[XTX] MARKETS

ANTI-LATENCY ARBITRAGE MECHANISMS: AN OVERVIEW

[LATENCY ARBITRAGE: THE REAL COST]

Speed races have gone way beyond the point of diminishing returns

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

Given that certain futures exchanges today measure jitter (fluctuations in processing time) in single-digit nanoseconds, "even tiny speed advantages" have become incredibly important. If you get your order to exchange ten nanoseconds (ten billionths of a second) before the next trader you will always win the race so shaving off ever smaller increments of latency is economically rewarded.

Ultra High Frequency Traders have long sought to gain a speed advantage by creating geodesic (the shortest path) proprietary microwave networks, well documented by the excellent Sniper in Mahwah blog. Now the same firms are even looking at Low-Earth Orbit (LEO) satellites as a method by which to gain a speed advantage.

LONDON
FRANKFURT
CHICAGO

Of course this monumental effort has a financial cost. And unfortunately, as we will demonstrate, this cost is ultimately paid by end-users of the market, such as asset managers, who couldn't care less about ten billionths of a second. We have long since passed the point of diminishing returns as David Olsen, President of leading uHFT, Jump Trading neatly observes:

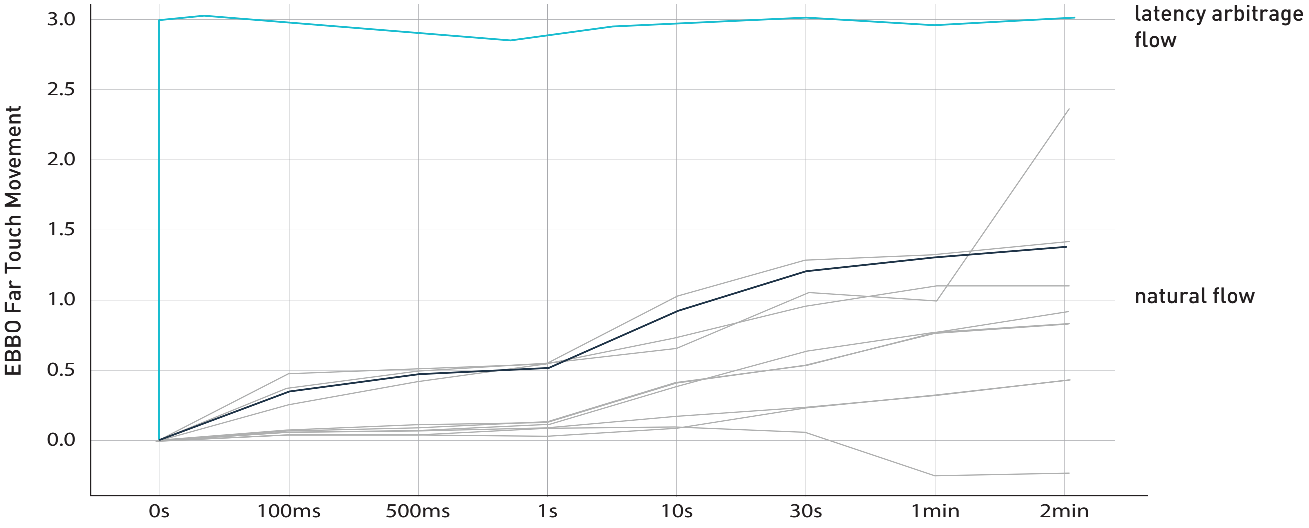
> *"Going from 80% efficient to 90% was pretty cheap and a fairly meaningful payoff, but going beyond 99.9% is incredibly expensive."*

High Frequency Traders are not charities and whatever costs they incur in sending satellites into space or tunnelling through mountains, they recoup from the market. There are several main ways in which this directly affects the buy-side.

1. Wider spreads and thinner books on exchanges because of the latency arbitrage component of flow

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.



Illustrative mark out chart. Natural flow and latency arbitrage flow are totally different. Because those providing liquidity to the market do not know which one they'll interact with they must show wider spreads and reduced depth. In this respect natural flow is subsidising the latency arbitrage 'pick offs' by having to trade on wider spreads and seeing reduced depth.

