CHICAGO BOOTH
The University of Chicago Booth School of Business

ERIC BUDISH
Professor of Economics and Richard N. Rosett and David G. Booth Faculty Fellow
The University of Chicago Booth School of Business
5807 South Woodlawn Avenue Chicago, Illinois 60637
Phone: (773) 702-8453 Email: eric.budish@chicagobooth.edu

Feb 5, 2016

Brent J. Fields
Secretary
US Securities and Exchange Commission
100 F Street, N.E.
Washington, D.C. 20549-0609

Re: Investors' Exchange LLC Form 1 Application (Release No. 34-75925; File No. 10-222)

Dear Mr. Fields:

I appreciate the opportunity to comment on Investors' Exchange LLC's ("IEX") application to become a national securities exchange. I am an economics professor at the University of Chicago Booth School of Business who researches market design – designing the "rules of the game" for markets – with a specific focus on the design of financial exchanges. Market design research assumes that participants in a market act optimally in their rational self-interest with respect to market rules, but takes seriously the possibility that the market rules themselves may be sub-optimal. I believe that this approach brings a useful perspective to the debate over IEX, and equity market structure more broadly. I write independently, without any financial involvement with any of the participants in this debate.

I support IEX's application to become a national securities exchange. IEX's proposed market design, while not a panacea and imperfect in ways I will describe below, has several innovative features that are constructive for investors, and based on my reading of the relevant law and regulations I do not see a compelling reason not to approve the application. I am also philosophically supportive of innovation in this space and agree with IEX's argument, expressed in its second comment letter dated 11/23/2015, that exchange design innovation aimed at the more negative aspects of high-frequency trading is important and should be allowed within the context of our current regulatory structure.

However, I caution that IEX's proposed exchange design, while constructive, does not fix the underlying problems with our current equity market structure. It only addresses latency arbitrage in a limited way, and does so in a way that free-rides off of price discovery on other exchanges rather than in a way that directly contributes to price discovery. And it does not reduce the overall complexity of our equity market structure, but arguably makes it yet more complicated (though I applaud its simplified fee structure and its overall approach to marketplace transparency). These limitations of the proposed design reflect IEX's attempt to devise an innovative exchange design within the severe regulatory constraints of Regulation National Market System ("Reg NMS"). A more complete solution will require both private-sector efforts analogous to IEX's as well as reform to the underlying regulation of the US equity market structure, which I will argue, while well intentioned, was critically flawed from the outset.

In the remainder of this comment letter I will make the following 4 points:

1. Latency arbitrage is “built in” to the continuous limit order book market design; this is the market design used by all 12 exchanges in the National Market System and the design implicitly assumed in Reg NMS. It is also the market design used by the displayed part of IEX.
2. IEX’s proposed market design addresses latency arbitrage for non-displayed pegged orders while doing nothing to address latency arbitrage for displayed limit orders. This is unfortunate since it is displayed orders that contribute to price discovery, while non-displayed pegged orders free-ride off of others’ price discovery; and ironic, because the whole point of becoming an exchange is the displayed orders, but for these orders IEX’s design is no better or worse than anyone else’s. These shortcomings likely stem from IEX’s efforts to design within the constraints of Reg NMS.
3. A more direct and complete way to eliminate latency arbitrage is to use a market design called frequent batch auctions. This requires treating time as discrete (e.g., in units of 0.001 or 0.01 seconds), analogously to how we treat price as discrete (e.g., in units of 0.01 dollars), and then processing messages that arrive at the same discrete time in batch, using an auction, rather than serially by arrival time. Frequent batch auctions eliminate latency arbitrage completely, not just for certain types of orders, and do so in a way that contributes to price discovery rather than free-riding off of others’ price discovery. However, frequent batch auctions would be in tension with Reg NMS.
4. Reg NMS is critically flawed. It was well intentioned but made two critical errors. First, it implicitly assumes that there is literally zero latency between exchanges. Not 0.0001 seconds, or 0.000000001 seconds, but literally zero. This conceptual error in Reg NMS explains why neither IEX’s supporters’ nor IEX’s detractors’ arguments make complete sense; they are debating the interpretation of a law that is flawed at its root. Second, it implicitly assumes that all exchanges use the continuous limit order book. This constrains market design innovation (including, but not limited to, frequent batch auctions). A more complete solution to the problems with today’s equity market structure will require not only private-sector efforts like IEX’s but also Reg NMS reform.

