Market Mechanism Refinement on a Continuous Limit Order Book Venue: A Case Study

HAYDEN MELTON
Deakin University & Thomson Reuters (Markets) LLC

This letter describes an exercise in market mechanism refinement that was recently undertaken on a major electronic trading venue: *Thomson Reuters Matching*. The exercise sought to address problems associated with continuous markets that have been described in recent literature. To this end, the design of the refinement is described, and its consequences are discussed.

Categories and Subject Descriptors: K.4.4 [**Computers and Society**]: Electronic Commerce

1. INTRODUCTION

Worldwide, electronic trading of most financial instruments occurs on venues implementing the *continuous limit order book (CLOB)* [Gould et al. 2013]. On these venues market participants oftentimes compete for resources in the CLOB [Melton 2017b]. When participants are *price-making* on an instrument, for instance, a resource they compete for is queue position for the bids and offers they submit at each price-level in the CLOB [Moallemi and Yuan 2016]. To the detriment of 'slow' participants, 'fast' participants are able to obtain earlier such queue positions and earn the profits associated with those earlier positions [Farmer and Skouras 2012]. Similarly, when participants are *price-taking*, a resource they compete for are the favorably-priced bids (or offers) in the CLOB against which they expect to be able to immediately sell (or buy) the instrument. Again, to the detriment of 'slow' participants, 'fast' participants can earn the profits associated with those favorably-priced bids (or offers) by matching against them (and thus causing their removal from the CLOB), leaving only unfavorably-priced bids (or offers) for the slower participants to match against. In modern financial markets mere nanoseconds—the time taken to send a single character over a gigabit network [CME Group 2017]—may separate the 'fast' from the 'slow'.

Despite its wide adoption in financial markets, two quite serious problems relating to the *continuous* (cf. *batch*) nature of the CLOB have been identified in the recent literature. The first is that it causes an ongoing, socially wasteful technology 'arms race' among participants where each seeks to be marginally faster at sending order-messages to a venue than their peers [Harris 2013; Budish et al. 2015]. It has been argued that this first problem is a form of the *prisoner's dilemma* because participants would be better off if they could all agree to limit their expenditure on technology in pursuit of speed [Budish et al. 2015]. The second is that it exacerbates

Author's address: hmelton@deakin.edu.au & hayden.melton@tr.com
Views expressed herein do not constitute legal or investment advice, and do not necessarily reflect those of the author's employer.

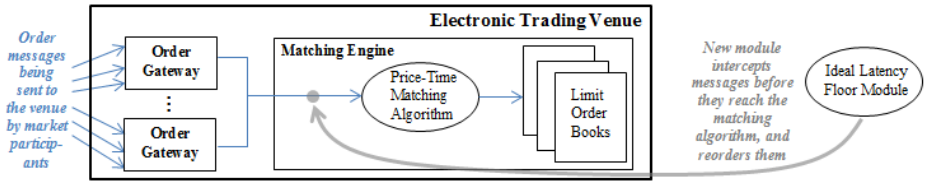

Fig. 1. Partial architecture of the TRM venue, and its subsequent refinement.

information asymmetries in market data distribution and order-message processing that inherently manifest in the technology used to implement these venues [Melton 2017b]. This second problem is one of *fairness*—if one views a venue as a racetrack, it is as if some participants in each race are being given 'head-starts' over others (by being sent market data updates earlier), and some participants are getting to run shorter distances than others (by their order-messages being subject to shorter processing delays by the venue) [Melton 2017a]. The seriousness of these problems is underscored by both the coverage they have received in mainstream media, and the value ascribed to the market inefficiencies they cause, which worldwide has been estimated in the hundreds of billions of dollars per year (see [Melton 2017b] and references therein).

To address the two problems described above, an exercise in market mechanism refinement was recently undertaken on a major electronic trading venue: *Thomson Reuters Matching (TRM)*[1]. What is interesting is that neither of the mechanisms proposed in the academic literature that seek to address these problems (see [Harris 2013] and [Budish et al. 2015]) was selected for this refinement to TRM. Instead, an entirely new mechanism—the *Ideal Latency Floor* [Melton 2015]—was designed from scratch with the explicit goal of minimizing its impact on characteristics of participants' existing trading strategies that do not relate to speed. In this regard use of the term *refinement* here is apt—radical departures from long-established mechanisms (such as the CLOB) may pose significant risks for market participants and venue operators alike, by rendering participants' existing trading strategies (in which they have likely made significant investments) unprofitable, and in turn causing them to stop trading on the venue [Phelps et al. 2010].