ALA mechanisms make it harder for latency arbitrageur taking strategies to perform latency arbitrage on liquidity providers. Once you remove the latency arbitrage, what's left is natural flow 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.

2. Costs are baked in to spreads

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.

LATENCY RACE
▼
HIGH FIXED COSTS
▼
CONSOLIDATION AND LACK OF NEW ENTRANTS
▼
MONOPOLISATION OF LIQUIDITY PROVISION
▼
HIGHER COST OF LIQUIDITY

3. High barriers to entry result in less competition and diversity amongst makers

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.

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

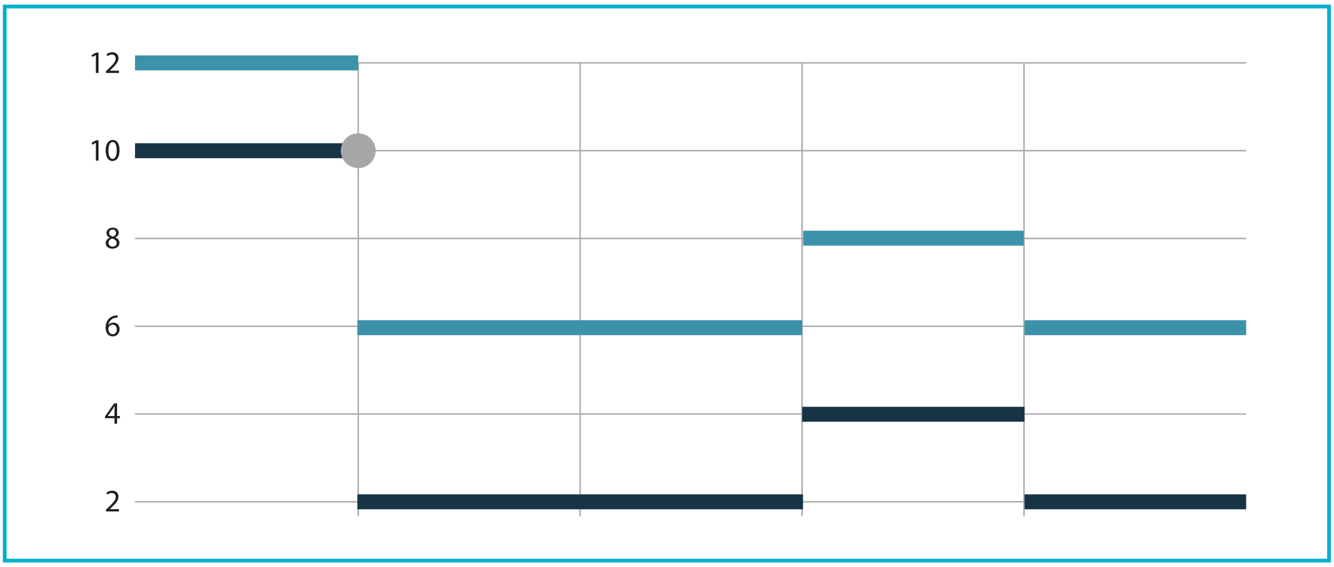
This leads to systemic risk, as a small number of HFT firms have limited risk absorption capabilities in relation to their outsized market share, and the failure or operational interruption, even if brief, of such an entity would have a disproportionate adverse impact on the market and liquidity relative 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, especially in times of volatility.

4. End-users get 'picked off' when using broker algorithms to peg to EBBO

Asset managers access the market via bank broker Smart Order Routers and algorithms. In fact CBOE explains that on its EDGA venue "two thirds of volume on the make side is attributed to agency brokers ". These algorithms will often try to trade passively by pegging to the EBBO on their chosen side.

However, these resting orders often get 'picked off' by uHFT aggressive orders who have observed a related market tick and are able to ship the data across and fill the orders before the broker algorithms can update. This greatly increases the overall cost of execution.

Consider the example below. The market is 10 /12.



A bid at ten gets filled in 1ms before the market reprices down to 2 / 6.

It is technically a passive fill at 10 but far from capturing the spread it has been latency arbitraged. Far better to avoid the initially adverse selection and sit on the bid at two or even 'pay the offer' at six.