1. Latency arbitrage is “built in” to the continuous limit order book market design.

The transition from human-based markets to electronic markets has on the whole been quite positive for markets, both for investors and for overall market efficiency. This is confirmed in the overall time series of transactions costs and by careful empirical studies that focus specifically on the transition from humans to computers.[1]

However, my research shows that in the transition to fully electronic continuous limit order books, exchanges made a subtle error, which causes latency arbitrage to be “built in” to the market design. The

[1] For overall time series evidence, see Andrea Frazzini, Ronen Israel and Tobias J. Moskowitz, “Trading Costs of Asset Pricing Anomalies,” Chicago Booth Research Paper No. 14-05, September 2015. For a study of the transition from human-based trading to algorithmic trading, see Terrence Hendershott, Charles Jones, and Albert Menkveld, “Does Algorithmic Trading Improve Liquidity?” *Journal of Finance*, Vol. 66(1), February 2011, pgs. 1-33.

flaw is that exchanges treat time as continuous – meaning, for any two orders received, one is first, even if by a nanosecond – and process messages serially, that is, one-at-a-time in order of receipt.

Here is the argument for why latency arbitrage is built in to the continuous limit order book market design.[2] Suppose there is a stock X, and a publicly observable signal, Y, about the value of stock X. Suppose for illustration that the signal is perfect in the sense that all market participants agree that X is worth exactly Y. Think of X and Y as a metaphor for either highly correlated financial instruments – X is an S&P 500 ETF and Y is the S&P 500 future. Or, X and Y represent exactly the same stock trading on different exchanges – X is the stock on NYSE or EDGE (or a dark pool), Y is the same stock on BATS or NASDAQ. Or, think of Y as representing public information that affects the value of X – company announcements, Fed announcements, SEC filings, etc.

Suppose that Y jumps. For instance, Y was at $10.00, and the market for X was at bid $9.99 – ask $10.01, and then Y jumps to $10.10. This jump in Y will generate a race to react in the market for X. On one side of the race are trading firms providing liquidity in the market for X – they will seek to cancel their old asks at $10.01, and also at $10.02, $10.03, etc., and replace them with new quotes. On the other side of the race are trading firms looking to buy X at the stale prices – to "snipe" stale quotes – before the quotes are canceled.

Here is the sense in which latency arbitrage is built in to the market design: even if the liquidity providers are at the cutting edge of speed, they still usually lose the race. (If they are not at the cutting edge of speed, they definitely lose). Because the continuous limit order book processes messages *serially, in continuous time* – that is, one at a time in strict order of receipt – each liquidity provider's request to cancel their stale quotes would have to reach the exchange before *all* of the stale-quote snipers' requests to trade at their stale quotes. If there are N trading firms at the cutting edge of speed, then since it's basically random who wins the race, a trading firm providing liquidity has only a 1/N chance of getting out of the way of the other firms who will try to snipe. Hence, the odds are stacked against liquidity providers, and they usually lose the race.

It is important to underscore how widespread the phenomenon of highly correlated assets is, and hence of latency arbitrage. Every ETF is highly correlated to other similar ETFs, to other similar futures, to a basket of its constituent components, etc.. Treasuries of any one duration are highly correlated to other similar duration treasuries and to similar duration treasury futures. Every option is highly correlated to its underlying and to other options for the same underlying. Every stock trading on any one exchange is *perfectly* correlated to that same stock trading on any other exchange. Essentially, any time any asset jumps in value – and assets are *supposed* to change in value in a well-functioning market, sometimes by large amounts! – there is a potential latency arbitrage opportunity.

