Latency is still a major principle in the world of MEV (Maximal Extractable Value). As block times get shorter on newer and newer chains, the blockchain emulates the first-come, first-served (FCFS) ordering system seen on off-chain centralized exchanges as searchers fail to keep up. Building low-latency systems for on-chain trading and arbitrage is a multi-faceted problem. However, likening it to traditional trading systems helps to understand each part.

The Traditional Trade Lifecycle

Traditional limit order book exchanges host two-sided auctions that represent the market of some asset (with respect to the value of another asset). Sellers create asks for the price they want to sell at, while buyers create bids to buy at a certain price. When sellers and buyers cross, the intents are matched and orders are generated, providing information that can be useful for predicting market trends.

The functional components of a trading system - Developing High Frequency Trading Systems by Rossier et al.

New ask and bid listings, along with information about the last trade price and size, are emitted as events/messages to connected clients, such as market makers and hedge funds. These clients generate a trade decision, be it an order placement or cancellation, based on the information received. The decision is then sent back to the exchange venue to be executed.

To simplify, the steps involved can be bullet-pointed as follows:

- Exchange sends events to connected trading systems
- Trading system processes events to generate the set of actions it wants to perform on the exchange to maximise returns
- Actions are sent back to the exchange to be executed

Exchange sends events to connected trading systems

Trading system processes events to generate the set of actions it wants to perform on the exchange to maximise returns

Actions are sent back to the exchange to be executed

The Trade Lifecycle On-Chain

Trading strategies executed on-chain typically follow a similar flow:

- Trigger propagation:

A node (or broadcaster) propagates either a pending transaction received from a user or a block of transactions proposed as the block proposer to connected peers and systems. This triggers the strategy to know when to run.

- Strategy execution:

Trading systems process events to generate a set of operations/transactions they wish to perform on the blockchain to maximise returns.

- Transaction propagation:

Transactions are propagated around the blockchain network or through other means (e.g PBS supply chain) to be received and executed by the next block proposer.

Trigger propagation:

A node (or broadcaster) propagates either a pending transaction received from a user or a block of transactions proposed as the block proposer to connected peers and systems. This triggers the strategy to know when to run.

Strategy execution:

Trading systems process events to generate a set of operations/transactions they wish to perform on the blockchain to maximise returns.

Transaction propagation:

Transactions are propagated around the blockchain network or through other means (e.g PBS supply chain) to be received and executed by the next block proposer.

A trading strategy may consider working off a pending transaction (e.g.,sandwiching, backrunning, etc.) or a proposed block (e.g., interest rate pushing borrow position into unhealthy position, rebase hopping) as its trigger event. Latency can play a monopolizing role in accessing the opportunity based on the time remaining to bid.

Solutions such as orderflow auctions (OFAs) and encrypted mempools aim to democratise both types of opportunities respectively by running sealed bid auctions off-chain and having arbitrageurs specify their intent to participate through bidding instead of latency-related investment.

Defining Performance Metrics

To remain competitive, trading firms monitor latency as one of their key performance indicators (KPIs). There are many sources of latency that exist in traditional trade lifecycles, including:

- Order validation by the exchange
- Exchange connectivity bandwidth
- Types of operations on the hot code path
- Latency from I/O

Order validation by the exchange

Exchange connectivity bandwidth

Types of operations on the hot code path

Latency from I/O

Similar sources of latency can be identified in blockchain-based systems as metrics to optimise. Below, I define four such metrics:

- Trigger propagation latency:

the time it takes for a trigger event to be sent from a node, block proposer, or off-chain service, and then received by the trading system.

- Tick-to-trade:

a metric borrowed from traditional finance. It measures the time it takes for a trading system to receive the trigger and then generate the trade decisions (transaction or transactions as a bundle) to execute.

- Transaction propagation latency:

the time it takes for a transaction or bundle to be sent from the trading system and received by the next block proposer (pre-PBS), the block builder (PBS), the rollup sequencer (layer 2) or auction decider (OFA).

- Supply chain latency (PBS):

the time it takes for a transaction or bundle, upon receipt by the builder, to be validated, included in the built block, and then sent across a relayer to the validator. This metric is recorded from the time of receipt by the builder to when the validator receives the block.

Trigger propagation latency:

the time it takes for a trigger event to be sent from a node, block proposer, or off-chain service, and then received by the trading system.

Tick-to-trade:

a metric borrowed from traditional finance. It measures the time it takes for a trading system to receive the trigger and then generate the trade decisions (transaction or transactions as a bundle) to execute.

Transaction propagation latency:

the time it takes for a transaction or bundle to be sent from the trading system and received by the next block proposer (pre-PBS), the block builder (PBS), the rollup sequencer (layer 2) or auction decider (OFA).

Supply chain latency (PBS):

the time it takes for a transaction or bundle, upon receipt by the builder, to be validated, included in the built block, and then sent across a relayer to the validator. This metric is recorded from the time of receipt by the builder to when the validator receives the block.

Time (in ms) that each node could see a transaction from the transaction originator - Strategic Latency Reduction in Blockchain Peer-to-Peer Networks by Tang et al.

Various approaches are being looked into to minimise each. For example,minimising the number of peer-to-peer (P2P) hops on the network layer reduces the latency induced by intermediate nodes that have to validate a transaction before rebroadcasting it. For traditional finance system engineers, this should bring to mind several techniques for optimising the network layer, such as co-location and adapted clients for networking, among others. Some research is being looked at to reduce latency further up the supply chain as well, for example, optimistic relays.

Optimistic Block Submission - Optimistic Relay Proposal by Mike Neuder, Justin Drake and AlphaMonad

Conclusion

This piece aims to discuss the possible segmentation of latencies that exist in the trade lifecycle for on-chain trading. These latencies can be defined as metrics that can be examined more closely. A number of optimisation techniques can be applied to each segment, which can be further discussed.

Acknowledgements

I would like to thank mempirate for comments and discussions on my earlier drafts.

I enjoy discussing topics such as this and am always happy to chat and collaborate on writing future pieces together.

My Twitter can be found here.