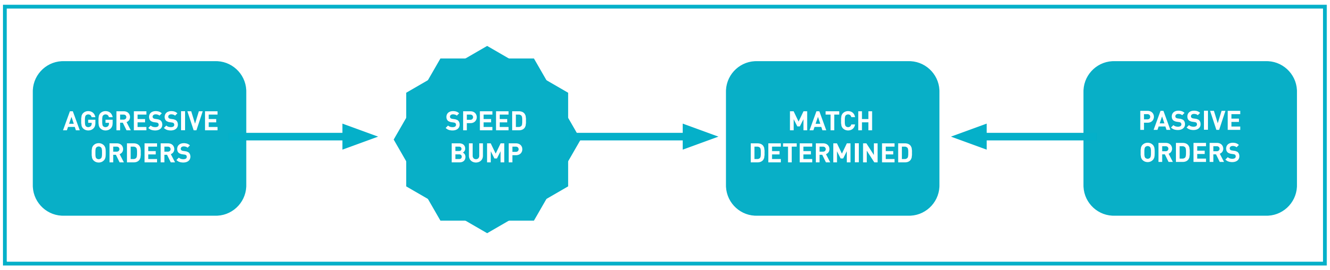
BACKGROUND: ANTI-LATENCY ARBITRAGE MECHANISMS

If long-term investors are getting the bill for the uHFT speed race, where exactly can I see it?

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.

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.

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.

