



Latency Arbitrage: The Real Power Behind Predatory High Frequency Trading

A Themis Trading LLC White Paper

By Sal Arnuk and Joseph Saluzzi

Introduction

In previous white papers, we have discussed several High Frequency Trading (HFT) strategies and tactics, such as Liquidity Rebate Trading, and more predatory styles, such as Flash, Dark Pool Pinging, and Predatory Algorithms (or algos). It is the predatory styles that first alerted us to HFT because of its detrimental impact on traditional institutional and retail investors. In this paper we address the practice of **Latency Arbitrage**. We believe Latency Arbitrage is more than a simple case of technological evolution, but raises serious questions about the fairness and equal access of US equity markets which have made them the envy of the world.

Most industry professionals have been aware of the term Latency Arbitrage as it applies to HFT. The common assumption is that it refers to the spending or “arms” race in which high frequency traders employ high-speed hardware, software and bandwidth, to execute orders as fast as possible, in order to gain an edge in trading. Since lower latency equals faster speed, Latency Arbitrage was viewed as natural technological evolution that eventually would translate into more efficiency in the marketplace as the speed cascaded down to all investors. Talk of latency measured in milli, micro and nanoseconds was considered background noise.

At the same time, most industry professionals have been aware of the term **Co-Location** as it applies to HFT. This is the practice of market centers, such as the major exchanges, like NYSE Euronext, NASDAQ, BATS, as well as alternative trading systems (ATS), like DirectEdge, to rent out space to HFTs to co-locate their servers next to the market center’s servers, in order to further reduce latency. While some industry officials have questioned the fairness of paying for such an advantage, there are no restrictions on which firms can co-locate. So, theoretically, any institutional or retail brokerage firm could do it, in order to serve clients better.

The Reality of Latency Arbitrage

In practice, however, Latency Arbitrage means something completely different to HFTs.

Firstly, it is about using cutting edge technology and co-located servers at exchanges and ATSs, combined with purchases of raw data feeds from these market centers, to create one’s own inside National Best Bid and Offer (NBBO) quote and depth of book substantially earlier than what is publicly available from the Security Information Processor, or SIP quote. The SIP feeds quotes seen on professional terminals, algo trading systems used by institutions for as much as 50% of their orders, and quotes seen by retail investors on Internet sites.

"It's like you're seeing the *Wall Street Journal* five microseconds into the future," said Kevin McPartland, a senior analyst at financial services research firm TABB Group, at a seminar in November sponsored by TABB and Switch and Data, a data center operator.¹

Secondly, it is about using HFT techniques, such as Predatory Algos, Immediate or Cancel (or "cancel and replace") orders, and Dark Pool Pinging, to determine what kind of institutional algo orders are in the market, such as those driven by commonly used volume weighted average price (VWAP) formulas, and how those orders will react if the bid / offer of a stock moves up or down. These techniques, all of which also benefit from co-location, were initially outlined in our December 2008 white paper, "Toxic Equity Trading Order Flow on Wall Street."²

Armed with all this information, HFTs are able to achieve "(almost) risk free arbitrage opportunities," according to a report by Jefferies & Company, Inc., an institutional brokerage firm.³ To put it in layman's terms, "those are the types of investment strategies that arbitrageurs and hedge-fund managers drool over," Richard Gates, a portfolio manager for TFS Market Neutral fund in West Chester, PA, who has studied latency arbitrage, told a blogger for *The Wall Street Journal*.⁴

Here's an example of how an HFT trading computer takes advantage of a typical institutional algo VWAP order to buy ABC stock:

1. The market for ABC is \$25.53 bid / offered at \$25.54.
2. Due to Latency Arbitrage, an HFT computer knows that there is an order that in a moment will move the NBBO quote higher, to \$25.54 bid / offered at \$25.56.
3. The HFT speeds ahead, scraping dark and visible pools, buying all available ABC shares at \$25.54 and cheaper.
4. The institutional algo gets nothing done at \$25.54 (as there is no stock available at this price) and the market moves up to \$25.54 bid / offered at \$25.56 (as anticipated by the HFT).
5. The HFT turns around and offers ABC at \$25.55 or \$25.56.
6. Because it is following a volume driven formula, the institutional algo is forced to buy available shares from the HFT at \$25.55 or \$25.56.
7. The HFT makes \$0.01-\$0.02 per share at the expense of the institution.

