American Economic Review: Papers & Proceedings 2014, 104(5): 418–424
http://dx.doi.org/10.1257/aer.104.5.418

FRONTIERS OF MARKET DESIGN[‡]

Implementation Details for Frequent Batch Auctions: Slowing Down Markets to the Blink of an Eye[†]

By Eric Budish, Peter Cramton, and John Shim*

Financial exchanges around the world predominantly use a market design called the *continuous limit order book* (CLOB). Our recent research, Budish, Cramton, and Shim (2013)—henceforth, BCS—argues that this design is flawed. In a continuous-time market, every time there is new public information that has implications for security prices—every change in the price of one security that has implications for the prices of correlated securities, every company announcement, etc.—there is a race to react. On one side of the race are high-frequency trading (HFT) firms looking to "snipe" stale quotes before they are adjusted. On the other side of the race are liquidity providers—typically, other high-frequency trading firms—looking to adjust their outstanding quotes to reflect the news before they are sniped. Since the CLOB processes traders' messages in *serial*—that is, one-at-a-time in order of arrival—liquidity-providing HFTs usually lose this race. Their one request to adjust their stale quotes would have to reach the exchange before *all* of the requests to pick off their stale quotes. The race thus creates an unnecessary, purely technical cost of liquidity provision—incremental to fundamental costs such as information asymmetries, inventory costs, etc.—which ultimately is passed on to investors in the form of wider spreads and thinner markets. The time-scale at which the race occurs has changed over time—in 2005, it was measured at the scale of 0.01 seconds; by 2008, 0.001 seconds; presently, 0.0001 seconds—but, under the CLOB market design, there always has been and always will be a race. A socially-wasteful and liquidity-reducing speed race is an equilibrium feature of the market design.

As an alternative, BCS propose *frequent batch auctions*—uniform-price sealed-bid double auctions conducted at frequent but discrete time intervals. BCS show that frequent batching eliminates the speed race, and the associated harm to liquidity and social welfare, for two reasons. First, modifying the market design *from continuous-time to discrete-time* substantially reduces the value of tiny speed advantages (e.g., 0.0001 seconds). In a continuous-time market, a tiny speed advantage is enough to always win the race; in a discrete-time market—even one as fast as the blink of an eye (roughly 0.5 seconds)—tiny speed advantages are orders of magnitude less valuable. Second, modifying the market design *from a serial process to a batch process* transforms the nature of competition—from competition on speed to competition on price. Rather than competing to be the first message processed, traders compete to be the most attractive quote. BCS show that these two

[‡]*Discussants:* James Andreoni, University of California-San Diego; Michael Ostrovsky, Stanford University; Larry Samuelson, Yale University; Ilya Segal, Stanford University.

*Budish: University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637 (e-mail: eric.budish@chicagobooth.edu); Cramton: Economics Department, 3114 Tydings Hall, University of Maryland, College Park, MD 20742 (e-mail: cramton@umd.edu); Shim: University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637 (e-mail: john.shim@chicagobooth.edu). Budish gratefully acknowledges financial support from the National Science Foundation (ICES-1216083), as well as the Fama-Miller Center for Research in Finance and the Initiative on Global Markets at the University of Chicago Booth School of Business.

[†] Go to http://dx.doi.org/10.1257/aer.104.5.418 to visit the article page for additional materials and author disclosure statement(s).

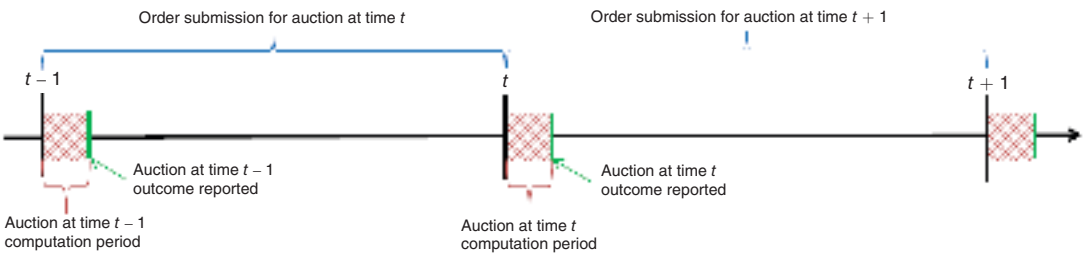

FIGURE 1. PROCESS FLOW FOR FREQUENT BATCH AUCTIONS

benefits of frequent batching translate to greater liquidity for investors and greater social welfare.