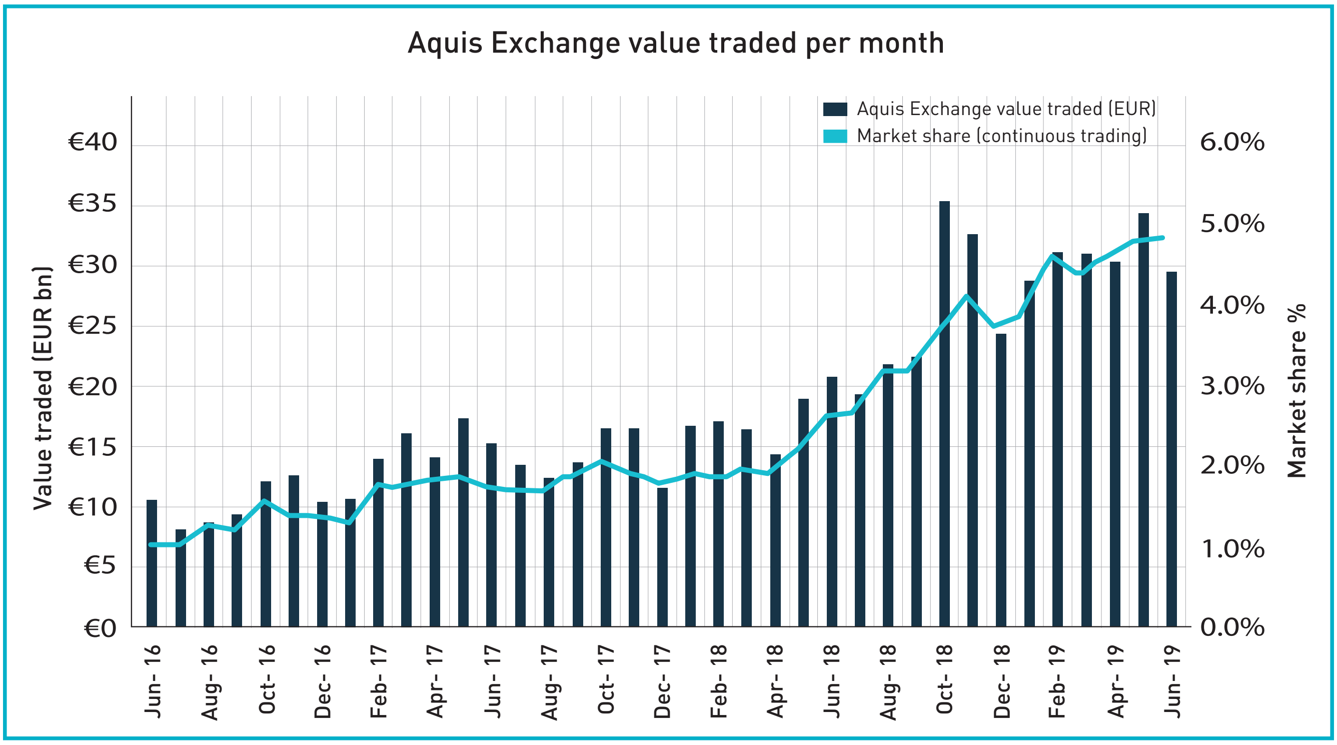


Chart taken from www.aquis.eu and shows the growth of market share on this venue because buy-side traders have benefited from its policy-based ALA.

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.

[GENERAL ACCEPTANCE OF ALA'S]

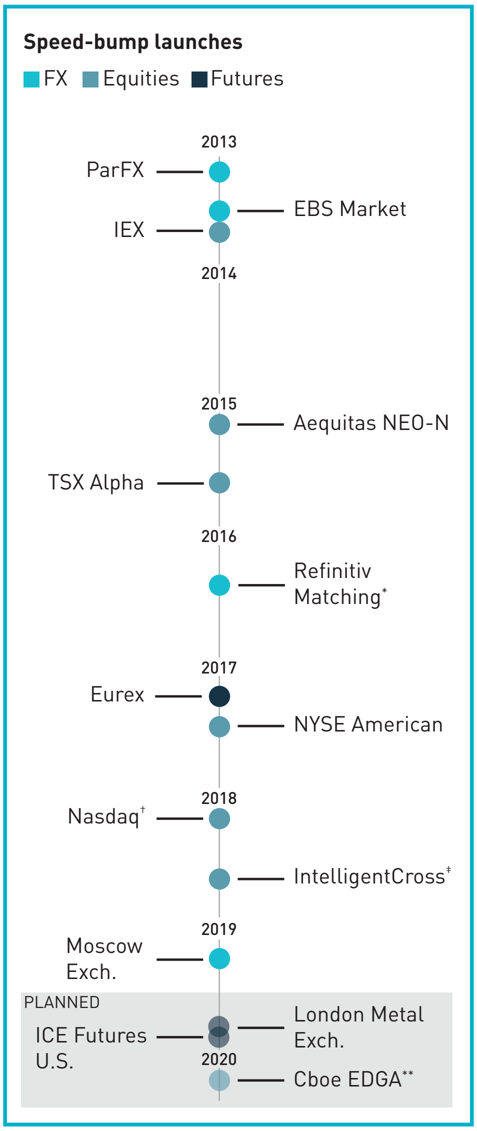
How do ALAs work?
Technology-based ALAs

The Wall Street Journal has created an excellent timeline showing the spread of ALAs across multiple geographies and asset classes. Clearly this is a natural response to an underlying issue – the destructive speed race – and we are likely to see more of them in future.

Taking a Pause

More than 10 markets have added speed bumps or similar features since 2013.

This timeline is not comprehensive: there are many other varieties on this theme. For example, the M-ELO order type which only becomes eligible for matching with other orders after a one-second pause.



Source: Wall Street Journal article: https://www.wsj.com/articles/more-exchanges-add-speed-bumps-defying-high-frequency-traders-11564401611

HOW ARE INSTITUTIONAL ORDERS PROTECTED BY ALA'S?

