# Recommendations for Equitable Allocation of Trades in High Frequency Trading Environments[1]

John McPartland

July 10, 2014[2]

## *Executive Summary*

Most industry observers and much of the academic research in this area have concluded that high frequency trading (HFT) is generally beneficial. Many institutional investors, however, argue that HFT places them at a competitive disadvantage.[3] Digital computers will always have some structural (speed) advantages over human traders. This is inevitable.

This paper 1) acknowledges and summarizes much of the relevant published research,[4] 2) discusses some of the HFT strategies that likely run counter to good public policy, and 3) makes nine recommendations that, if implemented, would likely restore the perception of fairness and balance to market

[1] The author is a senior policy advisor in the Financial Markets Group of the Federal Reserve Bank of Chicago. He wishes to acknowledge the very significant contributions that David Marshall and Rajeev Ranjan made to this paper. He also wishes to thank the many industry professionals who reviewed the document and its predecessor prior to publication. Any opinions expressed in this paper are those of the author, and those opinions do not necessarily reflect the opinions of the Federal Reserve Bank of Chicago or the opinions of the Federal Reserve System.

[2] This policy paper expands upon the identically titled prior work, dated July 25, 2013.

[3] Andrew M. Brooks, 2012, "Computerized trading: What should the rules of the road be?," testimony of vice president and head of U.S. equity trading, T. Rowe Price Associates, Inc., before the United States Senate, Committee on Banking, Housing, and Urban Affairs, Subcommittee on Securities, Insurance, and Investment, September 20, available at www.banking.senate.gov/public/index.cfm?FuseAction=Files.View&FileStore_id=4ce0eb65-ae54-45ab-82fa-072c3ee7236f. See also Charles Schwab Corporation, 2014, "High-frequency trading is a growing cancer that needs to be addressed," company statement, San Francisco, April 3, available at www.aboutschwab.com/press/issues/statement_on_high_frequency_trading

[4] See, for example, Anton Golub, 2011, "Overview of high frequency trading," Manchester Business School, April 15, and Investment Industry Regulatory Organization of Canada, 2012, "The HOT Study: Phases I and II of IIROC's study of high frequency trading activity on Canadian equity marketplaces," report, Toronto, December 12.

participants that would be willing to expose their resting orders to market risk for more than fleeting milliseconds.

Readers should avoid the tendency to review this working paper only within the framework of their own nationality and market domain. The paper is meant to be global in scope. Some HFT practices that may be inappropriate (or banned) in some markets in some countries are alive and well in other markets in other countries.

An exceptionally abbreviated summary of the nine recommendations follows.

1. Where appropriate, utilize a new trade allocation formula that is intermediate between the Pro Rata trade allocation formula and the Price/Time or FIFO (First In, First Out) trade allocation formula.
2. Create a new, optional, term limit order type that, as part of the trade allocation process, would reward traders for the time that their orders are committed to be resting in the order book.
3. Completely dark orders or the hidden portion of resting orders that are not fully displayed (lit) in the order book should go to the very end of the queue (within limit price) with respect to trade allocation.
4. Prior to trade allocation, multiple small orders from the same legal entity entered contemporaneously for the sole purpose of exploiting the rounding conventions of a trading venue should first be aggregated as a single order and should carry the lowest allocation priority time stamp of all of the orders so aggregated.
5. Rather than running a continuous trade match, trading venues should divide their trading sessions into discrete periods of one half second. At a completely random time within each half second period, the single-price market-opening trade match and trade allocation algorithms should be run once.
6. Visibility into the order book should be no more granular than aggregate size at each price point. Market participants should not be able to view the size of individual orders or any other identifiers of any orders of others. This more granular information is not information that any market participant needs to make a fully informed economic decision as to the instantaneous value of the financial instrument being traded.
7. Under normal operating conditions, no market participant should be permitted to cancel an order before first obtaining an acknowledgement

from the trading venue that the original order was received.[5] We can envision no legitimate trading strategy where the practice of cancelling an order in this way would be necessary and any number of intentionally deceptive trading strategies where it would.

8. Each automated trading system (each individual algorithm) that has the capacity to generate, modify, or cancel orders without human intervention should have a unique identifier. That unique identifier must be known to every trading venue where the trading system can direct, modify, or cancel an order. Trading venues must ascribe the unique identifier as a critical information element of all relevant orders and matched trades throughout the audit trail.
9. Relevant authorities should assess and, if appropriate, seek public comment on precisely when trade information becomes generally available to the public at large. Organizations that colocate in the data centers of trading venues should not be receiving trade information from the trade match engines but should be receiving such information from the same ticker plants from which the general public receives trade information. The issue is whether some firms have access to—and can trade on—information that has not yet reached the public domain.

## *Background*

Some twentieth-century financial markets had their origins in physical trading halls. The design of the physical trading floors and the rules of these exchanges provided the exchange members with a time, place, and informational advantage over the order flow. In turn, members, specialists, or market makers were expected to maintain continuous auction markets (presumptive responsibility[6]). By the 1990s, open outcry markets had largely given way to modern screen-based electronic markets—so-called click trading. Before click trading had largely given way to today's automated markets, no single class of market participants had a time, place, or informational advantage over all other classes of market participants. All market participants

[5] This assumes that the trading venue is not experiencing technical difficulties that would prevent it from promptly sending drop copy confirmations to market participants, confirming receipt of orders.

[6] U.S. Commodity Futures Trading Commission, Technology Advisory Committee, Market Access Subcommittee, 2002, "Best practices for organized electronic markets," final report, Washington, DC, April, p. 4.

enjoyed an equal opportunity to buy at the bid and sell at the offer—and to do so anonymously.

As algorithmic trading became far more prevalent, investment managers increasingly discovered that the market neutrality of the click trading era had been lost; that today's algorithmic traders, or algo traders had assumed a dominant market making role; and that role and its twenty-first-century version of presumptive responsibilities came with a time, place, and informational advantage. While some investment managers might have thought that the phenomenon of market neutrality had been taken from them, market neutrality was never theirs in the first place.

Algorithmic trading is quite simply more competitive, and it has changed the landscape and structure (and the public perception) of today's modern financial markets. In some sense, today's perception that today's markets may be unfair seems to be associated with the "loss" of the market neutrality that was present during the click trading era.

Many industry observers seem to believe that HFT offers many benefits to organized financial markets and to society, including improved liquidity, tightened bid/ask spreads, and a decrease in intraday price volatility. This working paper describes some of the HFT techniques that have developed in electronic markets around the world, as well as their effects.

Different financial centers have different rules and regulations regarding the appropriateness of some HFT techniques. This working paper is intended to be global in its scope and in its recommendations. All of its nine recommendations might not be appropriate for every electronic trading venue in every financial center. Throughout the working paper, when discussing different trade allocation methodologies, we refer to "shares," "futures," and "lots," which are three terms we use interchangeably.

## *Review of the Academic Literature*[7]

Brogaard, Hendershott, and Riordan (2013) analyzed NASDAQ and NYSE high frequency trading data[8] that show high frequency traders increase price

[7] See Investment Industry Regulatory Organization of Canada (2012, appendix A, pp. 51-56).

[8] The HFT data represent a sample of 120 randomly selected stocks listed on NASDAQ and NYSE for all of 2008 and 2009. Trades are time-stamped to the millisecond and identify the liquidity demander and supplier as a high frequency trader or non-high-frequency trader.

efficiency by trading in the same direction of permanent price changes and trading in the opposite direction of transitory pricing errors on normal trading days and on days with the highest price volatility. In contrast, liquidity-supplying nonmarketable orders executed via HFT are adversely selected in terms of the permanent and transitory components as these trades are in the direction opposite to permanent price changes and in the same direction as transitory pricing errors. HFT predicts price changes in the overall market over short horizons measured in seconds. HFT is correlated with public information, such as macro news announcements, marketwide price movements, and limit order book imbalances.[9]

Jones (2013) notes that the volume of HFT has increased sharply over the past several years, has reduced trading costs, and has steadily improved liquidity. The main positive is that HFT can intermediate trades at lower cost. However, HFT speed could disadvantage other investors, and the resulting adverse selection could reduce market quality. Ideally, research in this area should attempt to determine the incremental effect of HFT beyond other structural and technological changes in equity markets. The best papers for this purpose attempt to isolate market structure changes that facilitate HFT. Virtually every time a market structure change results in more HFT, liquidity and market quality have improved because liquidity suppliers are better able to adjust their quotes in response to new information. Jones cites the concern that HFT may not help to stabilize prices during unusually volatile periods and notes that there is a potential for an unproductive arms race among HFT firms for speed.[10]

Cartea and Panalva (2012) conclude that the presence of high frequency traders increases the price impact of liquidity trades and that this price impact increases as the size of the trades increase. High frequency traders increase microstructure noise of prices and increase trading volume. High frequency traders and non-high-frequency professional traders coexist as competition drives down profits for new HFT entrants while the presence of high frequency traders does not drive out traditional professional traders. Finally, the paper concludes that high frequency traders clearly generate costs, but they also

[9] Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan, 2013, "High frequency trading and price discovery," University of Washington, University of California, Berkeley and University of Ontario Institute of Technology, working paper, April 22.