[2] For the full mathematical details of the argument, see Section VI of Eric Budish, Peter Cramton and John Shim, "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *Quarterly Journal of Economics*, Vol. 130(4), November 2015, pgs. 1547-1621. Available under Open Access license at: http://faculty.chicagobooth.edu/eric.budish/research/HFT-FrequentBatchAuctions.pdf.

This built-in latency arbitrage is like a tax on liquidity provision that causes markets to be less liquid than they otherwise would be.[3] In particular, it is especially costly for liquidity providers to offer a deep book – a lot of depth at the bid and the ask – because the cost of getting sniped scales linearly with depth, whereas the benefits of providing a deep book scale less than linearly. If a trading firm offers a deep book in stock X, and there is a jump in signal Y, the firm will get sniped for the full amount offered. Investors frequently complain about the difficulty of trading in size; latency arbitrage is part of the reason why.

This built-in latency arbitrage is also the underlying driver of the "arms race" for speed among high-frequency trading firms over the past decade – first measured in increments of 0.01 seconds, then 0.001 seconds, then 0.0001 seconds, then 0.00001 seconds, now in increments of 0.000001 seconds or finer. Stale-quote sniping requires speed to win the race to snipe stale quotes, and liquidity provision requires speed to get out of the way of the snipers. It is also the underlying driver of the competition among exchanges to offer faster and faster connectivity to their markets, and to charge accordingly – colocation services and proprietary data feeds are arms in the arms race.

2. IEX's proposed market design addresses latency arbitrage only in a limited way.

IEX's market design applies a 350 microsecond delay to all message traffic between traders and IEX. On its own such a delay is pointless, because it has no effect on the relative order in which messages are received. To borrow the analogy of Hudson River Trading (comment letter dated 12/4/2015), if you add 0.00035 seconds to the times of all participants in a 100-meter dash, you don't change the finishing order. What is clever about IEX's design is that, in parallel to applying this 350 microsecond delay to requests to trade, IEX obtains information from other exchanges in the National Market System without the delay. Because of the geography of New Jersey data centers, and the speed at which information can travel, IEX is able to obtain information from most other exchanges in about 200 microseconds.

This combination of the 350 microsecond speed bump and the 200 microsecond information from other exchanges allows IEX to prevent latency arbitrage for pegged orders on its market. If prices jump on other exchanges (Y in the language of the previous section), then IEX can detect that jump in 200 microseconds and automatically adjust pegged quotes on the IEX market. Consequently, any firm that tries to snipe the pegged quotes will be too late – it will take a trading firm at least 350 microseconds to snipe, because of the delay, but in 200 microseconds IEX can adjust the quote to reflect the new information from other exchanges. This is very clever and should eliminate latency arbitrage for pegged orders.

However, there is a critical limitation: IEX's design only prevents latency arbitrage for pegged orders, which are non-displayed, and does not prevent latency arbitrage for standard limit orders, which are displayed. This is not an oversight, but rather reflects a fundamental tension in IEX's design. IEX's method of preventing latency arbitrage relies directly on price information coming from other exchanges, so IEX is only able to prevent latency arbitrage for orders that are explicitly pegged to prices

[3] Note that many HFT firms engage in both liquidity provision and stale-quote sniping. The argument still holds exactly. Sniping causes liquidity provision to be more expensive than it otherwise would be. HFT firms aren't charitable organizations that provide liquidity at a loss because they are making a lot of money sniping.

discovered elsewhere. For standard, plain-vanilla limit orders – which contribute to price discovery, rather than being pegged to prices discovered elsewhere – IEX's design has no effect on latency arbitrage. The displayed part of IEX's market is a standard continuous limit order book (as essentially mandated by Reg NMS), with latency arbitrage built in. The only difference is that the race between liquidity providers and stale-quote snipers is delayed by 0.00035 seconds out of the starting gate.

