

The End of Time

June 13th 2011

Once Upon a Time... our markets consisted of a single CLOB (Central Limit Order Book) by country. Absent competition, these order books were slower and participants' fees were higher. Such CLOB markets worked with strict 'Price-Time' priority. Additionally, in many continental European markets there were concentration rules mandating broker-dealers to execute client orders in the CLOB – so the Price-Time priority in the CLOB was to a significant extent the only game in town (although with support for iceberg orders, some would characterize it as 'Price-Display-Time').

As we're all familiar with, MiFID changed our market structure by sweeping away the concentration rules mandating use of a single CLOB by country, and allowing the emergence of multiple competing PLOBs (Public Limit Order Books). This has resulted in a period of intense competition and rapid innovation, driving dramatic reductions in trading tariffs, huge improvements in system performance and capacity, and the emergence of dark Midpoint orders books, and so on and so forth. And as brokers have developed the technology to participate in multiple lit PLOBs and dark midpoint books, they have also deployed internal crossing networks where they seek to internalise customer flow before or in parallel to routing it to external venues.

How have all these developments impacted Price-Time priority in our market?

Most individual PLOBs operate with Price-Time priority (we'll get to those that don't a bit later), but the fact that there are more than one means that brokers need to calculate in which price-time queue the speedy execution of their own limit orders is more certain.

Imagine a multi-lane motorway (each lane is a Price-Time queue of a PLOB), with traffic (limit orders to Buy) queuing up to pass through a set of toll booths (marking the Best Bid in each venue). Coming from the other direction are the Sell orders, also queuing in the same number of lanes for each booth. Booth operators (exchanges and MTFs) are supposed to keep their particular queue moving in a fair and orderly manner. Happily, the collision of a Buy and Sell order results in a Trade (which is published) rather than a car crash, and such collisions happen when somebody pays the toll (the venue's fee and the spread) to cross and meet the oncoming other queue. So, if you're in a hurry you can jump to the front of the queue by setting a new best Bid or best Offer, or you can pay a premium for immediacy by submitting an aggressive order.

Where there are multiple venues with the same displayed price, how do brokers decide which ones to access? We should expect a SOR's venue selection to be driven by cost, certainty of execution, and (possibly) market impact.

- Cost includes the explicit tariff for aggressive flow and also the related post-trade costs.
- Certainty of execution determined by a variety of factors including the broker's latency to the venue (both for inbound market data and order routing), the average lifespan of limit orders at each venue, and the share volume or number of different orders/participants at the BBO. To estimate certainty, some brokers measure the historical success rate of capturing a targeted bid/offer price when routing to each venue, whilst others use displayed size or a venue's share of trading as proxies for this.
- Market impact in this context depends on whether there are differences (by venue) in the propensity of prices to rebound (or fade further) after being hit by a marketable order.

Assuming you don't want to cross the spread, and you have decided to join the queue with a particular limit price, which queue is the right one to post your limit order in?

In the beginning brokers either chose which venue(s) to post to on the basis of each venue's overall share of trading in the stock (or group of stocks) – a bit like joining a queue because everyone else is (very English), and trusting in the wisdom of the crowd. Others set specific ratios for each queue, and in doing so some brokers were doubtless influenced by the payment of rebates for getting to the front of MTF queues and by a desire to stimulate and support competition amongst venues.

But as both brokers and their clients have become more sophisticated, so the criteria for venue selection have evolved. It's increasingly the case that brokers are applying a variety of predictive signals to determine which queue they can get to the front of quickest. What factors are they considering, and how do they capture these in their SOR decisions? Basically – how long is each queue and how fast is it moving?

- When setting a new EBO, brokers can jump to the front of any queue they choose. But their choice still matters, because if their new price is subsequently matched on other venues, they want to ensure that they're in a queue where the book is moving quickly – a venue that's reliably attractive to contra-side aggressive flow.
- If joining an existing queue, they need to consider the length of the queue in relation to the speed at which it's moving. This similarly depends on the arrival rate of contra-side aggressive flow.

In choosing where to post their limit orders, brokers have to predict the behaviour of prospective counterparties aggressing the market. For example, markets with lower take-fees, more participants and lower latency may enjoy more success in attracting aggressive flow, and hence become more attractive venues for posting.