[10] Charles M. Jones, 2013, "What do we know about high-frequency trading?," Columbia Business School, research paper, No. 13-11, March 20.

generate benefits, and that the net effect requires more precise empirical analysis.[11]

The Litzenberger et al. (2010) paper concludes that overall market quality has improved significantly, including bid/ask spreads, liquidity, and transitory price impacts (measured by short-term variance ratios). Studies using proprietary, exchange-provided data that identify the trades of high frequency trading firms show that HFT firms contributed directly to narrowing bid/ask spreads, increasing liquidity, and reducing intraday transitory pricing errors and intraday volatility.[12]

Wah and Wellman (2013) evaluate allocative efficiency and market liquidity arising from simulated order streams in fragmented financial markets. They find that market fragmentation and the presence of a latency arbitrageur reduce total surplus and impact liquidity negatively. By replacing continuous trade matching with periodic batch auctions or call markets, latency arbitrage opportunities are eliminated and further efficiencies are achieved by aggregating orders over short time periods.[13]

Budish, Cramton, and Shim (2013) use actual millisecond quotation data to show that the prices of related financial instruments are highly correlated at human-scale time horizons but that these correlations break down completely at the single-digit millisecond level. The lack of price correlation at the millisecond level can be arbitraged away profitably if a market participant can act faster than other market participants similarly engaged in latency arbitrage. Their theoretical model shows that that quest for speed is not only wasteful but can lead to wider bid/ask spreads and thinner markets for fundamental investors than would be otherwise. They then use their model to show that frequent batch auctions can reduce the value of tiny speed advantages because it forces completion that was previously based on speed into competition to be based on price instead. They conclude that frequent

[11] Álvaro Cartea and José Penalva, 2012, "Where is the value in high frequency trading?," University College London and Universidad Carlos III, Madrid, working paper, February 17.
[12] Robert Litzenberger, Jeff Castura, Richard Gorelick, and Yogesh Dwivedi, 2010, "Market efficiency and microstructure evolution in U.S. equity markets: A high-frequency perspective," RGM Advisors LLC, working paper, August 30.
[13] Elaine Wah and Michael P. Wellman, 2013, "Latency arbitrage, market fragmentation, and efficiency: A two-market model," *EC '13: Proceedings of the 14th ACM Conference on Electronic Commerce*, New York: ACM, Inc., pp. 855-872.

batch auctions can lead to narrower bid/ask spreads, deeper markets, and greater social welfare [14]

## *Questionable HFT Techniques*

Notwithstanding the evident benefits of HFT in electronic markets, many market participants have argued that some HFT practitioners utilize trading techniques that are detrimental to the well-functioning of financial markets.[15] Some of the trading techniques generally considered to be detrimental and not capital formative are *spoofing, layering*, and *quote stuffing*.[16]

*Spoofing* and *layering* are not at all unique to HFT. Both almost always involve feigning to be a buyer when one is really a seller or vice versa. Algorithmic HFT has, however, allowed these two strategies to be taken to new levels. FINRA states the following about spoofing and layering:

> Generally, spoofing is a form of market manipulation which involves placing certain non-*bona fide* order(s), usually inside the existing National Best Bid or Offer (NBBO), with the intention of triggering another market participant(s) to join or improve the NBBO, followed by canceling the non-*bona fide* order, and entering an order on the opposite side of the market. Layering involves the placement of multiple, non-*bona fide*, limit orders on one side of the market at various price levels at or away from the NBBO to create the appearance of a change in the levels of supply and demand, thereby artificially moving the price of the security. An order is then executed on the opposite side of the market at the artificially created price, and the non-*bona fide* orders are immediately canceled.[17]

---

[14] Eric Budish, Peter Cramton, and John Shim, 2013, "The high-frequency trading arms race: Frequent batch auctions as a market design response," University of Chicago Booth School of Business and University of Maryland, working paper, December 23.

[15] See, for example, German Federal Ministry of Finance, "Speed limit for high-frequency trading–Federal government adopts legislation to avoid risks and prevent abuse in high-frequency trading," press release, Berlin, September 26, available at www.bundesfinanzministerium.de/Content/EN/Pressemitteilungen/2012/2012-09-26-speed-limit-for-high-frequency-trading.html, and Australian Securities & Investments Commission (ASIC), 2012, "Australian market structure: Draft market integrity rules and guidance on automated trading," consultation paper, No. 84, Victoria, Australia, available at www.asic.gov.au/asic/pdflib.nsf/LookupByFileName/cp184-published-13-August-2012.pdf/$file/cp184-published-13-August-2012.pdf.

[16] See the proposed amendments to the ASIC Market Integrity Rules (ASX Market) to preclude market misconduct, manipulation or false trading in Australian Securities & Investments Commission, 2013, "Dark liquidity and high-frequency trading," report, No. 331, Victoria, Australia, March, p. 10.

[17] Financial Industry Regulatory Authority (FINRA), 2012, "FINRA joins exchanges and the SEC in fining Hold Brothers more than $5.9 million for manipulative trading, anti-money laundering,

Quote stuffing is unique to algorithmic HFT. Regarding this third dubious technique, Egginton, Van Ness, and Van Ness (2012) state the following:

> Quote stuffing is a practice in which a large number of orders to buy or sell securities are placed and then canceled almost immediately. During periods of intense quoting activity stocks experience decreased liquidity, higher trading costs, and increased short term volatility.[18]

Imagine that you are bidding at an art auction, and the serious bidders are now reduced to two or three. One of the persons pretending to be an interested bidder is really the owner of the art piece currently being auctioned off. It is to their advantage to get the bona fide bidders to pay as much as possible for their art piece. Bidders indicate their willingness to bid to the auctioneer by raising the bidder numbers assigned to them by the auction house. The *spoofing* equivalent in this physical environment would be if the seller of the art piece, pretending to be a buyer, raised his or her bidder number one last time, solely to get the last remaining buyer to pay more than they otherwise would be willing to pay. Granted, the *spoofer* in this case is absolutely at risk of buying their own art piece unless their *spoofing* strategy is successful and a bona fide bidder betters the *spoofer's* bid. While the practice of allowing sellers to masquerade as buyers is probably not allowed at proper art auctions, its electronic equivalent is permitted and well practiced among some HFT practitioners. Make no mistake, HFT *spoofers'* bids and offers are exposed to market risk as much as the bids and offers of click traders, even if they are often so exposed for only milliseconds. Spoofing is intentionally designed to be deceptive and, at a minimum, frustrates fair value investors' ability to determine the true market value of the instruments that are being traded.

*Layering* is only a slightly different technique, designed to similarly deceive market participants' perception of the aggregate size of the bids and offers in the order book. By entering thousands of bids or offers, and then cancelling them virtually immediately, but only after they have been acknowledged as having been present in the order book, HFT practitioners can successfully create the illusion of greater size at the bid (or offer) than is realistically executable. Investment managers often refer to this phenomenon as "phantom liquidity" as the visible liquidity is often not there when one goes to hit the bid

---

and other violations," press release, Washington, DC, available at www.finra.org/Newsroom/NewsReleases/2012/P178687.

[18] Jared Egginton, Bonnie Van Ness, and Robert Van Ness, 2012, "Quote stuffing," Louisiana Tech University and University of Mississippi, March 15.

or lift the offer. Not unlike the massive white clouds in the sky, they are actually nothing more than thin water vapor that simply gives the illusion of being huge, massive objects. Frequently, high frequency traders *layer* quotes on the bid side of the market, in an attempt to attract other bidders, and then hit the bid side of the market as a seller in size. *Layering* is designed to be intentionally deceptive and similarly frustrates fair value investors' ability to ascertain the fair market value of the instruments traded. It also intentionally and unduly complicates order execution.

