TL;DR: I propose a simple model of competition for earlier transaction execution to make sense of several aspects of transaction ordering policies:

Can we avoid latency competition by using bidding to determine the ordering and inclusion of transactions? (not really)

How does latency competition look like in a batch auction world? (we have zero average profit for bidders in equilibrium when accounting for the cost of latency investment)

What are the performance differences between batch auctions and hybrid formats such as the time boost proposal for Arbitrum? (depends on the parameterization, but in general they look very similar)

The results are particularly relevant for roll-up sequencing but should also inform the broader transaction ordering discourse for L1s or off-chain aggregators.

Many thanks to Quintus, Akaki, Sergioand Alejofor discussions and feedback on earlier versions. Forumfor discussion and comments.

Suppose you want to choose a transaction ordering policy for your new roll-up sequencer, L1 blockchain, DEX aggregator, or new financial exchange. The following categories seem to cover your available options pretty well:

First Come First Serve orders transactions by time stamp of arrival at whatever server orders the transactions. Questions of decentralized implementation aside, this policy seems appealing and intuitive to many. It is the go-to policy in traditional finance and hence users are used to interacting with it. FCFS appeals to basic intuitions about fairness. But there is also an efficiency argument to be made: FCFS provides an incentive to incorporate new external information quickly into the state of the system.

Bidding Based Ordering : Orders are processed in batches or blocks. Transactions within a batch interval are ordered according to a function of the attached bids . The function can (hypothetically) be arbitrarily complex: we could order transactions by bid, we could auction individual slots in the batch, or even allow combinatorial bidding where users express preferences over the entire content of the block. Also the “bid” can sometimes be interpreted broadly, for example in Ethereum block building, the bid could contain the amount of MEV the user allows the block builder to extract from him.

Random ordering or other non-conventional policies : Questions of implementation aside,randomnessis a mean s to achieve ex-ante fairness when ranking transactions. In a different direction, verifiable sequencing rulesare designed to make it detectable if a sequencer deviates from the rule. But, so far, they only have been designed to order AMM transactions for a single trading pair.

Hybrid policies : The above policies can be mixed and matched. For example, FCFS can be implemented with discrete buckets where orders within a bucket are ordered randomly. A recent proposalby Offchain Labs, orders transactions by a scoring rule called “time boost” that scores transactions by a combination of time stamps and bids.

While these policies look very different from each other, a substantial aspect of all of them is that they organize a contest for earlier transaction execution among those users that care about it (arb traders, liquidators, etc.). The term contest here has the usual meaning from economics: users exert effort (investing in latency, spending money on bidding, spamming your server with transactions) to produce a signal (a timestamp, a bid, a set of transactions IDs) and based on these signals, we decide which transactions to include and in which order (and therefore decide who wins the different contests).

The framing as a contest is helpful, in so far as it gives us an indication of what it means to choose a transaction ordering policy: we organize a contest among users and users maximize whatever signal maximizes their chances of winning the contest. Thus, we need to decide on what dimension we want them to compete: latency investment, expenditure on bidding, the number of transaction requests you receive, increasing entropy, a mixture of all of them, etc?

In the following, I want to focus on the first two categories of transaction ordering policies, time stamp and bidding based policies, and hybrid policies mixing between the first two categories. This is because these policies have an implicit or explicit focus on efficiency (broadly understood) which is desirable.

To motivate the following discussion I would like to start with three somehow obvious observations:

A pure bidding mechanism is impossible: This simply follows from the fact that the bidding phase of an auction cannot run forever. Thus, necessarily some bids end in different batches than others, and time plays a role.

Latency competition necessarily happens, even in a pure batch auction : This is a corollary of the first point. In a batch auction, there still is an advantage for low latency bidders when approaching the end of the batch. This has for example been well-documentedin Ethereum block building, where searchers specialized in CEX-DEX arbitrage need low latency to be competitive in “top of block” MEV.