The purpose of the present paper is to describe the practical implementation details of frequent batch auctions in more detail than was possible in BCS. We see this exercise as in the spirit of Roth's (2002) vision of "The Economist as Engineer," taking theoretical ideas and translating them for use in the real world by paying close attention to the relevant institutional and computational details. We note open questions throughout.

I. Frequent Batch Auctions: Process Flow

Figure 1 depicts the process flow for frequent batch auctions. There are three components: order submission, auction, and reporting. We describe the design details for each component in turn. This section focuses on a non-fragmented market; we discuss how to augment frequent batch auctions to accommodate fragmented markets (e.g., US equities markets) below.

Throughout, design details are chosen to minimize the departure from current practice, subject to realizing the benefits of frequent batching. This is both to reduce transition costs and to limit the scope for unintended consequences.

A. *Order Submission*

Orders in a batch auction consist of a direction (buy or sell), a price, and a quantity, just like limit orders in a CLOB. During the order submission stage, orders can be freely submitted, modified, or withdrawn. If an order is not executed in the batch auction at time t, it automatically carries over for the next auction at time $t + 1$, $t + 2$, etc., until it is either executed or withdrawn.

Orders are *not displayed* during the order submission stage.[1] This is important to prevent gaming, and is why we describe the auction as "sealed bid." Orders are instead displayed in aggregate at the reporting stage, as described below.

An important open question is the optimal tick size in a batch auction. The simplest policy would be to mimic the tick size under the current CLOB, which in the United States is \$0.01 by regulation (\$0.0001 for stocks less than \$1 per share). However, we note that one of the arguments against finer tick sizes—the explosion in message traffic that arises from traders outbidding each other by economically negligible amounts—is moot here, because orders are opaque during the batch interval. We also note that the coarser the tick size, the more important a role rationing will play. For these reasons, we conjecture that the optimal tick size in a frequent batch auction is at least as fine as in the continuous market.

B. *Auction*

At the conclusion of the order submission stage, the exchange batches all of the outstanding orders, and computes the aggregate demand and supply functions from orders to buy and sell, respectively. There are two cases: supply and demand cross, or they do not. See Figure 2.

[1] This is a common misconception about frequent batch auctions. For instance, the Chicago Mercantile Exchange's objection to frequent batch auctions in its December 2013 comment letter to the Commodity Futures Trading Commission focused extensively on the gaming issues that would arise if bids in a batch auction were displayed before the auction is conducted (http://comments.cftc.gov/PublicComments/ViewComment.aspx?id=59456). We agree that this form of batch auction would be a bad idea.

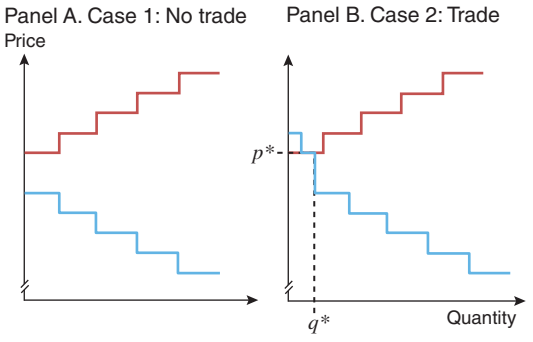

FIGURE 2. ILLUSTRATION OF SUPPLY, DEMAND, AND AUCTION OUTCOMES

Case 1: If supply and demand do not cross—the highest bid is lower than the lowest ask—then there is no trade. All orders remain outstanding for the next batch auction. We expect this case to be quite common, especially for securities that are thinly traded.[2] The supply and demand curves in this case represent liquidity that is being offered to the market by trading firms—this is analogous to how the limit order book in the continuous market represents liquidity provision by trading firms.[3]

Case 2: Supply and demand cross. Typically, if supply and demand cross they do so horizontally. Let p^* denote the unique price and let q^* denote the maximum quantity. In this case, all orders to buy with a price greater than p^* and all orders to sell with a price less than p^* transact their full quantity at p^*. For orders with a price of exactly p^* it will be possible to fill one side of the market in full whereas the other side will have to be rationed. We suggest the following rationing rule: pro-rata with time priority across batch intervals but not within batch intervals. That is, orders from earlier batch intervals are filled first, and in the single batch interval in which it is necessary to ration, this rationing is pro-rata treating all orders submitted within that interval equally. This policy encourages the provision of long-standing limit orders, as in the CLOB, but without inducing a race to be first within a batch interval.

There is also a knife-edge case in which supply and demand intersect vertically instead of horizontally. In this case, quantity is uniquely pinned down, and the price is set to the midpoint of the interval. There is no need for rationing in this case.