It is currently estimated that HFT accounts for 60% of all share volume.⁵ Based on our experience trading, we estimate that at least 10 percentage points of that is of a predatory nature. Based on current average daily volume of about 10 billion shares, that means approximately 600 million shares per day are subject to predatory HFT. At \$0.01-\$0.02 per share, predatory HFT is profiting \$6-\$12 million a day. At about 250 trading days a

¹ Nina Mehta, "High Frequency Trading Is a Tough Game," *Traders Magazine Online News*, November 24, 2009, November 30, 2009 <<http://www.tradersmagazine.com/news/high-frequency-trading-tough-game-104672-1.html>>

² Sal L. Arnuk and Joseph Saluzzi, "Toxic Equity Trading Order Flow on Wall Street: The Real Force Behind the Explosion in Volume and Volatility," Themis Trading LLC, December 17, 2008, November 30, 2009 <http://www.themistrading.com/article_files/0000/0348/Toxic_Equity_Trading_on_Wall_Street_12-17-08.pdf>

³ "A Report on Information Arbitrage (IA) & Its Impact on Execution Quality," Electronic Trading Desk, Jefferies & Company, November 3, 2009, Page 2

⁴ Rob Curran, "Measuring Arbitrage in Milliseconds," *The Wall Street Journal Online*, March 9, 2009, November 30, 2009 <<http://blogs.wsj.com/marketbeat/2009/03/09/measuring-arbitrage-in-milliseconds>>

⁵ Ivy Schmerkin, "High-Frequency Trading Controversy Continues to Swirl," *Wall Street & Technology*, November 23, 2009, November 30, 2009 <<http://www.wallstreetandtech.com/electronic-trading/showArticle.ihtml?articleID=221600663>>

year, that's \$1.5-\$3 billion in profit generated from traditional institutional and retail investor assets under management.

The Speed Game

Today, the bulk of trading is not done by professionals analyzing fundamentals, charts or patterns. It is not done by floor traders, agency traders, sell side block traders or institutional buy side traders. Instead, it is done by proprietary trading firms, where computers analyze quote changes and trade prints. These firms need to receive and analyze an immense amount of data as fast as possible, and they need to generate and deliver orders to market centers equally as fast.

As a result, latency / co-location has become one of the fastest growing mini-industries on Wall Street. Latency has been steadily decreasing as hardware, software and networking have improved and through the isolation of inefficiencies in circuits and cabling. There is now a wide variety of consultants available to develop ways for corporations and trading firms to reduce latency from endpoint to endpoint. There also is a continuous need to upgrade equipment.

Physical laws apply; reducing this distance can accelerate data input and output. Thus, HFT firms need to be as close as possible to all trading destinations. Jefferies estimates co-location creates a "100-200 millisecond advantage – over a regular vendor based market data provider."⁶ This is why the NYSE is constructing a 400,000 square foot facility in Mahwah, NJ and another one outside of London, at a combined cost of \$500 million.⁷ This is why NASDAQ and DirectEdge have facilities larger than multiple football fields, as well. These facilities all house servers and technology of outside trading firms, as well as the market centers' own data and matching engines.

According to a recent TABB report, "the financial services industry spends \$1.8 billion for co-location and private facilities to support fast direct access to market centers. Broker-dealers account for half...exchanges 23%, proprietary trading firms 13%, asset managers 10% and hedge funds 4%."⁸

Questions

We believe Latency Arbitrage raises three serious questions about market integrity.