Illustration of the latency game at the end of a batch auction, source: 0xpandebug onArbitrum Research

Bids are more informative signals of users’ preferences than time stamps : If time stamps are used to order transactions, then users are incentivized to create early time stamps. This can lead to investment into colocation and latency reduction. While the negative effects of excessive latency competition has beenrecognized, investment in latency is not wasteful per

se. It incentivizes quick information incorporation in the state. Moreover, it allows users to express their value for transaction inclusion; users who can generate more value from fast inclusion will invest more in latency reduction. However, everything else being equal, this signal is necessarily less information than a bid because in contrast to bidding, it happens ex-ante before the precise value for transaction inclusion is known to users.

If we agree with the three statements, it seems necessary to focus on hybrid policies that take both bidding and time into consideration: batch auctions and other bidding based procedures are in reality a particular instances of hybrid policies and FCFS is, in most cases , not appealing. But how should we choose, on economic grounds, among different policies that use combinations of time stamps and bids to order transactions?

A first step is to analyze the equilibrium bidding and equilibrium latency investment of users interacting with the policy and to derive the equilibrium welfare and revenue achieved in different designs.

Some disclaimers: The following analysis is purely economic. I abstract away from questions of consensus and implementation and assume that the policies considered can be implemented, because they are run by a trusted centralized sequencer or because we know how to decentralize these policies in a satisfying way. I abstract away from incentive compatibility problems (MEV extraction, censoring etc.) on the side of the party that implements the transaction ordering policy and assume that the policies are implemented as stated.

Equilibrium Analysis of Bidding and Latency Investment

The starting point of my analysis is a bidding and/or latency race between two bidders, who each want their transaction to be executed before the other bidder's transaction (I would expect similar results to hold for more than two bidders). A typical situation that triggers such race could, for example, be an arbitrage opportunity arising through a price discrepancy between an off-chain CEX and an on-chain DEX. Another typical example would be a competition for executing a liquidation. While there are other MEV games played in reality, where bidders have more complicated preferences over transactions orderings than just about how two transactions are ordered relative to each other, these atomic contest for earlier inclusion constitute a large fraction of trading activity on most platforms. Moreover, many other strategies contain an element of it, as it might be a necessary part of the execution of a more complicated trade.

I assume that the value of earlier execution can be different for the two bidders, for example they could have different amounts of liquidity deployed on different platforms so that they can extract different amounts of value from an arb. However, both of the two bidders should have a non-negative value for having their transaction be executed first. The situation is a race between the two bidders in the sense that the bidder, whose transaction is included later, cannot extract any value .

I assume that there are two sources of uncertainty for the two bidders:

The bidders are uncertain about the competitor's value of winning.

The bidders are uncertain about each others' latency. Timestamps are uncertain and random but correlated with (ex-ante) latency investment decisions.

Batch auctions

In a batch auction, all transactions arriving within a pre-specified time window (according to some time stamping scheme) are ordered according to their attached bid, with the highest bid transaction being executed first (if feasible), the second highest bid transaction second (if feasible) etc. A bid cut-off (or reserve price) can be used to bound the total number of transactions being included. A variety of payment rules can be used. In the following, I assume that the competition between the two bidders follow a first price auction format with two possible interpretations: 1) the payment rule is pay-as-you bid, but the capacity is bounded so that the lower bid transaction is not executed 2) the lower bid transaction is reverted if it does not land in the higher slot.

Qualitatively very similar results would also hold for an all-pay batch auction.

For the analysis I normalize time so that bidding happens during a unit time interval [0 , 1] [0,1] [0 , 1] . The timing is as follows:

An arbitrage opportunity arrives uniformly at random in the unit interval. When the arbitrage opportunity occurs, bidders learn their valuations. A player i i i has a valuation v i v_i v i to have his transaction included first, where v i v_i v i is distributed according to F i (x) F_i(x) F i (x) .

Bidders send a bid for inclusion. Depending on their latency and time of observing the arb, their bid gets included in the current or in the next batch.