A related concern is that, because IEX only prevents latency arbitrage for non-displayed pegged orders, and because non-displayed orders have lower priority than displayed orders even if the displayed orders are submitted later in time, the orders for which IEX does prevent latency arbitrage seem likely to be subject to disproportionate amounts of traditional adverse selection. If an investor uses a pegged order on IEX, the order will only trade if all liquidity provided by traditional limit orders is consumed first. So, investors using pegged orders on IEX only trade if either trading firms providing liquidity using limit orders do not want to be in front of them in line (and, because displayed orders have higher priority even if they are submitted later, they can move to the front of the line whenever they like), or if someone trades a large enough block that it sweeps up whatever liquidity is provided by limit orders and the investor's order as well.

These design choices reflect the very real constraints faced by IEX in architecting its market within the confines of Reg NMS. But there is a certain irony here. IEX is asking for official exchange status, which is essentially the right to have displayed limit orders that count towards the National Best Bid and Offer and can earn order protection. But for this part of its market IEX's design doesn't actually do anything to prevent latency arbitrage. And for the orders for which IEX's design does prevent latency arbitrage, it does so in a compromised way because of the excessive adverse selection described above.

When I described my concerns about IEX's market design to a colleague, he said it reminded him of the old Soviet joke:

> Soviet Patriot: "The USSR will invade and conquer every country in the world, except New Zealand."
> Curious Observer: "Why leave New Zealand out of the global communist economy?"
> Soviet Patriot: "So we can find out the market price of goods."

IEX's method of preventing latency arbitrage is extremely clever, but the essential limitation is that it relies on looking to other exchanges to find out what the prices are.

3. A more direct and complete way to eliminate latency arbitrage is to use frequent batch auctions.

The continuous limit order book market design used by all 12 exchanges in the National Market System (and used by the displayed part of the IEX market design), and which is implicitly assumed by Reg NMS, treats time as continuous. Continuous time means that it is economically meaningful for my order to arrive 0.000001 or 0.000000001 seconds earlier than yours – my order is earlier (even if for essentially random reasons), hence is processed first and gets priority. Prices, on the other hand, come in discrete units – for instance, stocks trade in units of $0.01 – and there are very good reasons for this. Markets would work less well if I could outbid you by $0.00000000001 (a "nanopenny") and gain priority.

My research[4] suggests that we should put time into units too. The unit of time should be long compared to the time it takes information to travel between exchanges, and long compared to the time it takes exchange computers to perform simple processing and communication tasks, but otherwise can be extremely short. For instance, 0.001 seconds would be long relative to the relevant latencies among exchanges located in New Jersey, and 0.01 seconds would be long if Chicago were included as well.

Once time is put into discrete units, it becomes possible that multiple orders arrive to the exchange "at the same time" (e.g., in the same millisecond). My research suggests that such orders be processed in batch, using uniform-price auctions, rather than serially (one-at-a-time) as in the limit order book. Intuitively, batch processing using an auction replaces speed competition with price competition, in the event that there are many traders responding to the same signal at around the same time. There is still such a thing as time priority, but to earn time priority an order has to be present in the book for more discrete units of time (more batch intervals), not just be a nanosecond earlier within the batch interval. Orders received in the same batch interval are treated equally.

This market design, called frequent batch auctions, eliminates latency arbitrage completely, not just for non-displayed orders pegged to prices elsewhere. Both market design differences versus the limit order book – discrete time, and batch processing using an auction – play a role in eliminating latency arbitrage. Discrete time gives liquidity providers a window of time to cancel their old quotes and replace them with new quotes in the event of news (e.g., a jump in a related asset or the same stock traded elsewhere); their canceled stale quotes are not even entered into the next auction. More subtly, the auction itself protects against latency arbitrage. Suppose an investor with a limit order in the book does not see some news event in time to react (e.g., he is not fast enough). Then the investor's order, at the stale price, will be entered into the next auction; but, if he trades, he will trade not at the price of his stale quote, but at the auction price, which reflects price competition in response to the new information. If many trading firms see the news and know that the investor's quote is stale, then rather than compete on speed to be first to trade at the stale quote, they compete on price.