*Quote stuffing* is roughly equivalent to driving a race car at 190 miles per hour, but preventing the other drivers from exceeding 160 miles per hour. By clogging a trading venue's outbound quotation system (or inbound order entry systems) with near worthless quotes, astute HFT practitioners can execute trades on another trading venue or on the same trading venue with some degree of confidence that at least a plurality of market participants (including many other high frequency traders) will, at best, be reacting to delayed quotes, creating an arguably unfair trading advantage for these HFT practitioners that can "slow down" the other traders by relatively increasing their own reaction times.[19] As CNBC noted in 2012, "the ultimate goal of many of these programs is to gum up the system so it slows down the quote feed to others and allows the computer traders (with their colocated servers at the exchanges) to gain a money-making arbitrage opportunity."[20] Price transparency is considered a public benefit of organized financial markets. It is difficult to envision that the practice of intentionally slowing down the dissemination of trade prices to the public is an activity that serves the public interest.

The thesis of this paper is that, rather than attempting to ban these techniques (which could likely be difficult to enforce in practice), one could change the character and economics of the trading environment so as to disincentivize these and similar undesirable trading techniques. Rather than propose solutions that might preclude specific HFT strategies, we propose to simply change the economics of the trading environment by modifying the criteria of order allocation priority and by discouraging certain questionable industry

[19] Quote stuffing is an offensive tool that high speed traders most typically use to gain a competitive advantage over other high speed traders. Click traders would not likely be adversely affected if outbound quotations were intentionally delayed by, say, 200 milliseconds, nor would they likely even be able to detect any such delay.

[20] John Melloy, 2012, "Mysterious algorithm was 4% of trading activity last week," CNBC, October 8, available at www.cnbc.com/id/49333454.

practices to strike a more equitable balance between the high frequency trading community and the investment management community.

## *Recommendations*

The proposal consists of nine recommendations that should be deemed as one complete set that should be considered and implemented as a whole, where appropriate. Several of the recommendations are admittedly rather complex, but so are the current electronic market structures in which we find ourselves. Our recommendations follow.

### 1. Trade Allocation with Cardinal Weighting of Time in the Order Book

The ideal trade allocation algorithm should be a combination of the Pro Rata trade allocation algorithm and the Price/Time or FIFO trade allocation algorithm.[21] Descriptions of the Price/Time and Pro Rata algorithms would seem to be in order.

*Price/Time or FIFO*

The Price/Time trade allocation algorithm is also known as the FIFO algorithm. The Price/Time trade allocation algorithm first prioritizes all bids and orders based on price, and within price, prioritizes orders (in an ordinal ranking) based on the time that each order was received. An order can always gain priority by bettering its price, while keeping its original time stamp. Within the best price, the Price/Time algorithm attempts to completely fill the order with the oldest time stamp, (the lowest ordinal ranking) with any residual contracts or shares subsequently allocated to the next oldest bid or offer, until the appropriate contracts or shares have been fully allocated.

The Price/Time trade allocation algorithm was the first algorithm utilized when the era of electronic trading was ushered in. Some market participants erroneously think that electronic markets still utilize the Price/Time trade

[21] There are at least half a dozen other trade allocation algorithms currently in use but not specifically referenced in this section. While it might be quite valuable for interested market participants to have a detailed treatise on the various trade allocation algorithms currently in use, that is not the objective of this section.

allocation exclusively—that is, that there *are* no other trade allocation algorithms. While it is equitable, some trading venues have diversified away from the Price/Time trade allocation as market participants tend to feel disconnected when they join the bid or offer, but are not senior enough to participate in any trade allocation. If there is a valid criticism of the Price/Time trade allocation algorithm, it is that it allocates trades based only on a simple ordinal ranking of bids or offers based on their respective time stamps. Basing the allocation of trades on a cardinal weighting (ranking) of trades based on their actual time stamps would seem to be a superior approach.

*Pro Rata*

In the Pro Rata trade allocation algorithm, all bids are allocated their pro rata share of the allocation of a matched trade based solely upon the lot size of their respective resting bid relative to the aggregate sum of all of the resting bids at the same price. For example, if there are a total of 2200 lots bid for at 12 and 220 offers hit the bid, each resting bid would be allocated 10% of the lot size for which they were bidding.[22]

If there is a criticism of the Pro Rata trade allocation logic, it is that many market participants are constantly bidding or offering unrealistically large quantities, often far greater than they could likely realistically absorb.

*NYSE/Liffe*

NYSE/Liffe has a hybrid trade allocation algorithm that assigns resting bids and offers with an ordinal ranking (based on their time stamp) and then allocates trades based on a combination of the Pro Rata approach and the ordinal ranking of the bids and offers.

In the formula that is equation (1), the first bracketed expression simply says that a market participant should be allocated the lesser of 1) the full amount of the quantity of his order or 2) a lesser quantity based upon where his respective order ranks in the order book, based on its time stamp, relative to

[22] This trade allocation process has been shown to be prone to the apparently unavoidable rounding error chicanery that occurs when dozens and dozens of one lot orders are intentionally entered by a single market participant, in the hope of being unjustly enriched by the trading system rounding of what would have been anything greater than 51/100ths of a futures contract or share to one full contract or share. See recommendation 4.

the time stamps of the other orders in the order book. The second bracketed expression determines the pro rata quantity of any given order relative to the aggregate quantity of orders at the same price. The third bracketed expression determines the ordinal ranking (by time stamp) of any given order relative to the time stamps of all of the other orders at the same price in the order book. It is this third bracketed expression that we believe could be improved.

NYSE/Liffe Time Pro-Rata algorithm[23]

$$A_n = Min\left( v_n, \frac{f_n}{\sum_{r=1}^{N} f_r} * L \right) \quad \text{where,} \quad f_n = \left[ \frac{v_n}{\sum_{r=1}^{N} v_r} \right] * \left[ \frac{(N+1)-n}{\sum_{r=1}^{N} r} \right] \qquad (1)$$

*N* - Total number of resting buy (sell) orders sorted by time, n = 1(oldest) to N (newest)
*n* - Individual order being considered
*r* - Ascending sequence, 1 to N
$A_n$ - Allocation for resting buy (sell) order, n
$v_n$ - Volume of resting order being considered, n
$f_n$ - 'Time Pro Rata Factor' calculated for resting buy (sell) order being considered, n
*L* - Incoming sell (buy) order volume

Recommended Trade Allocation Algorithm

When allocating trades, the instant proposal places a greater weighting on the time that an order is exposed to market risk. We extrapolate from the NYSE/Liffe model and assign a cardinal ranking, rather than an ordinal one,[24]

[23] Fractional allocations are rounded down to the nearest integer for all allocations greater than 1 and rounded up to 1 for all fractional allocations less than 1. For equally sized fractional allocations, priority is granted to the oldest order. If any volume remains unallocated following this sequence (for instance, as a result of rounding or when the calculated allocation for an order is constrained by the Min function in the NYSE/Liffe Time Pro-Rata algorithm), then a further pass of the sequence will occur.

[24] Ordinal ranking of resting orders involves creating a simple ranking, not unlike athletes who finish first, second or third in an athletic contest. That is, no consideration is given to the *difference* in athletic performance between the first and second finisher and the second and third finisher. A cardinal ranking would involve assigning a numeric value to the performances of the athletes, not unlike in gymnastics, where there is a quantitative evaluation of individual performances. In the instant proposal, our recommendation is to allocate matched trades based on the actual time that the orders have been resting in the order book relative to the times that

to resting bids and offers based on the actual length of time that bids and offers have been resting in the order book, relative to the time that all of the other orders have been resting in the order book. This is accomplished by raising "time in the order book" (Tau) to a low but effective exponential power (α).

$$A_n = Min\left( v_n, \frac{f'_n}{\sum_{r=1}^{N} f'_r} * L \right) \quad \text{where,} \quad f'_n = \left[ \frac{v_n}{\sum_{r=1}^{N} v_r} \right] * \left[ \frac{\tau_n^{\alpha}}{\sum_{r=1}^{N} \tau_r^{\alpha}} \right] \qquad (2)$$

$\tau$ - Time duration (in milliseconds) for every resting order (time difference between the time of trade match and an incoming order's time stamp)

$N$ - Total number of resting buy (sell) orders sorted by time, n = 1(oldest) to N (newest)

$n$ - Individual order being considered

$\alpha$ - A constant parameter set by the trading venue

$A_n$ - Allocation for resting buy (sell) order, n

$v_n$ - Volume of resting order being considered, n

$f'_n$ - 'Proposed Time Pro Rata Factor' calculated for resting buy (sell) order being considered, n

$L$ - Incoming sell (buy) order volume

Note that equation (2) is identical to equation (1) except for the third bracketed expression. Whereas the third bracketed expression in (1) is based on a simple ordinal ranking of time in the order book, the third bracketed expression in (2) raises time in the order book ($\tau$) to an exponential power ($\alpha$). Increasing $\alpha$ causes a nonlinear marginal increase in the number of lots a longer-duration order is allocated, compared with a shorter-duration order, based on the time that the order has been exposed to market risk. This additional weighting (resulting in a greater allocation of trades) could be set by the trading venue by increasing the parameter $\alpha$.