At the end of the batc h, bids are evaluated according to a first price auction. The higher bid transaction is executed and pays the attached bid.

Assume that there are two bidders with valuations v 1 v_1 v 1 and v 2 v_2 v 2 , distributed, for simplicity, i.i.d. uniform on the unit interval. Qualitatively, the analysis carries over to non-uniform valuations. The assumption of independent valuations models the case where there is heterogeneity between the two bidders, while the common component in valuations is known with certainty.

Assume that bidder i i i has probability of T i (τ) T_i(\tau) T i (τ) of submitting a bid at time 0 ≤ τ ≤ 1 0\leq \tau\leq 1 0 ≤ τ ≤ 1 which is included in the current batch and probability 1 − T i (τ) 1-T_i(\tau) 1 − T i (τ) of being included in a later batch.

Probability T i T_i T i of inclusion in the current batch. Dotted lines indicate batch cut-offs.

W e can interpret T i (τ) T_i(\tau) T i (τ) either as objective probabilities or as bidders' (subjective) belie fs. I don't make any particular assumption on T i T_i T i other than that 1 − T i (τ) 1-T_i(\tau) 1 − T i (τ) is a differentiable CDF, so that T i (0) = 1 T_i(0)=1 T i (0) = 1 and T i (1) = 0 T_i(1)=0 T i (1) = 0 . It is straightforward to calculate equilibrium bidding strategies, both for 1) the situation where each bidder is certain about whether or not their own bid gets included in the batch but uncertain about the competitor's bid, and for 2) the situation where bidders are uncertain about their own bid and the competitor's bid:

Equilibrium derivation

In equilibrium,

bids are lower than in a standard first price auction,

bidders shade bids according to the odds 1 − T (τ) T (τ) \tfrac{1-T(\tau)}{T(\tau)} T (τ) 1 − T (τ) that the competitor gets included in a later batch,

the payoffs for bidders are higher than in a standard first price auction.

the revenue for the auctioneer is lower.

Ex-ante Latency Investment for Batch Auctions

Next let us look at the latency investment game induced by the batch auction design. Assume that prior to bidding, bidders invest in latency reduction. The bidders do not know the arrival time and the value of the arb when making the investment decision. The investment is therefore at the ex-ante stage. When making their investment decision, the bidders only know the expected ex-ante profit from the auction as analyzed above.

Assume that it costs C (Δ) C(\Delta) C (Δ) for a user to invest into latency so that he can send messages with delay Δ \Delta Δ . We assume that

it is impossible to send arbitrarily fast messages, lim Δ → 0 C (Δ) = ∞ \lim_{\Delta\rightarrow 0}C(\Delta)=\infty lim Δ → 0 C (Δ) = ∞ ,

it is possible without any additional investment into latency to send a bid at the beginning of the batch that is included in the batch, C (1) = 0 C(1)=0 C (1) = 0 ,

lower latency is more expensive than higher latency, i.e. C (Δ) C(\Delta) C (Δ) is strictly decreasing,

there is increasing marginal cost of a marginal time advantage, i.e. C C C is convex.

Suppose that bidder i i i invests into latency with delay Δ i \Delta_i Δ i so that he is able to bid on all arbs arriving up to time 1 − Δ i 1-\Delta_i 1 − Δ i . Assuming uniform arrivals of arbs, defining the probability that bidder j j j has delay at most Δ j \Delta_j Δ j by σ j (Δ j) \sigma_j(\Delta_j) σ j (Δ j) , bidder i i i 's profit of investing into latency to obtain a delay of Δ i \Delta_i Δ i is (see the analysis in the previous section):

∫ 0 1 − Δ i (1 2 − 1 3 T (τ)) d τ − C (Δ i) = ∫ Δ i 1 (1 2 − 1 3 σ j (Δ)) d Δ − C (Δ i) . \int_0^{1-\Delta_i}(\tfrac{1}{2}-\tfrac{1}{3}T(\tau))d\tau-C(\Delta_i)=\int_{\Delta_i}^1(\tfrac{1}{2}-\tfrac{1}{3}\sigma_j(\Delta))d\Delta-C(\Delta_i). ∫ 0 1 − Δ i (2 1 − 3 1 T (τ)) d τ − C (Δ i) = ∫ Δ i 1 (2 1 − 3 1 σ j (Δ)) d Δ − C (Δ i) .