2. THE REFINEMENT

At a high-level the manner in which the refinement to the market mechanism was implemented on the TRM venue is shown in Fig.1. In particular, the new (software) module implementing the Ideal Latency Floor 'intercepts' order-messages before they reach the instruments' limit order books, imposes short deliberate delays upon them so as to buffer them, and then releases them from this buffer to the existing price-time matching algorithm in a generally different (temporal) order from that in which they were received by the venue. When described at this level of abstraction

[1]TRM is one of just two major, long-lived interbank spot foreign exchange (FX) trading venues. It launched in 1992, connects participants in more than 50 countries, and oftentimes trades in excess of $100 billion USD notional value daily. From its launch until June 2016 it implemented the CLOB.

the mechanism hardly seems different to those proposed by [Harris 2013] where a random delay between 0-10ms is imposed on order-messages before they are subject to matching, and by [Budish et al. 2015] where order-messages are batched every 100ms before being subject to matching. In all these mechanisms the advantage of being the 'fastest' participant is reduced because (unlike in the CLOB mechanism) it is not necessarily the order-message that reaches the venue first that is subject to matching first.

The specification of the Ideal Latency Floor mechanism is more involved than the mechanisms of [Harris 2013] and [Budish et al. 2015], and a plain-English description of it as it was deployed into production on TRM in June 2016 is as follows:[2]

—Upon receipt of an order-message (but not a cancel-request), the limit price of that order is compared to the best bid or offer in its instrument's limit order book to determine if it is *marketable*. If it is, it is put in either the 'taker as buyer' buffer or 'taker as seller' buffer for the instrument, depending on its side (buy or sell). If the order is *not* marketable, it is put in the 'maker bids at X' buffer or 'maker offers at X' buffer for the instrument, where X is the order's limit price. (In this manner, there are a plurality of buffers for each instrument on the venue, and crucially there is precisely one such buffer for each resource on the venue for which participants may compete).
—Each buffer has an associated timer that begins counting up to 3ms whenever it receives an order-message that causes it to transition from being an empty buffer to non-empty buffer. Upon its timer reaching 3ms the buffer is drained (i.e., rendered empty again) by removing all its order-messages and subjecting them to matching against its instrument's limit order book.
—The specific procedure for draining a buffer involves determining a random ordering on the participants that sent the order-messages in it, then repeatedly iterating over that random ordering of participants to remove the single oldest such remaining order-message in the buffer for that participant. To illustrate: if order-messages of the form $Participant_{OrderIdentifier}$ are *received* by the buffer in the following sequence $[A_1,B_1,B_2,A_2,A_3,C_1]$, and the random ordering of participants yields $[B,A,C]$, then the *draining* of order-messages from the buffer will be in the sequence $[B_1,A_1,C_1,B_2,A_2,A_3]$.
—Neither cancel-requests nor immediate-or-cancel (IOC) order-messages that are unmarketable upon their receipt are subject to buffering so are both processed immediately against their instrument's limit order book as they would otherwise have been in a CLOB. (Cancel-requests enable participants to remove bids and offers they previously submitted that would otherwise would remain active in the limit order book; IOC orders, by definition, can only be used for price-taking so can never appear in the limit order book as bids or offers).

3. IMPACTS ON EXISTING TRADING STRATEGIES

The complexity evident in the design of Ideal Latency Floor described in the previous section was gradually and deliberately introduced over many iterations in its

[2]A more formal and highly-parameterized specification of the mechanism is provided in [Melton 2015]. The description provided here is intended to be more gentle, and to illuminate the specific parameters chosen for its deployment on TRM.

design process by repeatedly asking two questions. Those two questions—noting 'behavior' refers to characteristics of participants' trading strategies—were:
1. *What new, undesirable behaviors might the mechanism incentivize?*
2. *What existing, desirable behaviors might the mechanism disincentivize?*
The answers to these questions were of course colored by the author's practical experience in the field electronic trading, and by the duty owed to his employer, the operator of the TRM venue. Regardless, the answers independently arrived at during course of the mechanism's design in 2013 in many cases seem to be consistent with those that have appeared in the academic literature. Along these lines, a brief discussion of each of the characteristics of the mechanism's design follows.