Figure 1 illustrates the progressive trade allocation results when $\alpha$ is set equal to or greater than zero and less than or equal to 2.3. As $\alpha$ is set to increasingly higher values, the weight associated with Tau (time in the order book)

---

other orders have been resting in the order book–and not based on the ordinal ranking of the respective time stamps of resting orders. Thus, in a cardinal ranking structure, an order that has been resting in the order book for four hours would be entitled to a far greater allocation of trades than an order that has been resting in the order book for only four seconds. In an ordinal ranking system, those orders would simply be ranked as #1 and #2 in priority.

increases exponentially and the actual time that an order has been resting in the order book becomes an increasingly dominant component when the algorithm allocates trades. It may be helpful to think of this recommendation as the introduction of perfect gradient shades of gray that lie between black (Pro Rata) and white (Price/Time).

***Market View – Top of Book***

Market View

| Bid Qty | Bid Prc | Ask Prc | Ask Qty |
|---|---|---|---|
| 600 | 99.69 | 99.70 | 900 |

| | |
|---|---|
| Incoming Sell Qty Order @ price 99.65 - | 300 |
| Number of orders resting at the price - | 5 |
| Incoming Order Timing - | 9.00.04.005 |

***Limit Orders – Listing all Bids at price of 99.69***

| Time | Buy/Sell | Qty | Price |
|---|---|---|---|
| 9.00.03.005 | B | 120 | 99.690 |
| 9.00.03.755 | B | 120 | 99.690 |
| 9.00.03.855 | B | 120 | 99.690 |
| 9.00.03.955 | B | 120 | 99.690 |
| 9.00.04.003 | B | 120 | 99.690 |

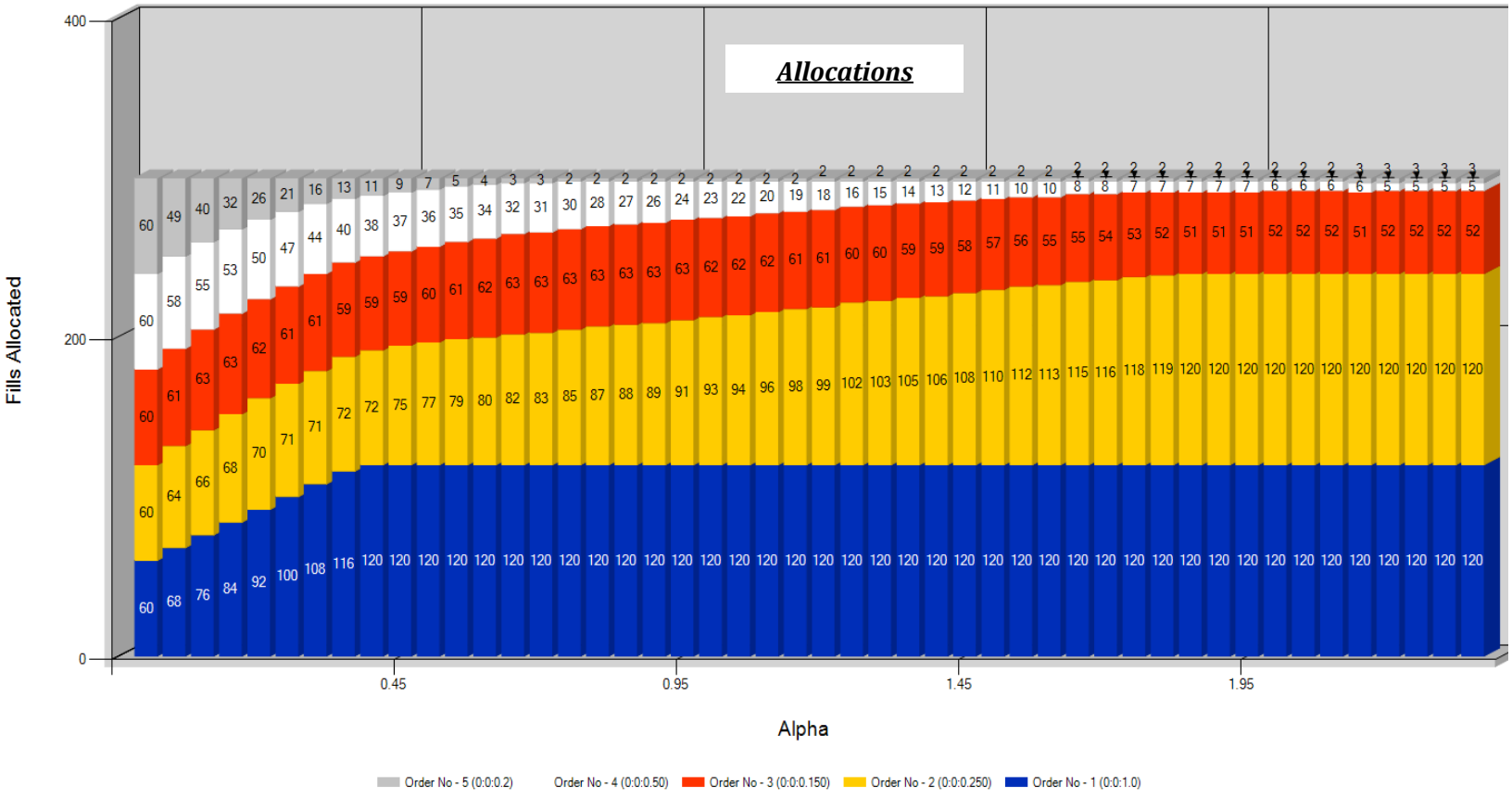


Figure – 1 (α: 0 – 2.3)

The horizontal axis reflects the value of α. The vertical axis reflects the quantities (lots) that would be allocated to resting orders based upon their respective time in the order book. When α is set to zero, each of the five resting bids would be allocated 60 lots, that is, an exact Pro Rata trade allocation (where time in the order book means nothing). As α is set to increasing levels, time in the order book receives more and more weighting. If α were set to a

value higher than 1.9, the recommended trade allocation algorithm closely approximates the Price/Time or FIFO trade allocation algorithm (where time in the order book means everything). This continuum approach would allow trading venues to select allocation outcomes of varying degrees between the Pro Rata and the Price/Time trade allocation outcomes.[25]

In the example in figure 1, there are 900 lots offered at 99.70 and 600 lots bid for at 99.69. The 600 lots on the bid side are comprised of five individual bids, each for 120 lots, all at a price of 99.69 but with different time stamps. There is a new incoming order to sell 300 lots at a price of 99.65, well below the resting bids at 99.69. The incoming order of 300 lots is therefore going to take out half of the 600 lots bid for at 99.69. The graphic demonstrates how 300 lots would be allocated to the five market participants bidding at 99.69 if $\alpha$ were set to various values between zero and 2.3.

Note that in this example, the oldest bid (in blue) resting in the order book is 750 milliseconds older than the second oldest resting bid (orange) but the remaining bids are separated by only about 10 milliseconds. The effect of this time differential can be dramatic, depending on the value of $\alpha$ that is selected by the trading venue.

If $\alpha$ were set to 0.40, the oldest resting bid would be allocated 100% of its bid quantity (120). This is largely due to the 750-millisecond time differential between the oldest bid and the second oldest bid. When $\alpha$ is set to 0.40, the remaining resting bids would be allocated 72, 59, 38, and 11 lots, respectively.

It is anticipated that, for every relevant instrument, a trading venue would select a fairly permanent value of $\alpha$ that strikes an equitable balance that rewards both liquidity providers and institutional market participants. When making such a determination, trading venues will undoubtedly consider the current preferences of market participants and the business risks and costs of changing trade allocation algorithms for legacy products.

[25] It is assumed that trading venues would change the value of $\alpha$ very infrequently, if at all, as market participants would need to have a clear understanding and expectation of the trade allocation process in order to correctly size the quantities of their bids and offers. Setting $\alpha$ to a value between zero and 2.3 would allow trading venues a continuum from which they could select the optimal trade allocation result for any financial instrument or product family.

2. Term Limit Order Type

Create a new, entirely optional noncancellable term limit order type—for example, Buy at 12, good for 4 seconds or Buy at 12, good for 4/10ths of a second. The order may not be cancelled during its stated term (see footnote 26) and would be displayed in the order book just as any other resting bid at 12. The term of the order is the *minimum* amount of time that the order would be exposed to the market. Like any other non-term-limit orders, a term limit order remains open until it is either filled (either during or after its stated term) or cancelled after its term has expired.[26] Importantly, the instant a term limit order enters the order book, it has the trade match and trade allocation priorities of having already been in the order book for the stated term of the order. For example, the instant that an order to Buy at 12, good for 4 seconds enters the order book, it would have the trade match priority and trade allocation priority identical to an order that has already been resting in the order book for 4 seconds.