From this we can construct a symmetric equilibrium of the investment game.

Cost of latency and the implied equilibrium latency distribution (prob that latency is smaller than Δ \Delta Δ).

Equilibrium derivation

Ex-ante investment into latency leads to zero average profit for bidders in equilibrium : The ex-ante expected gains from bidding in the batch auction are equal to the cost of latency investment of the bidder independently of the cost function.

Hybrid Policies: Time Boost

In a recent proposal, Offchain Labs describes a hybrid transaction ordering policy that orders transaction by time stamp of arrival at the (centralized) sequencer, but allows bidders to get a time boost by paying an additional fee. The time boost for paying a fee of F F F is

π = g F F + c , \pi=\frac{gF}{F+c}, π = F + c g F ,

with parameters g g g and c c c , and transactions are ordered by the score

π – t \pi-t π – t

where t t t is the timestamp of arrival of a bid at the sequencer and π \pi π is the time boost. The parameters have the following interpretation: Parameter g g g gives the maximal time boost, a user can get from bidding. In particular, transactions finalize after waiting g g g . Parameter c c c is the marginal cost per unit of time (normalized by g g g). The auction is all-pay: Bidders need to pay the time boost fee no matter how transactions are ordered. However, the analysis for a standard winner-pay auction looks very similar (I comment on this below).

Time boost formula (source: Arbitrum research)

Let us again consider the scenario where arb opportunities appear uniformly at random on a time interval and assume that the arb is only good for the earlier transaction. If a bidder has a lower score he will get a payout of zero, but still needs to pay.

The analysis of time boost is more complex than the previous analysis for the batch auction, since now the precise time stamps of the two bidders matter and not only whether the time stamp is before the batch cut-off or not. Thus, making the same analysis as previously, where the values and time stamps of the bidder follows some distribution is generally intractable. However, we can analyze the two extreme cases, where 1) the value of the arb is commonly known by the bidders, but the time stamps are uncertain, 2) the value of the arb to the other bidder is uncertain, but the time stamps are certain.

Known value, uncertain time stamp

Let us first analyze the case where time stamps are uncertain. They are generated according to a probability distribution, which is a function of the latency of the two bidders. The value v v v of winning is commonly known and the same to both bidders. Bidder 1 1 1 wins if and only if his score is higher

π 1 – t 1 ≥ π 2 – t 2 ⇔ π 1 + Δ ≥ π 2 , \pi_1-t_1\geq\pi_2-t_2\Leftrightarrow \pi_1+\Delta\geq \pi_2, π 1 – t 1 ≥ π 2 – t 2 ⇔ π 1 + Δ ≥ π 2 ,

where Δ : = t 2 – t 1 \Delta:=t_2-t_1 Δ := t 2 – t 1 is the difference in time stamps. Thus, it suffices to know the bids π 1 , π 2 \pi_1,\pi_2 π 1 , π 2 and the difference Δ \Delta Δ in time stamps to determine the winner. In particular, it suffices to know the distribution F Δ F_{\Delta} F Δ of Δ \Delta Δ (and not necessarily the individual distributions of the time stamps t 1 t_1 t 1 and t 2 t_2 t 2). I analyze the symmetric case where F Δ F_{\Delta} F Δ is symmetric and unimodal with mode 0 0 0 , which captures the case of equal latency distributions. Inverting the time boost formula, the fee F F F for buying a time increment of π \pi π is

F = c π g – π . F=\frac{c\pi}{g-\pi}. F = g – π c π .