Participant-wise draining of a buffer. A criticism of the mechanism of [Harris 2013] is that a 'fast' participant could continue to exploit their speed by sending multiple, redundant copies of the same order-message to a venue implementing it [Budish et al. 2015].[3] Various solutions to this problem exist—for instance, imposing a tax on duplicate messages, introducing a rule on the venue to prohibit the behavior, and so on, but the specific solution implemented here was to allocate a resource not on a per order-message basis, but rather on a per participant basis. In this way participants are not encouraged to alter their behavior to send multiple copies of the same order-message to the venue, which among other things would disadvantageously cause additional load on it.

One buffer per resource.[4] Operators of venues tend to value participants engaged in *market-making* on it quite highly, because without the bids and offers they are continually submitting other participants whose strategies mostly involve *price-taking* would have nothing to match against [Mizuta and Izumi 2016]. The most active of market-makers tend to submit bids and offers at multiple price-levels in an instrument's limit order book 'all at once' and it is important that they are not disincentivized from doing this. If in the mechanism there were only one buffer per instrument (and not one buffer per resource) then, ceterus paribus, a market-maker would be disadvantaged in their allocation of individual resources on the venue relative to other participants submitting relatively few order-messages at a time.[5]

A 'short', order-message triggered delay. Many financial markets are *fragmented* in that the same instrument trades on a plurality of competing venues. The sudden imposition of long delays in order-message processing by a venue may disadvantageously cause participants to change their behavior to instead route their order-messages to competing venues exhibiting shorter delays [Donier and Bouchaud 2016]. The length of the delay here (3ms) was thus derived to be as the smallest value that adequately addressed the issues of fairness and participants' speeds on the venue (see [Melton 2017b; 2017a]). Further, by virtue of the delay's timer being triggered by the receipt of a first order-message in the 'race' for a re-

[3]See [Melton 2017b] and references therein for a documented, real-life example of this phenomenon.
[4]To be sure, it is *not* the financial instruments that trade on the venue that are, per se, the resources here. Rather, the resources are various properties of each instrument's limit order book for which participants on the venue compete. These are: favorably-priced bids or offers when they are price-taking, and queue position at each price-level when they are price-making.
[5]See [Melton 2017b] for an example illustrating this.

source, participants are not incentivized to alter their behavior to withhold their order-messages from a venue until (say) just before a batch auction window (as in [Budish et al. 2015]) closes. Indeed, mechanisms where participants are incentivized to submit order-messages only at known points-in-time so as to minimize the delays they experience may also disadvantageously incentivize ongoing investments in speed-related technology [Mizuta and Izumi 2016].

Participant-wise retention of temporal ordering in a buffer. Market-making participants oftentimes submit a plurality of small bids (or offers) 'all at once' at the same price-level. In a CLOB, a market-marker can discern which of their own order-messages are nearer the front of the queue at the price-level, and which are nearer the back by simply keeping track of the order in which they were submitted. When, due to a change in market conditions a market-maker wishes to reduce quantity bid (or offered) at the price-level, they will likely do so by first canceling their bids (or offers) that are occupying less valuable positions nearer the back of the queue. The draining technique employed here ensures this existing behavior by market-makers is unaffected. It similarly ensures that participants who are accustomed to price-taking multiple price-levels in a CLOB by sending one order-message per price-level 'all at once' advantageously do not have to alter their behavior.[6]

Removal of only a single order-message per participant per iteration. While participants do not know the information content of future events that will cause them to compete for resources on the venue, sometimes they do know the timing of such events (e.g., regularly scheduled economic data releases). In light of this, an undesirable behavior a fast participant may engage in is to send a 'phony' order-message just before the occurence of such an event to strategically start a buffer's timer 'early'. To the detriment of 'slow' participants, this behavior will effectively shorten the period the timer runs for *after* the actual event. The draining technique employed here disincentivizes this behavior by ensuring that if a 'fast' participant were to do this, a second equally fast participant would be guaranteed to have his *legitimate* order-message that was sent in response to the information contained in the event subject to matching by the venue before the second, legitimate message sent by the first participant. Indeed, the one buffer per resource approach further defends against this same undesirable behavior because one cannot, for instance, strategically submit a 'phony' order-message for a low-cost/low-return resource (e.g., bidding well beneath 'the touch') to strategically start the timer for a distinct high-cost/high-return resource (e.g., buying 'at market').[7]

No delays for cancel-requests. The predator-prey relationship that oftentimes exists between 'fast' takers and 'slow' makers is well-documented in the literature [Farmer and Skouras 2012; Budish et al. 2015]. By not imposing delays on cancel-requests the approach ensures more favorable treatment of makers because they are

[6]In terms of price-taking here, if the following order-messages mirror exactly the state of the offers in the limit order book, any sequence of them different from [buy 1 unit at 1.00, buy 1 unit at 1.01, buy 1 unit at 1.02] at matching will *not* be completely marketable.