This order type should have the potential to provide more balance between the interests of institutional market participants and HFT practitioners, whose orders are often in the order book for only a few milliseconds. When combined with recommendation 1, allocation of matched trades should be directly related to how long a resting order was exposed (or committed to be exposed) to market risk. Orders that are resting (or committed to be) in the order book for a material amount of time are exposed to market risk, provide tangible price transparency to the public, and deserve a higher trade match priority and trade allocation priority than orders that have barely been in the order book for a few milliseconds and have provided only marginal (if not intentionally deceptive) pricing information to the public. The combined effect of recommendations 1 and 2 should increase the likely allocation of lots[27] to orders that are exposed to market risk for a greater period of time, at the expense of orders that are exposed to market risk for only a few fleeting milliseconds.

---

[26] A trading venue might elect to allow a term limit order to be amended to a price *better* than the original price of the term limit order during the period in which the order could not otherwise be cancelled. Lot size could not be amended. This should only be allowed if the limit price of the original order was initially entered to join the best bid or offer (the "top of the book") and remained at the top of the book, should the market move. Allowing such a trade (with the better price) to keep its original time stamp would seem equitable and consistent with public policy objectives.

[27] Lots could mean shares, options, futures, swaps or any other specialized descriptor of traded financial instruments. Our intention is not to limit the scope of the applicability of the recommendations.

There is no apparent reason this optional order type could not be used in a fragmented market structure where any number of trade allocation methods are in use.[28] Where financial markets are fragmented into multiple trading venues, different trade execution venues could have different trade allocation formulas. This proposed new order type would have no effect at all on trades allocated on trading venues that use the Pro Rata trade allocation method, as time in the order book would still be given no weight. It could, however, have a material effect on trades allocated on trading venues that use the Price/Time algorithm or which might adopt the trade allocation formula that we are recommending.

Implementing any new order type would, of necessity, involve implementation costs for trading venues, trade intermediaries, and, to a lesser degree, some end-user market participants. It may be advantageous for one or more trading venues to inaugurate pilot implementation programs for a limited number of traded products to better gauge potential commercial acceptance of this concept and to make a more informed business decision regarding wider (or universal) implementation of the proposed new term limit order type.

3. Time Stamp Conventions of Dark or Unlit Orders

According to the Australian Securities & Investments Commission, "dark liquidity refers to orders that are not known to the rest of the market before the orders are matched as executed trades. Such trades, known as 'dark trades,' can occur on exchange markets … and in venues other than exchange markets."[29] Orders of all types (except those noted in recommendation 2) should have a time stamp reflecting the start of the period during which that order was continuously visible in the order book to all market participants. Said another way, an order originally entered as an unlit order should have no valid time stamp as long as that order remains unlit and not visible in the order book to all market participants. Without a time stamp, all unlit orders at a limit price should stand behind all lit orders at the same limit price with respect to trade allocation.

[28] In the fragmented U.S. equities market, Regulation NMS would still require that an order initially be routed to the trading venue with the best price. That trading venue may or may not be utilizing a trade allocation method that places a value on the time an order has been resting in the order book.

[29] Australian Securities & Investments Commission (2013, p. 12).

If the offer were to go through all valid lit bids in the order book at the best bid limit price and should there remain unlit orders resting in the order book at the same limit price, the trading venue should allocate the residual amount of lots to the unlit orders under either one of two protocols, both of which appear to be equitable. The first protocol would simply be to allocate the residual amount of unlit orders on a Pro Rata basis. The second protocol involves ranking the unlit orders by the time stamp of the lit portion of their respective orders. Once in the proper sequence, lots are allocated to those orders but only for the quantity specified in the lit portion of each order. This process would continue iteratively until all of the lots had been allocated or until there were no longer any remaining unlit orders that had not been fully allocated. Following either of these protocols should be consistent with principle 3 of the of the 2011 report on dealing with dark liquidity by the Technical Committee of the International Organization of Securities Commissions:

> Principle 3: In those jurisdictions where dark trading is generally permitted, regulators should take steps to support the use of transparent orders rather than dark orders executed on transparent markets or orders submitted into dark pools. Transparent orders should have priority over dark orders at the same price within a trading venue.[30]

It is unfortunate that some trading venues allow traders to submit orders that are not visible to others and then modify the order while retaining its original time stamp. This practice runs counter to all principles of fairness. While this recommendation would not ban dark or unlit orders, it should provide an appropriate economic disincentive to utilize them to any great extent.

4. Aggregation of Consecutive Small Orders from the Same Legal Entity

Some traders have exploited trade allocation formulas that round fractional lot allocations up to the next integer. For example, a buyer of one 100 lot order might be entitled to an allocation of 62 futures or options contracts based on the applicable trade allocation formula. If the buyer were to have entered his 100 lot order as 100 *one lot* buy orders and if the trading venue rounds fractional allocations up to the next integer, the trade allocation formula would allocate 62/100ths of a futures or options contract to each one lot order. Since the trading venue cannot allocate 62/100ths of a futures or options contract,

[30] Technical Committee of the International Organization of Securities Commissions, 2011 "Principles for dark liquidity: Final report," Madrid, Spain, May, pp. 28-29.

the trader may be unjustly enriched by a rounding convention (that rounds up fractions greater than 1/2) only because the trader entered a 100 lot order as 100 one lot orders. Taken to the extreme, the trader could be allocated as many as 100 lots (one for each one lot order) while really only being entitled to 62 lots.

Prior to allocating trades, trading venues should first aggregate all matched trades submitted by the same legal entity to mitigate the potential for gamesmanship due to rounding conventions of one lot orders.[31] The aggregation routine should run once, every time that the trade allocation algorithm is run and would involve only orders that would be entitled to a fill or partial fill and only orders that appear to have been intentionally entered sequentially, by their respective time stamps. This should leave unaffected the bona fide trades of unequal quantities entered minutes apart by the same legal entity. For example, it should be easy enough to discern between the two or three unequal resting buy orders from a grain elevator, all entered within five minutes[32] and 100 one lot orders from an algorithmic trader, all entered three milliseconds apart.

The aggregated order should be assigned the worst time stamp of all of the multiple orders that it comprises.[33] There seems to be no reason to run the aggregation routine continuously or prior to running the trade allocation algorithm (note recommendation 5). This recommended procedure should eliminate much of the potential for gamesmanship, provided that market participants do not then violate other rules of trading venues by lying about their true identity.

5. Random Timing of the Trade Match Algorithm

Trading venues should divide their trading sessions into time periods of one half of one second. At a completely random time during each one half second trading period, the trading venue should run its trade match algorithm (allocating trades utilizing the cardinal ranking of resting bids and offers as

[31] In the alternative, trading venues may elect to round down, rather than round up, allocating one lot trades that would otherwise be allocated a fractional lot, nothing, as at least one trading venue already does.

[32] The orders from the grain elevator should not be affected and should each severally retain their own original time stamp.

[33] This is not intended to affect all trades from the same legal entity. This recommendation is intended only to mitigate the unjust enrichment associated with multiple and sequential one lot and, perhaps, two lot trades.

described in recommendation 1 earlier) *once*.[34] This procedure has several advantages. Because high frequency traders would never know when the trade match algorithm would be run during any one half second time interval, the value of entering thousands of quotes for only a few milliseconds should be at least partially diminished. Additionally, as many of those quotes are typically not actually intended to be matched, they would occasionally, under this proposal, become swept up and executed in the trade match algorithm, because of its random timing within each half second period. Under this proposal, high frequency traders could continue the practice of entering thousands of quotes per second, only with more substantial potential financial implications.

There is some anecdotal evidence that suggests that most humans can read, recognize, and process two to three numerical quantities per second.[35] Dividing the trading session into half second periods should provide human institutional traders with useful visual information on their trading screens as fast as that information can reasonably be comprehended. It may be helpful to imagine that if such an electronic market had an audio attribute, the market would trade and outbound quotations would be disseminated about as fast as a professional auctioneer can speak. Budish, Cramton, and Shim (2014) have suggested that such rapid-fire quotation dissemination (for matched trades) would constitute sufficient pretrade price discovery. That is, if the market were

[34] The practice of trading venues to display the probable single opening price (as the market is about to open) usually comes with a policy that precludes the cancellation of orders within half of a second prior to the opening time. Market opening trade match algorithms determine a single price at which the maximum number of trades would optimally be matched. From a technical perspective, it will likely then be impossible to run the market opening trade match algorithm at a random time within every half second trading interval and display such a single price before the fact. The instant proposal is simply to run the trade match and trade allocation algorithms once, not necessarily with all of the bells and whistles that come with the single-price market-opening trade match algorithm. HFT market participants and click traders should be able to deduce whether the trade match algorithm will match trades at the highest bid or the lowest offer by watching the size at the bid and offer.

