



Big bucks from small change



Matteo Aquilina



Peter O'Neill

Article

🕒 7 mins

10 17

Share



27 January 2020

Latency races are worth a few dollars a time and take place in a timescale of micro-seconds, but for high frequency traders the stakes soon add up. For the first time, economists at the FCA and the University of Chicago have put some hard numbers on the costs.

In 2014 the arcane topic of high frequency trading (HFT) became a best-seller with Michael Lewis' book *Flash Boys*.

Its central premise, that high speed trading technology meant the US stock market was 'rigged for the benefit of insiders', caught the public imagination.

Insight

Defenders of HFT dismissed the book's central thesis as 'a myth.'

—In the years since the book's publication in 2014 the academic literature on high-frequency trading has continued to be quite active and has not found empirical evidence consistent with the most extreme or alarmist readings of Flash Boys.

But what has so far been conspicuously absent from the debate has been empirical evidence of what is at stake. How much profit is being made through exploiting tiny speed advantages and what, if anything, is it costing other investors?

We believe we have an answer. After years of work and analysis over billions of high frequency data points on a new kind of dataset, we conclude the sums at stake are about \$5 billion a year in global equities, and that the cost of liquidity for investors could be reduced by 17% by addressing the problem.

Beware – snipers!

The importance of speed in modern markets is undeniable. By some estimates, HFT firms account for over 50% of trading volume. Significant sums of money are spent to facilitate high speed trading, from micro-wave links between markets (data travels faster through air in a straight line than through fibre-optic cables) to locating trading operations next-door or even on the rooftops of exchanges. And of course, salaries for the best technologists that money can buy.

The core reason why speed is so valuable is that many financial instruments' prices are highly correlated, i.e., they move in lock-step.

For example: a stock that trades on multiple exchanges; or a stock and its options; or government bonds with slightly different maturities; and so on. Some statistical relationships are completely obvious whereas others require considerably more sophistication.

When asset prices are highly correlated, and the price of one related instrument moves, natural market forces mean other related instruments' prices will swiftly move to keep in line.

HFT firms can profit by moving even more quickly – taking advantage of the brief gap that exists, typically for a matter of microseconds (millionths of a second), before the prices realign. For those fleeting moments, some traders will still be offering prices that have yet to catch up with the change.

While economists call this latency arbitrage – the market jargon is 'sniping'.

Previous work by Prof. Eric Budish of the University of Chicago (one of the authors of this article) explains how this process has become an arms race that will never reach a conclusion. Latency arbitrage is hard-wired into any financial markets that allow continuous trading- there will always be a race to see who can shoot first. The result is an endless pursuit of speed, in which the difference between winners and losers is today measured mostly in single microseconds.

And latency arbitrage makes it more difficult, and more expensive, to provide liquidity in the market.

Any market participant who wants to provide liquidity in any asset will worry that they will be a tiny fraction of a second too slow to react to a price change in a related asset. They could then be traded against at a price that no longer most accurately reflects fair market value.

In fact, even liquidity providers at the very cutting edge of the speed race will sometimes get sniped, by other HFTs who are similarly fast.

Various proposals have been made, and in some cases implemented, to slow down this race for speed.

The London Metals Exchange, for example, is one of a number of trading venues that have proposed 'speed bumps' - delays in trading that attempt to level the playing field between fast and the not-quite-so-fast traders.

Some HFT operators are themselves in favour of such speed bumps – a clear sign that they are not always the automatic winners of the speed race.

Eric Budish's previously mentioned paper on HFT proposed a more complete solution - simply stop trading being continuous and implement Frequent Batch Auctions. This would involve trading being broken into discrete windows of time, so removing the incentive to trade any faster than the size of the batch window.

Insight

So, should financial markets be considering speed bumps or batch auctions, or indeed any other measures, to change the structure of trading and so curb the speed race?

The answer to that question depends on whether the current structure of markets, and the HFT arms race it is unwittingly encouraging, is causing significant costs to efficient markets.

Counting the losers and the losses

The challenge is how to measure empirically the market costs of latency arbitrage. Typically, one might assess market activity by examining the limit order books - such as the US "Trade and Quotes" (TAQ) dataset. These datasets are a record of every order that successfully changes the order book, including trades, new bids and new offers.