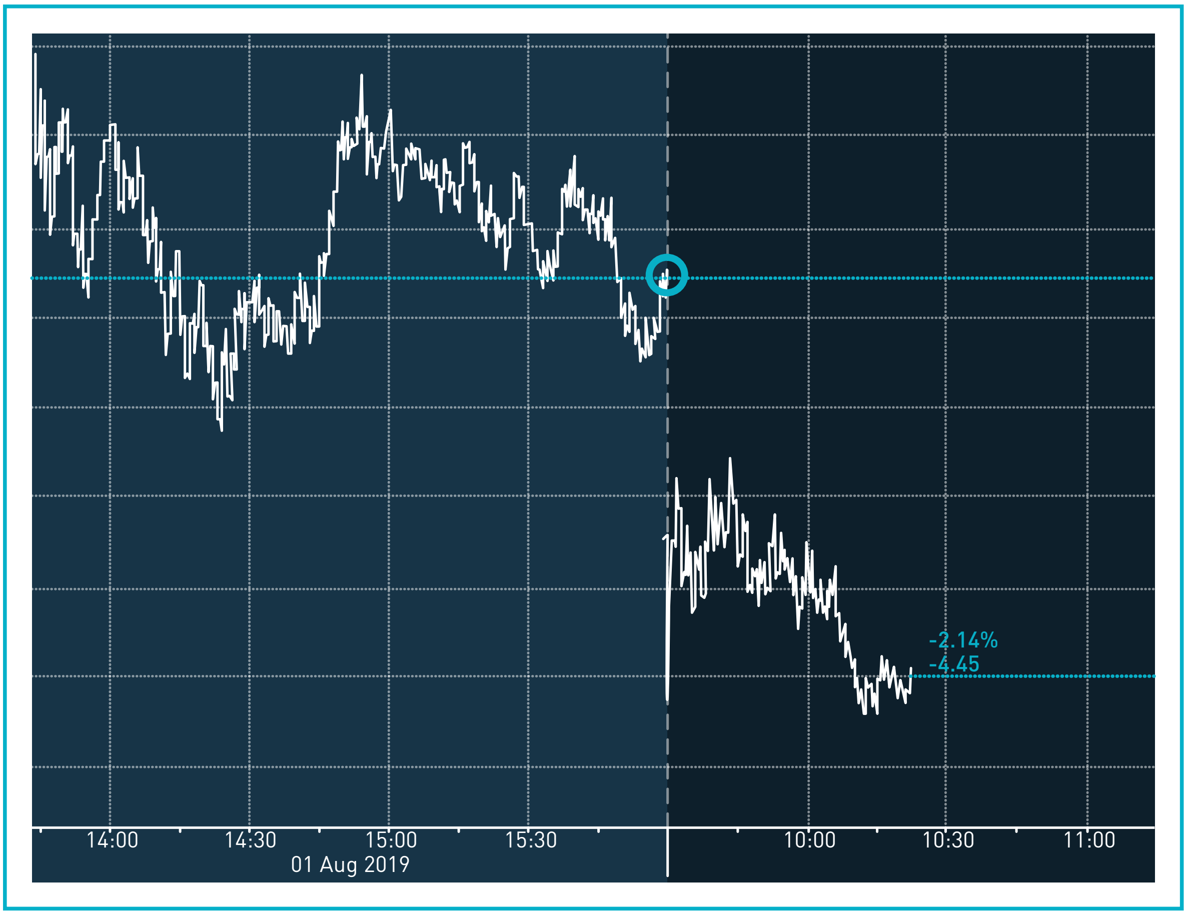
How do ALAs work?

Policy-based ALAs

Asset managers access the market via bank broker Smart Order Routers and algorithms. In fact CBOE explains that on its EDGA venue "two thirds of volume on the make side is attributed to agency brokers ". These algorithms will often try to trade passively by pegging to the EBBO on their chosen side.

However, these resting orders often get 'picked off' by uHFT aggressive orders who have observed a related market tick and are able to ship the data across and fill the orders before the broker algorithms can update. This greatly increases the overall cost of execution.

Imagine a VWAP that is buying APPL and getting filled just as the market ticks lower.



Illustrative chart of AAPL price action. Note the pick off – the buyside floating bid has got filled (just as the market ticks lower) by an HFT who received the market data a fraction of a second before the buyside's algorithm provider could see it and react.

The important thing to note is that a well designed ALA mechanism will allow the typical commercially available algorithm plenty of time to benefit from this protection against latency arbitrage. As a result, end-users will benefit from reduced adverse selection and lower overall execution costs whenever they are working orders passively.