Alternatively, trading venues could elect to have the next subsequent one half second trading interval begin as soon as practicable after the prior trade match and trade allocation algorithms have been run for the prior trading interval. Doing so would compound the randomness of running the trade match and trade allocation algorithms. One potential downside of doing so might be that occasionally, the trade match and trade allocation algorithms could be running in excess of twice per second. Most human click traders would not be able to read, digest, and respond to price and quantity information at such a fast pace.

[35] See Kimron L. Shapiro, Karen M. Arnell, and Jane E. Raymond, 1997, "The attentional blink," *Trends in Cognitive Sciences*, Vol. 1, No. 8, November, pp. 291-296, and Jane E. Raymond, Kimron L. Shapiro and Karen M. Arnell, 1992, "Temporary suppression of visual processing in an RSVP task: An attentional blink?," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 18, No. 3, August, pp. 849-860.

to disseminate the single auction price as fast as an auctioneer can speak, market participants would likely have sufficient price discovery information to make informed economic decisions on the fair market value of the financial instrument being traded and would no longer need a view into the order book.[36]

Derivatives exchanges often group related product types—for example, equities, interest rates, foreign exchange, agriculture, energy, and precious metals—on their trade match engines. Under this proposal, derivatives exchanges would run the trade match and trade allocation cycle, by product type (on each server) at a completely random time, once during each half second time period. Doing so should preserve the so-called implied functionality execution functionality—for example, the soybean crush spread, the crude oil crack spread, or simple calendar spreads. Again, if performed properly, this should also randomize the sequence in which the servers run their product specific trade match and trade allocation cycles.

Equity trading venues also typically run multiple trade match engines—for example, all stocks that begin with the letters A through F might trade on one server. It is envisioned that such a trading venue that has *n* number of servers would run one trade match and trade allocation cycle at random times during each half second period on each server. If done properly, this would also randomize the sequence in which each of the *n* number of servers would run their respective trade match and trade allocation algorithms.

The additional computational processing that would likely be required to assimilate both the new term limit order (recommendation 2) and the allocation of lots based on a cardinal ranking rather than a simple ordinal ranking (recommendation 1) should be partially, if not totally, ameliorated by only having to run the trade match and trade allocation algorithms once every half second. Doing so should allow the trading venues to conserve significantly on network bandwidth as outbound quotations would only be disseminated once every half second—and then, only in batches.[37] One firm that was invited to comment on a prior draft of this paper indicated that implementing this recommendation could potentially free up much of its annual information

[36] See Eric Budish, Peter Cramton, and John Shim, 2014, "Implementation details for frequent batch auctions: Slowing down markets to the blink of an eye," *American Economic Review*, Vol. 104, No. 5, pp. 418-424.

[37] In theory, disseminating outbound quotations in batches might also thwart the practice of quote stuffing, as market venues would push quotation data out in data "packages," which might make quote stuffing a less effective strategy to intentionally clog up trading venues' outbound quotation systems.

technology (IT) budget—currently dedicated to buying more servers to handle the tsunami of price and quantity (and related) data newly available by the millisecond and to retransmit them just as quickly to their customers who could not possibly react to them—to instead develop truly value added features to their client user interface. It is entirely possible that these potential cost savings could similarly be realized by a broader section of market participants.

Moving modern electronic markets away from continuous trade matching to discrete auction processing might also improve the technological framework within which national supervisory authorities will be held responsible for providing supervisory market oversight now and for many years to come. Implementation of this recommendation would materially reduce the amount of quotation and trade match data that make up the audit trail for today's modern electronic financial markets.[38] Common sense would argue that the probability of achieving some success in this regulatory area might be greater if the challenges of doing so could be made less formidable.

Millisecond-by-millisecond trade matching has also driven the average lot size of equity and futures transactions down dramatically. Maintaining the IT infrastructure to process a plethora of small lot fills does greatly increase the operating expenses of trading venues, clearing organizations, and trade intermediaries because the relevant operating expenses are driven by the *number of transactions*, not the *number of shares or futures contracts* of those transactions. Processing a small lot order consumes just as much bandwidth and just as many IT resources as processing one 10,000 lot order that is matched and clears as one 10,000 lot order. A serendipitous result of implementing this recommendation could be a material reduction in the operating expenses of trading venues, clearing organizations, and trade intermediaries that must scale their infrastructure to handle the *number of transactions* that they process or retransmit.

At least two leading electronic foreign exchange trading markets have implemented processes to slow down or randomize incoming orders. ParFX currently imposes randomized pauses on incoming orders.[39] EBS accumulates orders by institution, and, after waiting for either one, two, or three

[38] Theoretically, if markets currently trade every millisecond, moving to a batch auction once every half second would reduce the volume of outbound quotation data by a factor of 500.
[39] Stephen Foley, 2013, "High-frequency traders face speed limits," *Financial Times*, April 28, available by subscription at www.ft.com/cms/s/0/d5b42402-aea3-11e2-8316-00144feabdc0.html; and Nicola Tavendale and Joy Macknight, 2014, "Will latency floors do the trick?," *Profit & Loss in the Currency & Derivatives Markets*, Vol. 15, No. 151, May, pp. 52, 54.

milliseconds, randomizes the order of the institutions. This creates a matrix of orders within an institution. After randomizing the ranking of the institutions, the trade allocation algorithm then allocates the highest priority order from each institution (one from each institution) until sufficient orders have been allocated. This process effectively diminishes the value of any speed greater than several milliseconds.[40]

Some lawmakers and regulators have suggested that quotations must be exposed to the market for a minimum amount of time.[41] Fifty to 500 milliseconds[42] is an eternity to a proprietary algorithmic trader. The likely (and logical) reaction would be for market makers to widen their respective bid/ask spreads to compensate themselves for the additional market risk to which their quotes would be exposed under any such minimum cancellation time regimes. We would argue that rather than dampen the quest for speed, a minimum cancellation time would have exactly the opposite effect. If market participants could not cancel their quotes for 500 milliseconds, a persuasive argument could be made that high-speed automated traders would be willing to pay huge sums of money to ensure that they were the very fastest, thus enabling them to pillage the quotes of slower traders who could not cancel their quotes (for 500 milliseconds) before being plundered.

Running the single-price market-opening trade match algorithm at a random time within every half second time period would provide the trading venue with the option of either providing or not providing market participants with a view into the order book prior to the trade match. A cogent argument can be made that providing single-price match trade information to the general public every half second is more than sufficient for market participants to make a fully informed economic decision as to the current fair market value of the financial instrument being traded, and is as fast as a human can comprehend such information. A suggestion that would have the equivalent effect would be to run the single-price trade match algorithm once every half second (on the half second) but not provide any visibility into the order book, before the fact. Either approach would eliminate much of the time, place, and informational advantage that high speed automated traders currently enjoy over all other classes of market participants.

[40] Ibid.
[41] This provision was ultimately not included in the German High Frequency Trading Act, referenced later in this paper.
[42] Australian Securities & Investments Commission (2013, p. 10).

Our recommendation 5 would still allow algorithmic traders to cancel their orders at any time and should thus render moot any potential arguments about having market makers' quotes unduly exposed to market risk. There is some potential that implementation of this recommendation would considerably dampen or possibly even end the incessant low latency arms race.[43]

Lastly, some algorithmic trading firms have called into question how interexchange arbitrage trades could be executed if some or all of the trading venues ran their single-price trade match algorithm at random times within one half second periods. Under those circumstances, interexchange trades would be executed the same way they were executed for decades before trading became dominated by algorithms and before trading became a millisecond-by-millisecond phenomenon. One leg of an arbitrage trade might be delayed, perhaps up to the blink of an eye before being filled.

Algorithmic trading firms seem to be of the view that the market structure of the future can only evolve from the (arguably remedial) market structure that we have today. We reject this premise. If the current public focus over the issue of "fairness" is generally reflective of the consensus of a democratic society, then why would a free and educated society knowingly limit the design for the structure of the financial markets of the future to a starting place that is considered remedial? Interexchange trading will survive and flourish as long as there is a profit motive driving it.

6. Granularity of Information in the Order Book

In general, all market participants should have access to the same information, as well as the same level of granularity of information, from the order book. Market participants should only have access to information that they legitimately need to make an informed economic decision on market depth, price, and liquidity. Market participants that have the ability to query the order book should ideally only be able to see the aggregate size at each bid and offer price points.