1. The primary response from HFTs or market centers is typically "a penny or two should not matter to long term investors; this is much ado about nothing," to paraphrase the CEO of a major ATS who was addressing a financial industry conference in New York City in early November. We disagree completely. It isn't \$0.01-\$0.02. It's \$1.5-\$3 billion. Which leads us to question number one: *Do HFT firms have an unfair advantage?* Most professionals on Wall Street have taken a standard from our past for granted, that everyone sees the same quote and market data at the same time. What if the time differential between what the HFTs see and what everybody else sees was 5 minutes instead of 5 milliseconds? Would that be acceptable? It is not the amount of time that matters. *It's that a differential exists at all.* Who would bet on a horse race if a select group already knew who won?

It is interesting to note that some of the exchanges make sure that each co-located customer receives equal amounts of connecting cable, so that a server at the northeast corner of a facility has the same latency as one at the southwest corner. It appears that "fairness" and the equalization of market data speed *among* co-located firms is an important "must" for the exchanges, but not so when it comes to all other institutional and retail investors.

⁶ "A Report on Information Arbitrage (IA)," 2

⁷ Scott Patterson And Serena Ng, "NYSE's Fast-Trade Hub Rises Up in New Jersey," *The Wall Street Journal*, July 30, 2009, Page C1, November 30, 2009 <<http://online.wsj.com/article/SB124890969888291807.html>>

⁸ Mehta

2. Latency Arbitrage has created a two-tiered market of technology enhanced insiders (comprised of a handful of large banks, brokerage firms and hedge funds) and the rest of us. To be clear, HFTs only enjoy this advantage because market centers are selling them the right to co-locate and access raw data feeds. As for-profit organizations, market centers are incentivized to do this. Which leads us to question number two: *Is it fair to sell these rights to the highest bidders when market centers are supposed to be protecting all participants' interests equally?* At the end of the day, aren't market centers charging HFTs a higher fee in exchange for giving them an advance look at the NBBO?
3. It is entirely one thing for an HFT firm to use proprietary algorithms to try to predict how an institution's algo will operate, so that the HFT can out-maneuver the institution. And it is the buy side trader's fiduciary responsibility to protect his/her firm's orders by adjusting execution methods and tactics regularly, in order to avoid predictability. Jefferies notes in its report that it "has made a number of adjustments to our algorithms to counter these issues."⁹ Which leads us to question number three: *When a market center provides an HFT with the ability to out-maneuver institutional orders, is not the exchange putting institutions and their brokers in breach of their fiduciary responsibilities, especially those institutions managing pension funds governed by Employee Retirement Income Security Act (ERISA)?*

Growing Institutional Concern

Institutional investors are slowly growing concerned about HFT.

- **Polls:** A recent Greenwich Associates survey that found that 45% of participating institutions believe HFT poses a threat to the current market structure, while 36% believe it benefits the market and investors by increasing overall liquidity. The balance say they do not know enough to judge.¹⁰
- **Research:** Quantitative Services Group LLC (QSG), a leading provider of equity research and trading analytics to institutional investors, completed a study that found significantly higher impact costs and trading velocity are incurred for VWAP algorithms when compared to Arrival Price algorithms, especially when applied to liquid, low price stocks. "This is contrary to the popular perception that VWAP strategies reduce trading costs through passive, less detectible, order placement strategies," QSG said. "The results suggest that High Frequency Trading strategies are materially contributing to these increased costs."¹¹ In addition to Jefferies, BMO Capital Markets has published research exploring the negative aspects of HFT.¹²
- **Media:** Trade media, such as *Securities Industry News*, have published by-lined articles, like "The Un(?)fair Advantage of Latency Arbitrage," by Ralph Frankel, CTO of Solace Systems.¹³ As a result, the mainstream media is also becoming more aware of HFT. For example, *The New York Times*, published a front page story on July 24, 2009, titled, "Stock Traders Find Speed Pays, in Milliseconds,"¹⁴ and the *Financial Post* in Canada on November 7, 2009, ran a story titled, "Traders of the Shadows."¹⁵