C. *Reporting*

When the auction stage is completed, the following information is announced publicly:

- Price: either the market clearing price, p^*, or the outcome "no trade."
- Quantity: the quantity filled, q^*.
- The aggregate demand and supply curves.

Additionally, the submitter of each individual order receives a private message reporting the outcome for that particular order. All else equal, we suggest that individual orders not be made public as this would add little to price discovery and strengthen incentives to hide or split orders.

We emphasize that the information policy we propose for frequent batch auctions is analogous to current information policy under the CLOB. In particular, in both the CLOB and the frequent batch auction, the following three things happen in sequence for each order: (i) the order is submitted to the exchange; (ii) the order is processed by the exchange; (iii) the relevant information is announced publicly. The key difference is that in the batch auction there is a small period of time that elapses between (i) and (ii), i.e., between when the order is submitted and when it is processed at the end of the batch interval.

[2] Over a handful of randomly selected days in 2011, we found that SPY had no trading activity in 14 percent of one-second intervals, JPM had no trading activity in 38 percent of seconds, and GOOG in 73 percent of seconds.

[3] An apparent dissimilarity is that the supply and demand curves in the frequent batch auction represent the liquidity that was offered in the *last* batch auction, not necessarily in the current batch auction, whereas in a CLOB the current state of the limit order book represents liquidity that is being offered "right now." However, this is not correct. Since there is latency in any continuous-time market, the best a trader can do in a CLOB is see the limit order book as it was a short time ago (e.g., 1ms ago) and act on this limit order book a short time later (e.g., 1ms later). Thus, just as in a frequent batch auction, liquidity perceived at the time of a decision to trade may or may not be there when one's trade request is actually processed. An entire new exchange, IEX, has been introduced to deal with problems that arise due to the difference between the limit order book as perceived by a trader and the limit order book as it actually is when the trader's order reaches the market.

II. A Modification to Account for Market Fragmentation and Reg NMS[4]

US equities markets are highly fragmented, meaning that the exact same security can be bought and sold on many different exchanges rather than on just a single listing exchange. In this section we discuss a modification to the auction stage of frequent batch auctions that could allow frequent batch auctions to operate in the context of a fragmented market in a manner that is both economically appealing and consistent with the spirit of the relevant regulation.[5] Roughly, the modification is to incorporate liquidity from the continuous market into the batch auction when advantageous to do so.

Regulation National Market System (Reg NMS), passed in 2005 and implemented in 2007, governs the interaction of exchanges in the fragmented market. Conceptually, it is useful to think of Reg NMS as attempting to synthesize a national CLOB market from multiple individual CLOB markets running in parallel. It does so with two key provisions. The Access Rule (Rule 610) requires displayed quotes from each exchange to be immediately accessible by other exchanges and prevents the national limit order book from being crossed. The Order Protection Rule (Rule 611) prevents an exchange from executing a trade at a price that is inferior to the top-of-book quote on any other exchange (and in particular from executing trades at a price that is inferior to the National Best Bid and Offer (NBBO)), without first attempting to execute on all other exchanges with a superior top-of-book quote. Otherwise, the exchange is said to be "trading through" a "protected quote," which is prohibited.

These provisions of Reg NMS raise two potential regulatory issues for frequent batch auctions. First, what happens if a participant in the batch auction submits a bid or ask outside the current NBBO—is there an obligation to prevent a nationally crossed book? Second, what happens if supply and demand in the batch auction cross at a price outside the NBBO—would executing trade at this price violate the prohibition against trade through?

The first concern—crossed books—is most likely a legal non-issue for a technical reason: the prohibition is against "*displaying* quotations that lock or cross any protected quotation in an NMS stock" (Reg NMS, Rule 610(d)(i), emphasis added). Since orders in a batch auction are kept sealed until the auction is conducted, there is no violation. We also note that this interpretation seems consistent with the economic spirit of the rule as well as current practice by exchanges (e.g., the use of non-displayed quotes to prevent a nationally crossed book).

The second concern—trade through—necessitates a modification to the auction stage of frequent batch auctions. Suppose that the NBBO at the end of the batch interval is bid \$9.99 and ask \$10.01. If the batch auction clears at a price inside the NBBO (e.g., \$10.00) or at the NBBO (e.g., \$10.01) then there is no trade through. Participants in the batch auction market are getting a price that is either strictly or weakly more attractive than the best price they could obtain on the continuous market. Suppose, however, that supply and demand in the batch auction cross at a price outside the NBBO, e.g., \$10.03. Executing trades at \$10.03 while there are asks available at other exchanges for \$10.01 would violate both the letter and spirit of the trade-through rule. As a remedy, we modify the auction stage by incorporating liquidity from continuous markets into the supply and demand curves in the batch auction when advantageous to do so. Specifically, in this example, we would augment the supply curve in the batch auction by including asks at the NBBO of \$10.01 (all of which are protected under Reg NMS), as well as asks at \$10.02 and \$10.03 (which may or may not be protected under Reg NMS) in the batch auction supply curve.