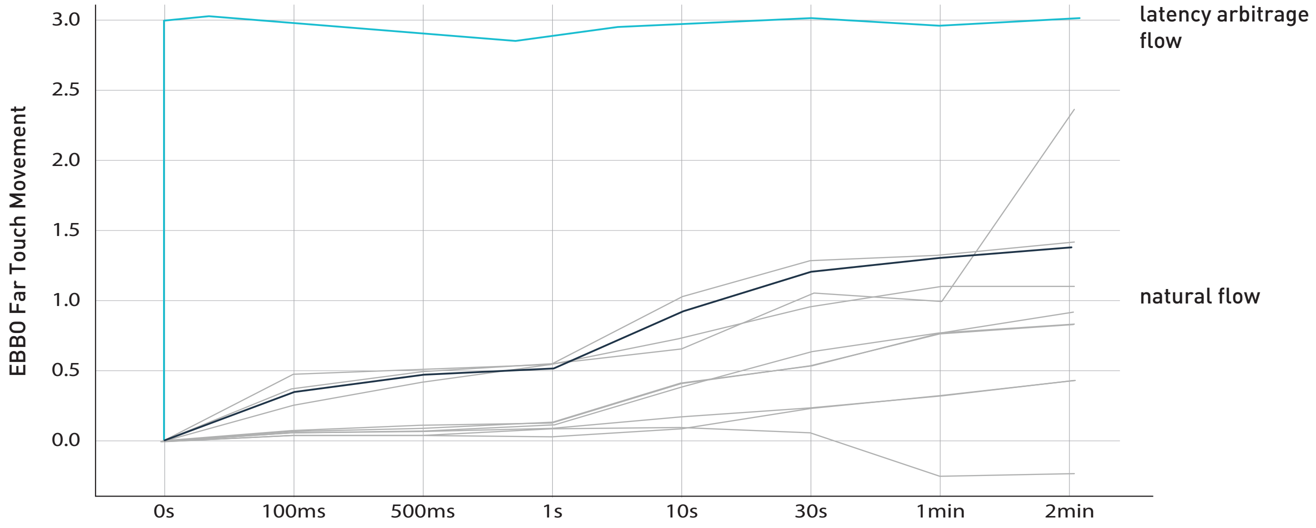
VASTLY DIFFERENT FLOW: ALA'S VS LATENCY ARBITRAGE

How might things work for an institutional investor on a venue with an ALA? Aggressing visible liquidity.

On a venue with an ALA institutional investors should expect to see both tighter spreads and deeper books.

The reason is simple. 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.



Illustrative mark out chart. Natural flow and latency arbitrage flow are totally different. Because those providing liquidity to the market do not know which one they'll interact with they must show wider spreads and reduced depth. In this respect natural flow is subsidising the latency arbitrage 'pick offs' by having to trade on wider spreads and seeing reduced depth.

ALA mechanisms make it harder for latency arbitrageur taking strategies to perform latency arbitrage on liquidity providers. Once you remove the latency arbitrage, what's left is natural flow 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.

Tighter spreads and deeper books are something worth having!

Retail investors may potentially benefit, too. In truth most retail flow is traded off exchange bilaterally – in the US under 'payment for orderflow' agreements with large HFTs. As such their flow wouldn't directly interact with an ALA on exchange and things would look much like they do today … except that the HFT might feel an obligation to price-improve relative to an exchange NBBO that is now tighter than it was before. Therefore retail investors would stand to benefit, even if they do not directly interact with the exchange in question.

[WHAT NEXT]

Is it really that simple?

No, it is not. We 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.

How, then, should exchanges proceed? 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.

[Q&A]

Devil's advocate

1. But I heard 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:

MARKET IS 10/11

▼

END USER SENDS BUY ORDER AT 11

▼

PRE-HEDGER PLACES ORDER INTO 100ms LAST LOOK WINDOW

▼

DURING THIS 100ms, PRE-HEDGER ATTEMPTS TO BUY IN THE MARKET BELOW 11

▼

IF FILLED BELOW 11, PRE-HEDGER FILLS END-USER AT 11 AND LOCKS IN A RISK-FREE PROFIT; IF UNFILLED, PRE-HEDGER REJECTS END-USER ORDER

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. Costs are baked in to spreads

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.