But that's not the whole picture – the twin forces of competition and technological innovation have changed the landscape in two ways that arguably reduce the certainty of execution for publically displayed limit orders (and hence reduce the incentives to post them):

- There are a whole bunch of other queues that you can't see – effectively another private motorway next to yours that you may or may not be entitled to use. You're waiting patiently in your queue on the public highway, and you see reports of other people's executions on the private motorway, but the public traffic doesn't seem to be moving.
- Having reached the front of your chosen public queue, you're expecting to trade when an incoming contra-side order crosses the spread. Instead, somebody sneaks ahead of you at the last second, or you get only a partial execution as some of it is allocated to the people behind you.

Both brokers and exchanges have started to monetize time/place priority as the valuable commodity that it is...

First, brokers...

- With SORs in place, brokers had the opportunity to introduce their own crossing networks (whether ATNs or BCNs) without doing a disservice to their clients (previously an order kept 'upstairs' could not easily also be represented in multiple other places). With "take" fees being high on US exchanges and ECNs (relative to equivalent fees in Europe), it made particular sense for brokers to internalise marketable flow. Some firms already had the necessary internal market making capabilities, some acquired the capability by buying specialists in the field, and others approached the big market makers active in public markets and encouraged them to do the same thing in their own pools.

- Whilst many market makers are pro-transparency and are, in principle at least, against internalisation, getting a chance to intercept liquidity before your competitors was a fairly compelling opportunity. Brokers realised that the customer flow they were executing was a valuable commodity that could be monetised by offering electronic market makers an earlier opportunity to interact - in essence selling those market makers 'time/place priority'. So whilst electronic market makers may have displaced the traditional market making businesses of many large banks, within brokers' own liquidity pools they have become a source of revenues and/or cost savings.
- The same is beginning to happen in Europe, although the ambiguity surrounding provision of 3rd party "non-discretionary" access to Broker Crossing Systems acts as a partial brake on some firms.

And then exchanges/ECNs and MTFs...

As competition has heated up, exchanges have also started to experiment with different routing or "allocation" models (most of which are breaks from the traditional price-display-time model).

- NYSE's model offers Designated Market Makers and Floor Brokers '*parity*' - the ability to participate in a trade even when they're not at the front of the queue. The exchange dilutes time priority for normal participants (as just under two-thirds of the liquidity they might have captured in a strict price-time model is instead allocated elsewhere) in return for the fees and committed liquidity they get from the DMMs. This model has recently attracted regulatory criticism, although it's worth noting that diluting price-**time** priority lowers the importance of pure speed as a market advantage (instead it's about your relationship with the exchange) – and hence does not necessarily favour HFT firms. Such models exist because the committed liquidity that DMMs bring can give the exchange a competitive advantage.
- DirectEdge's "flash orders" were seen by some as an attempt by the market to subvert price-time priority and instead 'sell' priority to a selective subset of customers.
- NASDAQ's PSX introduces size priority, allocating incoming contra-side shares pro-rata to the displayed size of participants at the BBO.
- The arrival of Taker-Maker books (in which a rebate is paid for *removing* liquidity) can be understood as an attempt to create a venue in which limit orders enjoy superior time (or price) priority over those posted in venues that charge for liquidity removal.

Some exchanges are arguing against internalization on the basis that (as a result of reducing the certainty of execution for public limit orders), it will reduce incentives to post limit orders in public order books and lead to a vicious circle of widening spreads and increasing internalisation. Basically they appear to be against the dilution of price-display-time priority. And yet the exchanges have also exacerbated the dilution of price-display-time priority through the launch of multiple competing order books and alternative "allocation" models.

All of this makes for a more complex market structure than either participants or regulators are accustomed to, and means we'll probably be debating these topics for some time.

- Since the regulators have clearly opted for a competitive (and thereby fragmented) market structure, should we worry about the *degree* of fragmentation?
- Is the continual evolution of technology leading us towards a much more distributed market model in which the role of displayed order books is less prominent?
- How much volume do we need in public order books to provide reliable price formation?
- To what extent does price discovery rely upon there being price-time priority within the market?

- And if the death of price-time priority does ultimately undermine the efficacy of the price formation process, then “who done it”?
 - Was it regulators with the introduction of RegATS, RegNMS and MiFID to stimulate competitive (and fragmented) markets?
 - Was it exchanges introducing innovations that that give ‘first look’ or privileged interaction rights to a subset of members?
 - Was it brokers leveraging their SOR investments by introducing their own dark pools to which they give precedence over public markets?
 - Was it exchanges operating multiple order books with differential tariff models?
 - Or, like Murder on the Orient Express, were we all guilty?