If latency is noisy enough the model has a pure strategy symmetric equilibrium:

Equilibrium derivation

In equilibrium,

for low valuations neither bidder will make a time boost bid,

for high valuations bidders produce the same score on average,

the value threshold at which bidders start bidding is increasing in c c c and decreasing in g g g and increasing in the variance of time stamps.

To maximize bidding revenue, the parameter c c c should be chosen small and the maximal time boost g g g large.

Unknown value, known time stamp

In the previous section I assumed that the value v v v is commonly known. Now I consider the private value setting, analogous to our analysis of the batch auction. For tractability, I assume that the latency of the two bidders is common knowledge among the two bidders. This simplification is not innocent, but it is a reasonable approximation of reality when bidders interact repeatedly and can estimate the expected time stamp of the competitor with very good accuracy. Second, for simplicity, I consider a linear approximation for the boost formula:

F = c π g – π ≈ c π g . F=\frac{c\pi}{g-\pi}\approx \frac{c\pi}{g}. F = g – π c π ≈ g c π .

For a reasonably large boost parameter, e.g. g = 10 g=10 g = 10 when c = 1 c=1 c = 1 the marginal cost c g (g – π) 2 \frac{cg}{(g-\pi)^2} (g – π) 2 c g (which is relevant for deriving optimal bidding policies) is approximated very well by c / g c/g c / g and we can expect to get qualitatively very similar results for the true boost formula, as long as g g g is sufficiently large. In the following, I assume g ≥ c g\geq c g ≥ c . The equilibrium analysis now follows an all-pay auction with a head start for the lower latency bidder:

Equilibrium derivation

In equilibrium:

For low valuations neither bidder will make a time boost bid.

For high valuations, the bidders produce the same score for the same valuation.

The value threshold at which bidders start bidding is increasing in c c c , decreasing in g g g and increasing in the latency difference.

A lower latency bidder underbids relative to a standard all-pay auction, a high latency bidder overbids relative to a standard all-pay auction.

To maximize bidding revenue, the parameter c c c should be chosen small and the maximal time boost g g g large.

Both versions of the model, the one with uncertainty about time stamps and the one with uncertainty about valuations make qualitatively similar predictions. An immediate implication is that the parameter c c c in the time boost formula should be selected small to increase participation and revenue. In a similar way, the maximal time boost g g g should be selected large. However, the latter comes with trade-offs, as finalization of bids is slower. I comment on these parameter choices and compare the performance in the next section. The equilibrium would look qualitatively similar (pooling for small valuations and separation for high valuations) for a winner-pay instead of an all pay-auction: The threshold at which bidders start bidding is smaller in that case (the square root disappears).

An analysis of the latency investment game in the same vein as above for the time boost proposal seems intractable. However, the same high level logic should apply: latency competition will lead to a race to the bottom where all profits are competed away. I therefore conjecture:

Conjecture: For the time-boost proposal and many other hybrid polices, ex-ante investment into latency leads to zero average profit in equilibrium: The ex-ante expected gains from competing in the time boost auction are equal to the cost of latency investment of the bidder.

Comparing batch auctions and time boost

It is instructive to compare the relative performance of the two auction formats. To make the batch auction comparable to the time boost proposal, I assume that if neither of the two bidders make it into the batch because they have too high latency, then their orders are both processed in the next batch. We can then look into the equilibrium in either model as a function of the realized difference in latency Δ : = | Δ 2 – Δ 1 | \Delta:=|\Delta_2-\Delta_1| Δ := | Δ 2 – Δ 1 | between the two bidders where Δ i \Delta_i Δ i is the delay of bidder i i i when sending a bid to the sequencer.