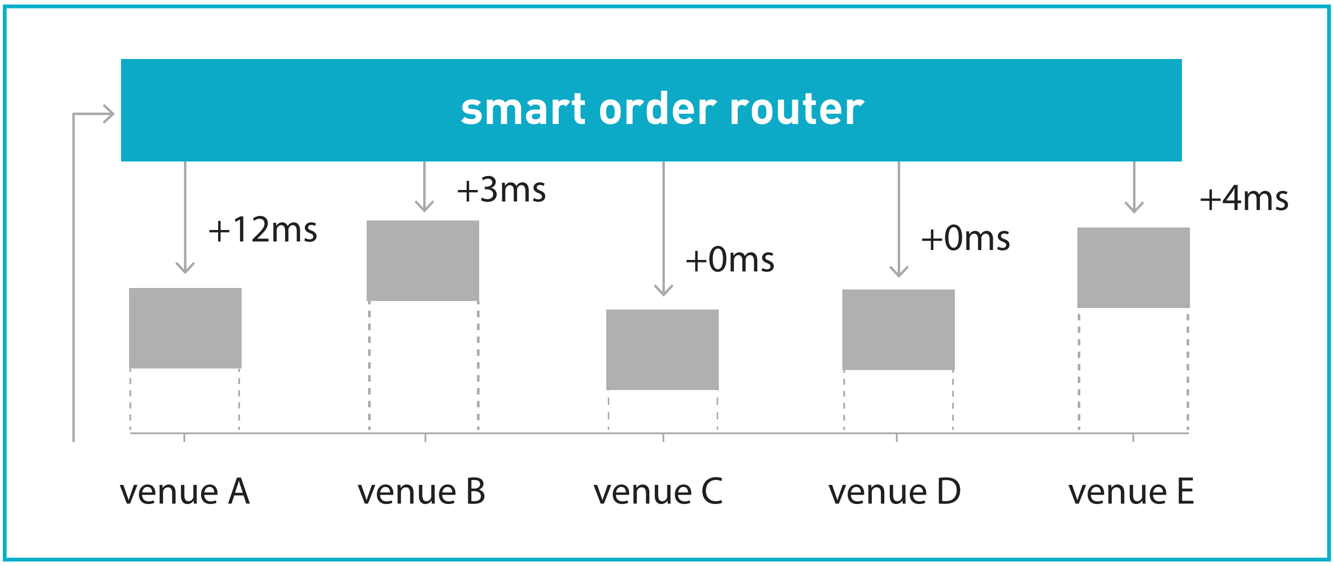
7. Wouldn't the fill rate go down for buy-side clients, too?

No. 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.

It is true that broker SORs may adjust their order routing logic in order to avoid swiping multiple markets simultaneously, which could result in lower fill rates. All the SOR has to do is simply stage the order – as they already do across exchanges in different locations – so that the child orders complete at the same time. For example if a venue has a deterministic 4-ms speedbump then the algorithm simply sends that order 4-ms before the others so that all the orders complete at the same time.



Bank smart order routers can (and do) sequence orders so that they all complete at different venues simultaneously. The SOR must simply account for geographic/telco latencies and any fixed delays when staggering its orders. This way the SOR and its buyside clients should expect to enjoy the same high fill rate on venues who have added a fixed length speedbump as they have historically. They should not expect any increase in rejections.

Similar concerns were raised in Canada when the TMX launched an anti-latency arbitrage mechanism like LP² on its Alpha exchange in 2015. The Ontario Securities Commission ("OSC") conducted a review of market quality post implementation of the changes and published its results in February 2018. With respect to this specific concern, the OSC found that dealers continued to route institutional investors' orders to Alpha, but the way they did so had changed:

> **In certain situations, fill rates on Alpha have decreased, often for orders that are expected to go through multiple price levels or need to be split and sent to multiple marketplaces simultaneously (e.g. institutional orders). Some dealers reported initial fill rates to be much lower on Alpha in these circumstances, and some have modified their routing strategies to achieve improved outcomes. For smaller orders that can be executed on a single marketplace, some dealers have experienced improved execution results that are consistent with observations of larger average trade sizes on Alpha.**

So, while accessing liquidity on an ALA venue may require a different approach, evidence suggests that market participants can adapt their routing strategies and enjoy high fill rates.