But the problem is we need to see not just the winners of the races, but also the losers.

The order book only records the winners, because they successfully change its status. How do we see and measure the losers - the attempts to snipe or to change an offer price that are too slow and so never appear in the order book?

The eureka moment was realising that there is another type of information that allows us to capture those failed attempts - message data.

Those who are too slow in a latency arbitrage race make no impression on the order book, but there is a record of their message sent to the exchange, and a reply from the exchange informing them that they were too late.

Our method relies on the simple insight that these failure messages are a direct empirical signature of speed-sensitive trading. If multiple participants are engaged in a speed race to snipe or cancel stale quotes, then some will succeed and some will fail. Only the winners are recorded in the order book - so it is hard to know that there was in fact a race - whereas both the winners and losers show up in the message data.

We obtained from the London Stock Exchange (by a request under Section 165 of the Financial Service and Markets Act) all message activity for all stocks in the FTSE 350 index for a nine week period in the autumn of 2015.

The messages are time-stamped at the exchange gateway with accuracy to the microsecond. Using these data, we can directly measure the quantity of races, provide statistics on how long races take, how many participants there are and the diversity and concentration of winners and losers.

And crucially, by comparing the price in the race to the prevailing market price a short time later, we can measure the economic stakes - how much was it worth to win?

Results

These sniping races are extremely frequent. Shares in the average FTSE 100 stock are subject to 537 latency-arbitrage races per day, or about one per minute.

Unsurprisingly given this rate of activity, the volume is also significant with these races accounting for 22% of average daily trading volume.

Comparing the timing of successful traders in these races with that of the losers, we can also see how small an amount of time makes a difference. Typically, the winner beats the first loser by just 5-10 microseconds; that's 0.000005 to 0.000010 seconds.

You might say it happens in the blink of an eye, but you'd be wrong - it is 1,000 times quicker than that.

This activity is also quite concentrated, with a handful of firms winning the bulk of the races. In fact, just 6 firms accounted for more than 80% of winners and losers of the races in the study.

Each race is typically worth a quite small amount. The average race is worth a bit more than half a tick, which comes to about £2.00.

Insight

So, the events are fleetingly brief, a small number of firms dominate the trade and the sums involved in each trade very modest. But there are tens of thousands of such races taking place every day and in aggregate, these small races add up to meaningful sums.

We find that the "latency arbitrage tax", defined as the ratio of daily race profits to daily trading volume, is 0.42 basis points (0.0042%).

If this still sounds like small beer, then consider it in real terms across the whole market. The annual sums at stake in latency arbitrage races in the UK equity market are about £60 million a year. Extrapolating globally, we find that the annual sums at stake in these races across world equity markets are about \$5 billion per year.

Even this is not quite the end of the story. Because being 'sniped' is an occupational hazard for trading firms, the risk is naturally included in the prices and quantities they submit to the market, reducing overall liquidity.

Our research suggests that eliminating latency arbitrage would reduce the cost of liquidity by about 17%.

Is this enough to matter?

Whether the numbers in our study seem big or small may depend on the vantage point from which they are viewed. As is often the case in regulatory settings, the detriment per transaction is quite small and a 0.42 basis points tax on trading volume certainly does not sound alarming.

But it all adds up. A 17% reduction in the cost of liquidity is undeniably meaningful for large investors, and \$5 billion per year is, as they say, real money.

And remember, we are talking here only about equities. The same phenomenon extends to other asset classes that trade on electronic limit order books such as futures, currencies, U.S. Treasuries and more.

The research indicates that ordinary households are not significantly impacted by the costs of this activity in their retirement and savings decisions.

Yet at the same time, flawed market design significantly increases the trading costs of large investors, and generates billions of dollars a year in profits for a small number of HFT firms, who then have significant incentive to preserve the status quo.

Keep reading

Insight



IEX President says tech can help solve market inefficiencies

Ronan Ryan, President of US stock exchange IEX, talked about fibre optic cables, regulation and Flash Boys at a recent Insight lecture on the future of trading. Here are the highlights.



Ronan Ryan

Speech

🕒 6 mins