Importantly, this market design protects both sophisticated algorithmic trading firms and ordinary investors from latency arbitrage, and it does so without pegging to prices discovered elsewhere – the whole point of an auction is to discover the market-clearing price! Notice, too, that investors are

[4] See Budish, Cramton and Shim (2015), *supra* fn 1, and also Eric Budish, Peter Cramton and John Shim, "Implementation Details for Frequent Batch Auctions: Slowing Down Markets to the Blink of an Eye," *American Economic Review*, 104(5), 418-424, May 2014. Available by permission of the journal at http://faculty.chicagobooth.edu/eric.budish/research/HFT-FrequentBatchAuctions-ImplementationDetails.pdf.

protected from latency arbitrage without having to sacrifice priority to a different class of orders – priority is still strictly price then time, with the nuance that time is now discrete.

For all of these reasons, frequent batch auctions are a more complete approach to the latency arbitrage problem than is IEX's proposed design. Once you realize that the mathematical cause of latency arbitrage is treating time as continuous and processing messages serially, it makes sense that the mathematical solution to latency arbitrage is treating time as discrete and processing messages in batch.

So, if discrete-time auctions are such a good idea, why aren't they being used anywhere? There are at least two obstacles. First, the fact that frequent batch auctions improve liquidity and market quality does not necessarily imply that they improve exchange profitability – in particular, exchanges would lose some of their power to charge high prices for latency-sensitive market data, for colocation, etc. Second, and perhaps more importantly, frequent batch auctions are in tension with Reg NMS.

4. Regulation NMS made two critical errors: it implicitly assumes that there is literally zero latency between exchanges, and that all exchanges use the continuous limit order book market design.

Regulation National Market System ("Reg NMS") was well intentioned but made a critical logical error: it only makes sense as written in a world with *zero* latency. Not 0.1 seconds, not 0.01 seconds, not 0.001 seconds, not 0.000001 seconds, not 0.000000001 seconds, but literally zero. In a world with latency between exchanges – and since communications and computers are not infinitely fast, there *has* to be latency between exchanges – Reg NMS as written is logically incoherent.

I suspect that the drafters of Reg NMS simply didn't foresee that millisecond and now microsecond-level latencies could possibly make a difference in financial markets – and it is really bizarre, stepping outside the market structure bubble, that they do – but, given the current design of financial exchanges, such latencies do matter, and given that they matter, Reg NMS is logically flawed. I think the reason why neither IEX's arguments nor IEX's opponent's arguments make complete sense is that they are arguing over the interpretation of a rule that at core doesn't make sense.

Here is the issue. In continuous time, with zero latency, it makes sense to ask "what is the best price across all exchanges, right now?" This is a meaningful object, because with zero latency you can both (i) measure what is the best price right now, and (ii) be assured that you can trade at this price right now if you want to. But, as soon as you live in a world with non-zero latency – and, because information can't travel faster than the speed of light, and because it takes computers non-zero time to process trade messages, the world will always have latency – this no longer works. The best you can hope for is (i) measure what the best price was a latency ago (e.g., 0.0002 seconds ago), and (ii) hope that you can still trade at this price a latency from now (e.g., 0.0002 seconds from now).

That's the intrinsic flaw at the core of NMS – it is written as if you can ask, what is the best price on every exchange *right now*, that I can trade at *right now*, but the best you can hope for, mathematically, is what is the best price on every exchange *a latency ago*, that I can try to trade at *a latency from now*. You will get different answers depending on your vantage point, and the prices that you think are best, based on where they were a latency ago, might not actually be prices that you can trade, because they won't be there when you go to trade with them a latency from now. But Reg NMS mandates that orders are routed based on these intrinsically latent prices.