More precisely, let t denote the end of the order submission stage, let S_t^{batch} and D_t^{batch} denote the supply and demand (step) functions from the orders submitted to the batch auction, and let S_t^{clob} and D_t^{clob} denote the supply and demand (step) functions formed by aggregating all asks and bids, respectively, from the member exchange CLOB markets given the state of their limit-order books at time t. Assume that the batch auction exchange can execute perfectly in

[4] We are extremely grateful to numerous industry participants and the fellows of the Program in the Law and Economics of Capital Markets at Columbia Law School for detailed discussions on the topics covered in this section. Any errors in interpretation of the relevant regulation are our own.

[5] We believe that the modification also is consistent with the letter of the relevant regulation, but this is a matter for regulatory clarification.

the CLOB markets in the sense that it can sweep in whatever of S_t^{clob} or D_t^{clob} is desired into the batch auction (we will discuss what happens under imperfect execution below). Let B_t and A_t denote the national best bid and ask at time t. Given generic supply and demand curves S and D, let the function $p(S, D)$ return the market-clearing price and $q(S, D)$ the market-clearing quantity. Our proposed modification to the auction stage is as follows:

(i) If the market clearing price of the batch supply and demand curves is within the NBBO (i.e., $p(S_t^{batch}, D_t^{batch}) \in [B_t, A_t]$) then proceed as in Section I.

(ii) If the market clearing price of the batch supply and demand curves is greater than the national best ask (i.e., $p(S_t^{batch}, D_t^{batch}) > A_t$), then:

a. Augment the supply curve: $S_t^{b\&c} = S_t^{batch} + S_t^{clob}$.

b. Compute the market-clearing price and quantity given batch demand and augmented supply: $p^* = p(S_t^{b\&c}, D_t^{batch})$ and $q^* = q(S_t^{b\&c}, D_t^{batch})$.

c. Compute the set of inter-market sweep orders to send to the CLOB exchanges to sweep in needed supply. Specifically, form $S_t^{sweep} \subset S_t^{clob}$ consisting of all supply from the CLOB exchanges with ask price strictly lower than p^* and the minimum necessary supply from CLOB exchanges with ask price equal to p^* such that $q(S_t^{batch} + S_t^{sweep}, D_t^{batch}) = q^*$. Send inter-market sweep orders corresponding to S_t^{sweep} to the CLOB markets.

d. Execute all batch orders with bid strictly greater than p^* and ask strictly lower than p^* at price p^*. For batch orders with bid or ask equal to p^*, it may be necessary to ration one side of the market as in Section I.

e. Any difference between p^* and the prices paid in the sweep of S_t^{sweep} generates a surplus. Distribute this on a per-share basis to all filled batch orders.

(iii) If the market clearing price of the batch supply and demand curves is less than the national best bid (i.e., $p(S_t^{batch}, D_t^{batch}) < B_t$), then proceed analogously to step (ii).

We make several clarifying remarks concerning this procedure.

First, we emphasize that this procedure is only necessary for the case where the clearing price for batch orders alone is outside the NBBO. We expect this case to be relatively rare.

Second, we believe that this procedure complies with the trade-through requirement of Reg NMS because, if the price in the augmented auction is p^*, step (ii, c) ensures that all CLOB orders with price strictly less than p^* are swept into the batch auction.

Third, step (ii, c) assumes that the batch exchange has perfect execution in the CLOB exchanges. In practice, small latencies could cause the batch exchange to sweep in less than S_t^{sweep}. There are two choices for how to deal with the possibility of imperfect execution. The first is to ration batch orders in step (ii, d) as needed to deal with the shortfall, keeping the price at p^*. The second, letting $\tilde{S}_t^{sweep}$ denote the supply successfully swept in step (ii, c), is to re-compute the clearing price and quantity as $p^* = p(S_t^{batch} + \tilde{S}_t^{sweep}, D_t^{batch})$ and $q^* = q(S_t^{batch} + \tilde{S}_t^{sweep}, D_t^{batch})$, and proceed accordingly. The second alternative is more economically appealing, because it generates larger gains from trade, but the second alternative may not comply with Reg NMS if the re-computed price p^* causes trade through of a protected quote. The first alternative complies with Reg NMS but at the cost of some gains from trade.