No market participants should be able to see any other identifying data in the order book that would reveal the identity or origin of the other market participants that have entered orders. No market participant should be able to

[43] See Chris Sparrow, 2011, "The failure of continuous markets," Market Data Authority, Tabb Forum, December 5.

see or otherwise ascertain the time stamps or the individual lot sizes of orders entered other than their own.[44] Such granular data is not information that any market participant legitimately needs to make an informed economic trading decision. It should be sufficient for trading venues to provide market participants with a graphic representation of where they stand in the order queue.

Many trading venues currently provide veritably all changes in the order book (new orders and cancelled orders) via the User Datagram Protocol (UDP). Any reasonably sophisticated trader can recreate the order book with precise detail by monitoring which trades were recently entered, their lot sizes, and their time stamps. By also capturing those orders that were cancelled, their lot sizes, and their time stamps, one can not only recreate a granular order book, but also determine the priority of their own orders and the priority and size of the orders of other market participants. Some equity venues provide some market participants with such granular information.[45] At least two major futures exchange tout their provision of such granular order book information as being fully transparent. We contend that this practice is fundamentally wrong.

One might attempt to argue that this recommendation goes against the following principle: Transparency in organized financial markets is beneficial and consistent with good public policy. We disagree with this conclusion about our recommendation. Dissemination of *granular* data from the order book allows algorithmic traders to gain inappropriate insights into the trading patterns of both algorithmic traders and click traders. That said, we have no qualms about HFT firms using *aggregated* data from the order book: High frequency traders should continue to have the unfettered ability to attempt to reverse engineer aggregated data and reach any conclusions that they may care to reach.

We note that there is an abundance of research that demonstrates a trade-off between market liquidity and transparency.

---

[44] This is based on the premise that traders join resting *prices*, not resting *times*. It may be helpful to approach the issue from the perspective of a click trader, rather than from the perspective of an algo trader.

[45] See Sal Arnuk and Joseph Saluzzi, 2012, *Broken Markets: How High Frequency Trading and Predatory Practices on Wall Street Are Destroying Investor Confidence and Your Portfolio*, Upper Saddle River, NJ: FT Press, pp. 102-103, and Charles Duhigg, 2009, "Stock traders find speed pays, in milliseconds," *New York Times,* July 23, available at www.nytimes.com/2009/07/24/business/24trading.html.

For instance, Lee (1998, p. 98–99) states:

> The choice by an exchange of what price and quote information to release is a central element of the wider decision as to what market architecture to adopt. Not only are there substantial differences between the types of data about prices and quotes that trading systems choose to release, there are also differences in the types of information that trading systems are able to deliver. … In no trading system are all these categories of price and quote information published. Indeed, the strategic non-disclosure of some types of price and quote information is a central element of all market architectures. For some of the information categories, the reason is a matter of confidentiality. In most markets, for example, investors are unwilling to countenance releasing information about what their trading policies have been or will be. The identities of market participants submitting quotes and participating in trades are therefore normally not publicly released. Sometimes, however, identities are concealed for commercial reasons. For example, although the identities of traders on Instinet were initially released, Instinet later decided against allowing this.[46]

And Pirrong (2010) notes:

> It is well known that transparency has costs as well as benefits. … Moreover, transparency isn't the only thing that matters to market participants. Other aspects of execution affect their costs and benefits as well. A myopic focus on transparency alone ignores these other relevant dimensions.[47]

Additionally, Madhavan, Porter, and Weaver (2005, pp. 286) state:

> Our findings are consistent with theoretical models in which traders adjust their trading strategies based on the level of transparency. Too much transparency increases the ''free option'' cost of limit-order providers, resulting in order withdrawal and a reduction in market depth. Thinner limit order books imply larger transitory price movements associated with order flows, increasing volatility and execution costs.[48]

---

[46] Ruben Lee, 1998, *What Is an exchange? The Automation, Management, and Regulation of Financial Markets*, Oxford, UK: Oxford University Press.

[47] Craig Pirrong, 2010, "What is a swap execution facility?," *Streetwise Professor*, July 1, available at http://streetwiseprofessor.com/?p=3964.

[48] Ananth Madhavan, David Porter, and Daniel Weaver, 2005, "Should securities markets be transparent?," *Journal of Financial Markets*, Vol. 8, No. 3, August, pp. 265-288, available at www.sciencedirect.com/science/article/pii/S1386418105000145.

It should become obvious that displaying the both the order sizes and the time stamps of all other orders in the order book can only have a detrimental impact on market liquidity. The granularity of order book information currently being provided has now exceeded all bounds of propriety, confidentiality, and common sense.

7. Improper Premature Cancellation of Orders

Unless the relevant trading venue is experiencing technical problems, market participants should always be prohibited from cancelling orders before they have received notification from the trading venue that the original order was properly received by the order book. It is particularly difficult to envision a legitimate trading strategy where one would need to cancel an order before receipt of the order was even acknowledged by the trading venue. Engaging in such a practice, however, would almost certainly enhance the effectiveness of both layering and quote stuffing—behaviors that are intentionally deceptive. Detecting this practice would arguably be made easier by having unique identifiers for each algorithm that can generate orders (recommendation 8).

Some might attempt to make the case that one should be able to cancel an order that was determined to have been sent in error. We agree. We also know of no algorithm that can 1) detect that an order was sent in error, 2) generate a cancellation message, and 3) release it faster than the time it takes for the trade match engine to acknowledge receipt of the original message.

Others argue that Regulation NMS (National Market System) or the fragmented U.S. equity markets somehow require market participants to send identical bids to multiple equity trading venues. Then, the argument goes, that if the order gets filled on trading venue #8, the algorithm might have to cancel the identical bids that were sent to all other trading venues and that, on occasion, that might entail sending a cancellation message to the other trading venues before receiving the acknowledgement messages back from those venues that the original orders were received. This argument requires one to believe that trading venue #8 can 1) receive an order, 2) send an acknowledgement message back that the order was received, 3) match that order against one or more resting orders, 4) send matched trade drop copy messages to all of the affected parties, and 5) send the information on the matched trade to its ticker plant faster than the other trading venues can acknowledge receipt of their respective original orders. This argument contradicts all common sense.

8. Unique Identifiers for Trading Systems that Generate Orders Automatically

Every automated trading system that is capable of generating, modifying, or cancelling orders without human intervention should have a unique identifier that distinguishes it to the trading venues where it has the capability to send orders. The process of designing, standardizing, and/or implementing a framework for assigning such identifiers should involve the proprietary trading companies and the relevant trading venues. We see no additional benefit associated with the involvement of the relevant supervisory authorities, which, nonetheless, could benefit from the establishment of such a framework.

Currently trading venues establish session IDs. Session IDs are like pipes through which orders flow into the order book. Dozens of different algorithms can be sending orders through the same session ID. This practice makes it difficult, if not impossible, for supervisory authorities to detect intentional or unintentional misbehavior of individual algorithms. This recommendation would provide trading venues with the ability to identify the firm associated with each session ID and within each session ID, each algorithm that is currently operating. If trading venues had such granular information, they could also potentially alert the trading firms of instances where their individual algorithms might be behaving out of pattern. This capability could be quite helpful to all concerned. It is truly astonishing with all of the human intellect and sophisticated technology that this industry has marshalled, that individual algorithms to this day are not individually identified at most trading venues.

9. Private Access to Trade Information before it is Generally Available to the Public at Large

Trading venues typically provide optional services that allow market participants and trade intermediaries to colocate their respective servers in the data centers of the trading venues. We take no issue with this practice provided that these colocation services are openly available and uniformly priced. Nor do we take issue with the ability of market participants and trade intermediaries to have a latency advantage when entering their orders, because of their colocation in the trading venue's data center. Decreasing the physical distance between one's servers and the trading venue's trade match engine reduces latency to the minimum dictated by the laws of physics.

The matter at issue is whether trading venues are providing trade information to firms that colocate in their data centers before such information is generally made available to the public at large and, if so, whether such a practice is appropriate from a public policy perspective. For risk-management purposes, after a trade is executed, the buyer(s) and sellers(s) that are direct parties to that trade should be so advised as promptly as twenty-first-century technology can.[49] All other market participants, including those that colocate (but are not direct parties to the trade) should be advised that this trade occurred *at the same time as the public at large,* regardless of whether they subscribe to the colocation services of the trading venue or not.