⁹ "A Report on Information Arbitrage (IA)," 7

¹⁰ "High-Frequency Trading: Lack of Data Means Regulators Should Move Slowly," Greenwich Associates, October 21, 2009, November 30, 2009 <http://www.greenwich.com/WMA/in_the_news/news_details/1,1637,1851,00.html?>

¹¹ "QSG® Study Proves Higher Trading Costs Incurred for VWAP Algorithms Vs. Arrival Price Algorithms, High Frequency Trading Contributing Factor," Quantitative Services Group LLC, November 24, 2009, November 30, 2009 <<http://www.qsg.com/PDFReader.aspx?PUBID=722>>

¹² "The Impact of High Frequency Trading on the Canadian Market," Quantitative Execution Services, BMO Capital Markets, July 22, 2009

¹³ Ralph Frankel, "The Un(?)fair Advantage of Latency Arbitrage," *Securities Industry News*, July 24, 2009, November 30, 2009 <http://www.securitiesindustry.com/issues/19_100/-23732-1.html>

¹⁴ Charles Duhigg, "Stock Traders Find Speed Pays, in Milliseconds," *The New York Times*, July 23, 2009, Page A1, November 30, 2009 <http://www.nytimes.com/2009/07/24/business/24trading.html?_r=1&scp=1&sq=%22Stock%20Traders%20Find%20Speed%20Pays,%20in%20Milliseconds%22&st=cse>

¹⁵ Karen Mazurkewich, "Traders of the shadows," *Financial Post*, November 06, 2009, November 30, 2009 <<http://www.nationalpost.com/related/topics/story.html?id=2194323>>

We believe that as more organizations explore HFT and its implications, awareness and concern among institutional investors, especially those responsible for public and private pension funds, will continue to gain traction.

Conclusion

Many professionals believed that Latency Arbitrage referred to the race by HFT firms to be faster than other traders in delivering their orders. We cared little about their technology war, as we thought their game was divorced from executing orders for traditional institutional and retail investors. We always thought HFT advancements in technology would benefit the market in general, as front end functionality, speed, and efficiency filtered down. We have since found this not to be entirely true.

HFTs use Latency Arbitrage to re-engineer the NBBO from end sources directly versus relying on the publically available standard SIP quote. HFTs are able to do this by paying exchanges and ATSS for the right to locate their servers next to market center data servers and matching engines and the right to access raw data feeds. As a result, HFTs know with near certainty what the market will be milliseconds ahead of everybody else – valuable knowledge that HFTs take advantage of when they trade thousands of stocks, thousands of times, every trading day. For HFTs, it is like shooting ducks in a barrel of honey. For all other institutional and retail investors, it is death by a thousand cuts.

To date, this situation has been tolerated because most investors were unaware that two different quotes existed and could not fathom that those in charge of overseeing the markets would allow this to happen. Why would most investors assume that they looked at different quotes? Why would most investors assume that they weren't watching the same horse race? Now we are at a crucial juncture in terms of our financial market's structure. The time has come for us to ask, "Is short-term fleeting liquidity (in our most liquid names) worth the tradeoff of accepting a multi-tiered and unequal market?"

Joseph Saluzzi (jsaluzzi@ThemisTrading.com) and Sal L. Arnuk (sarnuk@ThemisTrading.com) are co-heads of the equity trading desk at Themis Trading LLC (www.themistrading.com), an independent, no conflict agency brokerage firm specializing in trading listed and OTC equities for institutions. Prior to Themis, Mr. Arnuk worked for more than 10 years at Instinet Corporation, where he headed the team responsible for equity sales and trading for major institutional money managers. Mr. Saluzzi worked for more than nine years at Instinet, where he headed the team responsible for equity sales and trading for major hedge fund accounts. For more information, call 973-665-9600.