[7]This same line of reasoning explains why IOC order-messages that are unmarketable upon receipt are not subject to delay—if they were, they could be used strategically (and at no economic cost to a participant) to start the timer for a buffer.

guaranteed to be able to cancel their bids or offers provided their cancel-request is received by the venue within 3ms of the receipt of the fastest taker's order-message. Alternate approaches to handling cancel-requests are described in [Melton 2015], but these may lead to (undesirable) strategic behavior of the form described earlier, where cancel requests can be used at low economic cost to strategically start a buffer's timer 'early'.

4. CONCLUSIONS

The market mechanism refinement described in this letter seems to provide a number of advantages over others that have appeared in the literature. Further, its success in actual use is reflected by (1) it being well-received by participants on TRM, and (2) there being no plans to remove it from the venue or to further modify it. While several other FX venues have also moved away from the CLOB [Detrixhe 2016], it remains to be seen the extent to which venues in other asset classes will follow suit. Finally, what underpins most of these mechanisms that are in actual use and that have received citations in the literature is the general technique of *buffering*; the extent to which non-buffering mechanisms [Melton 2016; Kyle and Lee 2017; OneChronos 2016] will come into actual use remains to be seen.

REFERENCES

Budish, E., Cramton, P., and Shim, J. 2015. The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics 130,* 4 (Nov.).

CME Group. 2017. Notice of summary access denial - NYMEX-16-0600. `https://goo.gl/stsSF3`. In *CME Group Disciplinary Notices* (online), accessed Jan 6 2017.

Detrixhe, J. 2016. Speediest traders becoming less welcome in currency markets. `https://goo.gl/Ris4s2`. In *Bloomberg News*, accessed Jan 9 2017.

Donier, J. and Bouchaud, J.-P. 2016. From walras auctioneer to continuous time double auctions: A general dynamic theory of supply and demand. *Journal of Statistical Mechanics: Theory and Experiment 2016,* 12.

Farmer, D. and Skouras, S. 2012. Review of the benefits of a continuous market vs. randomised stop auctions and of alternative priority rules (policy options 7 and 12). `https://goo.gl/QDpgSP`. *UK Government Office for Science*, accessed Jan 7 2017.

Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J., and Howison, S. D. 2013. Limit order books. *Quantitative Finance 13,* 11, 1709–1742.

Harris, L. 2013. What to do about high-frequency trading. *Financial Analysts Journal 69,* 2.

Kyle, A. S. and Lee, J. 2017. Toward a fully continuous exchange. *Available at SSRN 2924640*.

Melton, H. 2015. Ideal latency floor. US Patent App. 14/533,543.

Melton, H. 2016. Systems and methods for obtaining and executing computer code specified by Code Orders in an electronic trading venue. US Patent App. 15/064,163.

Melton, H. 2017a. A fairness-oriented performance metric for use on electronic trading venues. Tech. rep., Deakin University. July. `https://doi.org/10.13140/RG.2.2.14427.05922/1`.

Melton, H. 2017b. Understanding and improving temporal fairness on an electronic trading venue. In *37th Int'l Conf. on Distributed Computing Systems Workshops*. IEEE, 1–6.

Mizuta, T. and Izumi, K. 2016. Investigation of frequent batch auctions using agent based model. Tech. rep., Japan Exchange Group. Dec. `https://goo.gl/37CYjs`.

Moallemi, C. and Yuan, K. 2016. A model for queue position valuation in a limit order book. *Available at SSRN 2996221*.

OneChronos 2016. *OneChronos* website. `https://www.onechronos.com/`. Accessed May 29 2017.

Phelps, S., McBurney, P., and Parsons, S. 2010. Evolutionary mechanism design: a review. *Autonomous Agents and Multi-Agent Systems 21,* 2, 237–264.