such as exchange market making schemes which may, for example, offer brokerage discounts of up to 90% but are designed to effectively apply to only a single liquidity provider.

9. ***It may contribute to market fragility and flash crashes.***

On the contrary.

An ALA mechanism is likely to encourage more resting orders into the market – since these orders are less likely to be adversely selected by latency arbitrageurs – providing a deeper market with greater price stability.

Similarly, increased diversity in liquidity providers is likely to increase the overall risk capital and absorption capabilities of a market. This diversity is the crucial point and this increased, more diverse liquidity should be perfectly accessible for long-term investors since they are not performing latency arbitrage.

***10. This would advantage a subset of highly sophisticated market makers but not the wider market.*** There are of course many benefits to end users, such as tighter pricing and more book depth, and those have been outlined in detail above.

Furthermore, the effect of an ALA mechanism is not to advantage a subset of highly sophisticated market makers. It has the exact opposite effect, since it reduces the gap between large and small liquidity providers. By lowering the barriers to entry for market makers it widens the possible pool of participants.

The current market structure rewards deep-pocketed and sophisticated market makers – i.e., those that can afford to spend tens of millions each year on microwave networks and can react rapidly to the market data that these networks transport. By levelling the playing field, small electronic trading firms and less technologically sophisticated yet well capitalized banks would also be able to compete as liquidity providers.

10. ***An ALA mechanism would provide its venue with a commercial advantage and may encourage other venues to react.***

This is a surprising argument in that it infers (correctly) that an ALA mechanism would result in better liquidity being available on a venue in the form of tighter pricing and a deeper book. It is true that clients may gravitate to the improved liquidity, but that is how innovation and competition are supposed to work!

There is nothing preventing multiple competing venues operating ALA mechanisms in an attempt to improve the trading experience for the end users of their venues.

**Why, Then, Don’t We See More Experimentation with ALA Mechanisms in Listed Products?**

Well, we do.

The concept is already popular in FX, with major venues such as Reuters Matching, EBS Market and ParFX all utilizing some form of tech-based ALA mechanism. Moscow Exchange has also announced that it will be adding an ALA mechanism to its Rouble market.

This issue is especially important in the highly fragmented, geographically diverse FX market. Prior to Reuters and EBS implementing their ALA mechanisms, liquidity provision was dominated by HFT firms.

Since implementing these protections, concentration among liquidity providers has substantially decreased, with multiple bank LPs re-entering the market. According to Alexis Atkinson, head of order-driven markets at EBS:

> *"Our latency floor mechanism has enabled a more diverse range of participants to make markets successfully. Enabling a wider range of participants to be successfully quoting builds a stronger market, improves price discovery and reduces maker concentration risk, which is particularly important in stressed market conditions."*

In European equities there is the previously mentioned policy-based ALA mechanism at Aquis. In North America there is TSX Alpha in Canada, which enjoyed remarkable improvements to its market quality as a result. The Bank of Canada found in its analysis that there was "no evidence Alpha harms overall market quality," while the execution size on Alpha became larger. The Investment Industry Regulatory Organization of Canada further notes that "execution size nearly doubles from 150 to 260 shares" and that, "following Alpha's redesign, both retail and buy-side heavy users have their orders filled more frequently, and order execution sizes increase."
Nasdaq launched a non-displayed order type named M-ELO which uses a form of ALA mechanism to incentivize and protect orders that rest in the book for longer and is enjoying healthy growth despite its 500-millisecond speed bump (such orders must wait half a second before being eligible to interact with other orders). According to the Wall Street Journal, CBOE is considering implementing an ALA mechanism on its EDGA market.

In commodities, the London Metal Exchange has announced its intention to implement a fixed delay to all orders other than cancellations for its precious metals futures, which is another form of ALA mechanism.

We are also seeing this innovation appear in futures markets. Eurex has deployed an ALA mechanism on its FX Futures market. Eurex has also announced its intention to implement an ALA mechanism on its German and French equity options venues. And ICE intends to launch an ALA mechanism on its precious metals contracts.

There is clearly an underlying problem to which these solutions are responding. Therefore, one might question why in fact we do not see ALA mechanisms on several of the remaining major markets.

From an exchange CEO's perspective, any change is risky – even when it will demonstrably benefit end users. If your top-five brokerage payers are speed-focused latency arbitrageurs and they all hate speed bumps (they would, as it would have the effect of levelling the playing field), it takes real confidence to implement one. If you're right, you've improved the business and end-user client experience; if you're wrong, you may lose your job!

This doesn't mean that exchanges haven't deeply considered adding speed bumps. See this patent, for example. This is because there are also great benefits to exchanges of speed bumps or latency floors.

Any business would much rather have its revenues diversified across many participants than have 50% come from the top 10, as this exposes them to less "key client" risk and allows them to develop products broadly without being beholden to a small special interest group.Furthermore, because the top participants typically receive meaningful volume discounts on brokerage, exchange brokerage revenues would likely increase if overall volumes were to remain constant but were diversified across many (individually smaller) liquidity providers.

Finally, this would improve conditions for end users of the exchange. Let's not forget them since they are, after all, the whole reason for markets existing.

**How Can Venues Proceed with ALA Mechanisms?**

Well, the easiest thing to do is to experiment, thoughtfully. We certainly agree with sentiments from a wide range of participants that market structure changes should be based on data collection, iterative experimentation and careful reflection.

Each venue and product has a different set of conditions (tick size, participant mix, regulations, proxy venues, etc.), so design decisions need to take these factors into account. In some cases – like one-tick markets, where the bid-ask spread is practically always at the minimum tick increment – there may be preparatory work required.
Determine a list of market/liquidity quality criteria and try adding a speed bump in a subset of products. Does the data indicate conditions have improved and holistic costs reduced for end users? How do activity levels change? Is activity more diversified?

If – and only if – it has the desired effects, continue to experiment more boldly and roll out across more products.

*Matt Clarke is Head of Distribution and Liquidity Management, EMEA, for XTX Markets, a leading quantitative-driven electronic market maker partnering with counterparties, exchanges and e-trading venues globally to provide liquidity in the Equity, FX, Fixed Income and Commodity markets.*