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 behaviour 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.

While it is true that liquidity providers tend to widen and, in some cases, withdraw their quotes during periods of market stress, one of the main reasons spreads widen is because the rate of latency arbitrage activity increases during these periods. Consequently, a marketplace that offers protection against latency arbitrage, and which enables liquidity provision from a wider group of market participants with diverse risk absorption capabilities and investment horizons, should offer relatively tighter spreads over volatility spikes.

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 on their aggressive flow, 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.

And let's remember the crucial point: institutional investors also leave passive orders via broker algorithms. They would also benefit from reduced latency arbitrage since the broker algorithm is now able to compete on a level playing field with the fastest uHFT takers. Given that on exchanges like EDGA a full two-thirds of passive volume is coming from these buy-side agency algorithms, it seems pretty clear that the benefits would be widespread.

11. 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.

12. Didn’t they try this in Canada?

The TMX Exchange Group operates an unprotected exchange, TSX Alpha, in Canada with 1-3 millisecond randomized delay on orders to remove liquidity that is reasonably comparable to EDGA’s proposed LP². Two commenters have seized on that comparison to cite academic research on the impact TSX Alpha has had on the Canadian equity markets to suggest that EDGA’s LP² would have a negative impact on liquidity. In both cases the commenters neglected to reference a subsequent academic study that found no evidence that TSX Alpha negatively impacted market-wide liquidity, market-wide trading costs or execution quality. In 2018, the Ontario Securities Commission published its own review of TSX Alpha’s effect on the Canadian equity market and found no negative impact to market quality.

13. Wouldn’t it just make market makers rich?

Nope. On a venue with an anti-latency arbitrage mechanism one would not expect that professional liquidity providers would suddenly be able to extract outsized rents from their market making activity. For existing market makers, the incremental cost of launching market making on such a venue is immaterial. If an ALA allows market makers to earn outsized returns, other market makers will enter the market and normalize it. In other words, market makers will compete against each other aggressively on the venue on both price and order size, putting the dollars saved by avoiding latency arbitrage into the pockets of investors.

14. How does this work in the US equities context? Should the quote be protected and how would it appear in the SIP?

US equities is indeed a special case because of things like the Order Protection Rule. It seems reasonable that venues with ALAs should choose to give up their protected status. This has the neat effect of meaning that such venues only win volume because they deserve it i.e. investors have a good experience when trading on there and choose to send flow to the venue. The venue succeeds or fails based on merit!

Inclusion of quotes into the SIP market data can be complicated. We take the view that it is reasonable for pegged orders to peg off reference to the protected BBO and ignore unprotected quotes, just as Canadian markets do today. Equally, it seems reasonable to us to exclude unprotected quotes from consideration for regulatory references such as Regulation SHO's price test.

With respect to concerns that including EDGA's unprotected quote on the SIP creates confusion over best execution obligations, we aren't certain we fully understand this argument. The SEC and FINRA have provided various best execution guidance and whether the quote is included in the SIP would not seem relevant to the question of whether EDGA's manual quote would need to be accessed to satisfy best execution requirements. The SEC noted in the Regulation NMS final rule order regarding the inclusion of manual quotes in the NBBO:

> **The Commission continues to be concerned that eliminating all manual quotations from the NBBO would exclude not only inaccessible manual quotations, but also manual quotations that truly establish the best available price for a stock... Such a result could lead to decreased execution quality for investors in these stocks by allowing broker-dealers to ignore the best available quotations when executing customer orders.**

The SEC further noted in the Reg NMS final rule order that the decision to access a manual quotation rests with the broker-dealer's review of execution quality:

> **The Commission continues to be concerned that eliminating all manual quotations from the NBBO would exclude not only inaccessible manual quotations, but also manual quotations that truly establish the best available price for a stock... Such a result could lead to decreased execution quality for investors in these stocks by allowing broker-dealers to ignore the best available quotations when executing customer orders.**

As such, it appears the Commission has squarely addressed the best execution concerns raised by commenters.