First, let us look at allocative efficiency: how likely is it that the higher valuation bidder wins the race? In the batch auction this can only happen if the two bidders end in different batches and the faster bidder has the lower valuation which happens with probability Δ / 2 \Delta/2 Δ/2 . Under time boost, this can only happen, if the slower bidder refrains from bidding (which in our equilibrium happens if his valuation is below the threshold u : = c g Δ u:=\sqrt{\tfrac{c}{g}\Delta} u := g c Δ) while having a higher valuation. This leads to the following observation:

Under the batch auction design, the likelihood that the high valuation bidder loses is half the latency difference Δ 2 \tfrac{\Delta}{2} 2 Δ .

Under the time boost design, the likelihood is half the cost of compensating for the latency difference c g Δ 2 \tfrac{c}{g}\tfrac{\Delta}{2} g c 2 Δ .

The likelihood of the higher valuation bidder winning is higher under time boost if and only if the marginal cost of bidding is small c g ≤ 1 \tfrac{c}{g}\leq1 g c ≤ 1 .

Next let us look at the welfare loss relative to the first best where the item is always allocated to the higher valuation bidder: under which design is this welfare gap larger? In the first best, the surplus is the expectation of the maximum of the two valuations which is 2 / 3 2/3 2/3 for the i.i.d. uniform case.

Derivation of welfare loss

Under the batch auction design, the welfare gap to the optimum is Δ / 6 \Delta/6 Δ/6 .

Under the time boost design, the welfare gap is 1 6 (c g Δ) 3 / 2 \tfrac{1}{6}(\tfrac{c}{g}\Delta)^{3/2} 6 1 (g c Δ) 3/2 .

The welfare gap is smaller with time boost, if and only if the marginal cost of bidding is small c g ≤ 1 / Δ 3 \tfrac{c}{g}\leq1/\sqrt[3]{\Delta} g c ≤ 1/ 3 Δ .

Next let's look at bidding revenue for the auctioneer. The ex-ante revenue from bidding is the sum of payments received.

Derivation of revenue

Under the batch auction design, the revenue is (1 – Δ) / 3 (1-\Delta)/3 (1 – Δ) /3 .

Under the time boost design, the revenue is (1 − (c g Δ) 3 / 2) / 3 (1-(\tfrac{c}{g}\Delta)^{3/2})/3 (1 − (g c Δ) 3/2) /3 .

The revenue is higher under time boost if and only if the marginal cost of bidding is small c g ≤ 1 / Δ 3 \tfrac{c}{g}\leq1/\sqrt[3]{\Delta} g c ≤ 1/ 3 Δ .

The comparisons indicate that the time boost proposal outperform the batch auction among several dimensions, if the marginal cost of bidding in time boost is small. There is, however, a catch to this: Decreasing the marginal cost by increasing the boost parameter g g g , leads to longer finalization times of transactions which has other downsides, some of which I have previously discussed. Moreover, it is a somehow unfair comparison, as I have normalized the size of batches to 1 1 1 , whereas the g g g parameter, which plays a similar role as the batch size, is allowed to be larger than 1 1 1 .

To get an intuition what happens with variable batch size ,note that if a unit time interval is subdivided into two batches while there is still one arb per unit of time, then the likelihood that the lower valuation bidder wins the race doubles. Similarly, the welfare gap grows and the revenue decreases in the number of batches. Thus, choosing larger batches has a qualitatively very similar effect as choosing a larger maximal time boost g g g .

The previous analysis is stylized but has immediate implications for economic design:

If the time boost design is implemented, special attention should be put on the choice of parameters. Choosing the marginal cost of bidding too high is detectable in equilibrium . In that case we can predict little use of time boost bidding (no bidding rather than producing low signals) and the parameters should be adjusted.

While latency competition is much less severe for the batch auction than in a FCFS design, there are still advantages of low latency bidding. Towards the end of a batch, bidders with a faster connection can underbid relative to optimal bidding in a standard first price auction , since there is a substantial likelihood that slower competitors do not make it into the current batch. This has an adverse effect on efficiency and revenue. Moreover, the equilibrium analysis predicts that all profits of the bidders are competed away through latency investment by the bidders in the ex-ante stage before the actual bidding.