Fourth, we emphasize that the sweep of orders from the CLOB market may generate surplus for the batch exchange, because the batch exchange pays the ask price for the orders in S_t^{sweep}, not p^*. This surplus can be rebated to participants in the batch auction.

Last, we note that an important open question for future research is the nature of equilibrium if investors, market makers, and exchanges each are endogenously choosing between continuous and batch markets.

III. Considerations for Determining the Duration of the Batch Interval

The theoretical analysis in BCS suggests that the batch interval should be long relative to the following two quantities:

(i) The average difference in reaction time to a commonly-observed piece of news between trading firms at the cutting edge of speed technology and trading firms not at the cutting edge (e.g., this year's cutting edge versus last year's cutting edge).

(ii) The maximum difference in reaction time to a commonly-observed piece of news among trading firms at the cutting edge of speed technology (i.e., how large is the stochastic component of speed among trading firms at the cutting edge).

Our understanding from conversations with industry participants is that, as of the present writing, (i) is less than 0.001 seconds and (ii) is less than 0.0001 seconds. BCS show that the batch interval should be chosen so that the ratios of (i) and (ii) to the batch interval are small.[6]

In addition to these items highlighted by the theoretical analysis in BCS, there are several additional practical considerations for determining the batch interval:

(iii) The auction stage of the batch interval should be long enough to allow the exchange to compute the auction outcome even in worst-case scenarios for message traffic.

(iv) The auction stage of the batch interval should be long enough to allow the batch auction exchange to sweep orders from CLOB exchanges as described in Section II.

(v) The reporting stage of the batch interval should be long enough to allow for information to travel round-trip from the batch exchange to other exchanges trading related securities.

(vi) The batch interval should be long enough so that trading algorithms have adequate time between receipt of time t auction outcomes and the close of the time $t+1$ auction to make time $t+1$ order submission decisions.

To assess item (iii), we ran computational simulations of frequent batch auctions based on the rate of message traffic during the flash crash (a reasonable worst case). Even with a very simple technology setup (C++ on an ordinary laptop) we were always able to compute the auction outcomes in under 0.010 seconds.

Regarding item (iv), the time required to sweep orders from other exchanges is small because all member exchanges' servers are located in a small geographical area in New Jersey. Our discussions with industry participants suggest that this latency is about 0.0005 seconds round-trip for the four most important exchanges (NYSE, NASDAQ, BATS, Direct Edge).

Regarding item (v), a rough upper bound on the round-trip information travel time between any two exchanges is the amount of time it takes information to travel around the world, which, at the speed-of-light, is 0.135 seconds.[7] If the batch interval is long relative to 0.135 seconds, batch auctions could be synchronized at various locations around the world so as to satisfy the following attractive property: traders in any one location (e.g., Chicago) wishing to participate in the time t auction in any other location (e.g., London) have information about the time $t-1$ auction outcomes from *all* locations (e.g., Chicago, New York, London, Tokyo), and have information about the time t auction outcomes from *no* locations.

Regarding item (vi), our point is simply that, once we subtract from the batch interval the time it takes to run the auction and the time it takes for information to travel, there should be some time left over for algorithms to process information and make trading decisions. We do not have a precise view on how long this additional time should be. The simplest trading

[6] More precisely, item (i) corresponds to δ in the BCS model, while item (ii) corresponds to ϵ in an extension of the BCS model described in footnote 40 of BCS. BCS show that the batch interval, τ, should be chosen so that the ratios $\frac{\delta}{\tau}$ and $\frac{\epsilon}{\tau}$ are small.

[7] We abstract from the potential for latency arbitrage between planets, or galaxies (cf. Krugman 1978).

algorithms need comfortably less than 0.001 seconds to make decisions, whereas there are of course more complicated trading algorithms that need amounts of time that are orders of magnitude larger.

REFERENCES

Budish, Eric, Peter Cramton, and John Shim. 2013. "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response." http://faculty.chicagobooth.edu/eric.budish/research/HFT-FrequentBatchAuctions.pdf (accessed January 13, 2014).

Krugman, Paul. 1978. "The Theory of Interstellar Trade." http://www.princeton.edu/~pkrugman/interstellar.pdf (accessed January 13, 2014).

Roth, Alvin E. 2002. "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics." *Econometrica* 70 (4): 1341–78.

Securities and Exchange Commission. 2005. "Regulation NMS." Release No. 34-51808. http://www.sec.gov/rules/final/34-51808.pdf (accessed, January 13, 2014).