This conceptual error at the heart of Reg NMS helps explain why, in the argument over whether IEX's 350 microsecond speed bump should be allowed within Reg NMS, neither side's arguments make complete sense. IEX points out that its 350 microsecond delay is an order of magnitude smaller than the delay caused by the location of the Chicago Stock Exchange; since the 4 milliseconds of latency imposed by Chicago's geographical location is allowed within the confines of Reg NMS, surely so should the 350 microseconds of latency imposed by IEX's magic shoebox. I find this argument by analogy persuasive, but it does beg the question of why locating in Chicago is allowed in the first place – could an exchange within the National Market System locate its servers in Hawaii? On the moon? IEX's opponents argue that "intentional" delays violate the letter and spirit of Reg NMS. This argument too sounds kind of persuasive; but it begs the question, what exactly is the difference between IEX's so-called intentional delay and the kinds of delays caused by sub-optimal code, sub-optimal communications infrastructure, longer-than-necessary cables, etc.?

Ultimately, neither side's arguments are fully persuasive – and that's because the underlying rule they are debating the interpretation of is flawed at its core.

The second critical flaw with Reg NMS is that it implicitly assumes that all exchanges use the continuous limit order book market design. This flaw is not an explicit issue in the IEX debate, because IEX architected its market design within the constraints of Reg NMS (as described above, the lit part of IEX's market is a continuous limit order book). But, this is the reason why market design innovation is so constrained, and in particular is a reason why it would be difficult for an incumbent exchange, or IEX for that matter, to try frequent batch auctions.

Chair White emphasized the importance of eliminating regulatory obstacles to market design innovations in her June 5, 2014 speech "Enhancing our Equity Market Structure":

> We must consider, for example, whether the increasingly expensive search for speed has passed the point of diminishing returns. I am personally wary of prescriptive regulation that attempts to identify an optimal trading speed, but I am receptive to more flexible, competitive solutions that could be adopted by trading venues. These could include frequent batch auctions or other mechanisms designed to minimize speed advantages. … A key question is whether trading venues have sufficient opportunity and flexibility to innovate successfully with initiatives that seek to deemphasize speed as a key to trading success in order to further serve the interests of investors. If not, we must reconsider the SEC rules and market practices that stand in the way.

Reg NMS constrains market design innovation; implicitly imposes that all exchanges use a market design that has built-in latency arbitrage; and implicitly assumes zero latency, which is physically impossible. To put it simply: the SEC should fix this flawed rule.

Conclusion

I support IEX's application to become a national stock exchange. I think they have correctly diagnosed many of the flaws with today's equity market structure, and I admire the tenacity with which they have pursued correcting those flaws, especially given the regulatory constraints and vested interests they are up against. I also think, as a matter of principle, that it is important to allow for innovation within the current regulatory environment that attempts to address the negative aspects of high-frequency trading.

But, in market design, details matter. And, given the details, IEX is a step in the right direction but is far from a solution to the problems at hand.

To go further will require a more comprehensive solution, involving some combination of private sector efforts and regulatory reform; private sector forces alone won't do the job, because the underlying regulatory structure is flawed. Two clear steps are for the SEC to reform Reg NMS; and for an exchange (possibly IEX itself) to try a market design such as frequent batch auctions that more comprehensively addresses the flaws with today's equity market structure.

The continuous-time, serial-process limit order book – used by all 12 exchanges in the National Market System, proposed by IEX for the displayed part of its market, and implicitly mandated by Reg NMS – is a flawed market design, with built-in latency arbitrage, and too often elevates speed above all other considerations. We would never trade stocks in nanopennies. We should stop trading stocks in nanoseconds.

I will be pleased to be of service to the SEC in these important matters in whatever way is helpful.

Kind regards,

Eric Budish