When is information on submitted orders and/or information on executed orders generally available to the public at large? Are firms that subscribe to colocation services currently gaining access to such information before it is generally available to the public at large? Public policy issues of fairness would seem to be mollified if firms that colocate were to receive trade information at the same time that such information were made available to the public at large. This would be accomplished if trading venues provided trade information from their respective ticker plants, rather than from their trade match engines, to firms that colocate in their data centers. Doing so would not only serve the public interest, but also likely encourage the operators of industry utility ticker plants to upgrade the technology of their ticker plants to current industry standards.

## *Implications*

Recommendation 5 (random trade match within half second time intervals) may have the greatest potential to disincentivize all three questionable behaviors—namely, spoofing, layering, and quote stuffing. If you don't know when the next trade match is going to occur, the downside risk of pretending to be a seller when you are really a buyer could leave a trader with a position that is exactly the opposite of his desired position. This could even more so act against the interests of a trader that engages in a combination of spoofing and layering, creating the illusion that there is size building on the bid side of the market, when the trader is really a seller. The trader could get stuck with a substantial position completely the opposite of what he actually wants.

[49] Ideally, this notification would be received via the User Datagram Protocol, which should only be sent to the actual parties to the trade. Please refer to recommendation 6.

One important potential implication of recommendations 5 and 7 is the possible elimination of quote stuffing as a strategy to slow down other algorithmic traders. As no one will know when the trade match and trade allocation algorithms will actually run, one would either have to abandon this strategy (simply considering it as being no longer effective) or attempt to clog the outbound quotation system continuously. Trading venues are reasonably adept at identifying and penalizing traders that have an exceedingly high ratio of quotes to trades (as continuous quote stuffing would undoubtedly require). Recommendation 8 would help do exactly that.

Perhaps most importantly, there is some possibility that recommendation 5 could dampen or even end the incessant low latency arms race. If the trade match engine only runs once every half second, and (assuming some trading venues might adopt recommendations 1 and 2) the allocation of orders would increasingly become a function of time in the order book, the so-called real money resting orders would receive increasing allocations and very short-term traders would receive decreasing allocations. As very short-term traders get allocated fewer and fewer lots, their respective quote-to-trade ratios would logically increase, which almost always carries penalties assessed by the relevant trading venues. If trading venues only matched trades and disseminated price and quantity information once every half second, there would arguably be considerably less financial incentive for all concerned to invest increasingly large sums in an effort to shave one or two milliseconds off a process that only occurs once every 500 milliseconds.

Recommendations 3, 4, and 6 would likely only indirectly disincentivize spoofing, layering, and quote stuffing. But those three trading strategies are not the only behaviors that should arguably be discouraged. The questionable behaviors addressed by recommendations 3 and 4 are obvious: using dark orders and gaming rounding conventions.

Recommendation 6 (providing only aggregated pretrade information from the order book) has more complex implications. Recent research papers by Weller (2012)[50] and by Baron, Brogaard, and Kirilenko (2012)[51] indicate that the fastest high frequency traders 1) are the most profitable and 2) tend not to have

[50] Brian Weller, 2012, "Liquidity and high frequency trading," University of Chicago Booth School of Business and University of Chicago, Department of Economics, working paper, November 10, pp. 42-43.

[51] Matthew Baron, Jonathan Brogaard, and Andrei Kirilenko, 2012, "The trading profits of high frequency traders," Princeton University, University of Washington, and Commodity Futures Trading Commission, working paper, November, pp. 20-21.

their trades match opposite other fast high frequency traders. So, at present, some high frequency traders seem to be taking advantage of their advance access to granular pretrade information from the order book to make large profits and to avoid each other as trading partners. While this phenomenon has been detected in futures contracts, where trades are completely anonymous, it is undoubtedly occurring in the equity markets, because some equity trading venues currently provide more specific trade identifiers—more than only the aggregate quantity bid or offered at each price point—in pretrade information made available to certain market participants.[52] As this information changes millisecond by millisecond, it cannot be of much value to human click traders; it can only be of value to high frequency traders. By obtaining this granular pretrade information, high frequency traders can 1) more efficiently reverse engineer the trading algorithms of their competitors and 2) more effectively discriminate among the counterparties whose resting orders are in the order book. It is difficult to see how either of these activities serves the public interest.

We continue to be of the opinion that trading decisions should be based on economic fundamentals. We appreciate that the decision to buy on one exchange and sell on another may in large part be based on the probability of being allocated lots on the first exchange—and that that will likely be a function of an order's priority in the order book. However, by making recommendation 6, we are challenging the entire premise that our modern electronic financial markets need to be necessarily synchronized to the millisecond. Financial markets have functioned reasonably well in the past at human-scale time horizons.

Recommendation 9 may be the most important of all. If, after a public debate of the issue, it is the consensus that all market participants should get access to trade information at the same time, then automated trading firms that colocate should no longer receive (and make trading decisions based on) trade information before it is generally available to the public at large. In other words, should this consensus arise, trading venues would need to cease providing colocating firms with trade information from their trade match engines and commence providing these firms with trade information from their ticker plants. A welcomed byproduct of this would be the likely deployment of better technology on trading venues' ticker plants.

[52] Arnuk and Saluzzi (2012, pp. 102-103).

## *Regulatory Initiatives*

Germany

On May 15, 2013, the German High Frequency Trading Act (*Hochfrequenzhandelsgesetz)* became effective. It places a number of requirements on exchanges and proprietary trading companies that operate in Germany.

HFT market participants[53] that trade for their own account need to have or obtain a license to do so from BaFin (*Bundesanstalt für Finanzdienstleistungsaufsicht*), the German financial supervisory authority. Firms that are located within the European Economic Area (EEA) can passport their MiFID[54] license to BaFin via their national authority. Non-German firms that are not located within the EEA must create a subsidiary or branch office in Germany in order to obtain a suitable license from BaFIN in order to trade for their own account.

Market participants are subject to maximum order-to-trade ratios that are calculated monthly by product. Violations can result in suspension from trading and/or fines not to exceed €250,000.

Exchanges are required to levy fees on market participants for excessive use of exchange systems, according to the German High Frequency Trading Act.

Importantly, HFT market participants are required to have each of their algorithms labeled with a unique identifier to enable the Trading Surveillance Office to identify manipulative or erroneous algorithms.

Exchanges are subject to a broad "orderly" requirement that would necessitate kill switches, volatility interruptions, and many other protections that are largely already in place.

Exchanges have the affirmative obligation to determine appropriate tick sizes to avoid negative implication to market integrity and market liquidity, according to the German law.

[53] BaFIN has established four criteria to clarify HFT proprietary trading: proprietary trading, a latency minimizing infrastructure, no human intervention, and high intraday message rates.
[54] MiFID is the European Commission's Markets in Financial Instruments Directive.

United States

On September 9, 2013, the U.S. Commodity Futures Trading Commission published its "Concept release on risk controls and system safeguards for automated trading environments."[55] This 137-page document covers a very broad range of suggestions for risk controls and asks whether rulemaking would be appropriate action for the commission to take. Dozens of public comment letters (some over 80 pages in length) are being reviewed by staff. It is not currently clear what ultimate regulatory actions might be forthcoming.

Canada

The Canadian Securities Administrators recently expressed their concern that the payment of trading rebates may be incentivizing behavior that may not serve the public interest. The director of market regulation at the Ontario Securities Commission indicated that "a pilot study of how portions of the market perform without maker-taker incentives will be conducted."[56]

## *Conclusion*

Using term limit orders and running the trade match algorithm at random times within half second intervals would seem to provide an equitable balance between human institutional traders and automated liquidity providers and could drastically reduce the current tsunami of data disseminated by trading venues. Allocating trades based on the actual time that orders have been exposed (or committed to be exposed) to market risk would appear to be a more equitable approach than some trade allocation algorithms currently in use. Implementing both term limit orders and the new trade allocation formula could return some equitability that some argue may have been lost.

Recommendations for establishing appropriate rounding conventions, preventing orders from being cancelled before they are even confirmed, appropriately treating invisible orders, tagging algorithms, and significantly reducing or eliminating the granularity of available pretrade information visible

[55] The release is available at www.cftc.gov/ucm/groups/public/@newsroom/documents/file/federalregister090913.pdf.
[56] Barbara Schechter, 2014, "Canadian regulators look at scrapping controversial fee model linked to high-frequency trading," Financial Post, May 15, available at http://business.financialpost.com/2014/05/15/canadian-regulators-look-at-scrapping-controversial-fee-model-linked-to-high-frequency-trading/.

in the order book are all approaches not inconsistent with sound and defensible public policy with respect to HFT.

Relevant authorities should assess and, if deemed appropriate, solicit public comment on when trade information should be deemed to be generally available to the public at large. National authorities and purveyors of modern electronic trading venues should consider these recommendations and the informed comments of interested market participants.