

## **F** What you see is not what you get The challenge of truly 'smart' order routing



**Will Winzor-Saile**  
Electronic Execution  
Product Specialist

In the world of equities, the multi-market landscape is well established. It goes without saying that before any order is sent to a market, the trader – whether human or algo – looks across all available pools of liquidity and decides on the optimum venue of execution. This decision used to be simple – which venue has the best price or which has the most volume? But as markets evolve, we can no longer rely solely on visible information to make these decisions. In the first paper of this series, *Shifting Sands – the harsh realities of executing in today's markets*, Will Winzor-Saile explored the issues around constructing a reliable, globally-consistent market access infrastructure, and looked at how the more enlightened brokers are starting to fundamentally rethink their approach to electronic execution. Here, he focuses on a fundamental part of this, the smart order router (SOR), and the challenges facing firms as they struggle to stay on top of the ever-changing market landscape.

There are many decision points in the life of an order. At the start of the day, the portfolio manager decides *what* to trade and passes this on to the trader – or the algo – who then breaks the order up over the course of the trading day making the decision about *when* to trade each slice. The job of an SOR is simple: take those slices and decide *where* to trade them.

Often the SOR will be coupled with, or indistinguishable from, an execution strategy which may hold on to the order until some optimal time, or test dark pools to determine their volume before releasing it. These actions are separate from the fundamental objective of an SOR which

is to find the best venue to trade an order right now. It needs to find the best price or the lowest trading fees, achieve the most rebates or minimise market impact, but above all else it has to get the order filled. Achieving this simple objective, however, is incredibly complex.

The SOR is a vital part of a firm's market access infrastructure and, as such, it needs to be global, provide a normalised view of markets across all regions and remove the multi-market complexities from upstream systems. It also needs to be regional, understanding the nuances of every market – the available order types, queue priority, varying liquidity of stocks,

and so on. Walking the line between global consistency and regional specificity is just one of the challenges facing SOR developers.

## Everything is relative

The world of science has a huge amount to offer financial markets. Techniques from quantum physics are being used to model and predict movements in stock prices, and neutrino-based communication promises to cut the latency between New York and Sydney from several hundred milliseconds to around forty. Little wonder, then, that in the UK at least, around 10% of all physics graduates end up working in finance, myself included.

Even on a day-to-day basis we can still learn from a few of the basic principles. Relativity teaches us that there is no such thing as a universal frame of reference; what you observe depends on where you are. Since an SOR relies on information from many different places, this makes life very difficult.

Most SORs start by combining order books from every exchange into a single, consolidated view of the markets. Since each underlying exchange is located in a different city, a different data centre, or at the end of a line with a different latency, the way these combine will vary between locations. Consequently, each SOR is looking at a completely different view of the market. This is not some minor eccentricity that can simply be corrected by comparing timestamps and measuring latency.

Where two exchanges are geographically apart, the idea of synchronicity becomes meaningless.

London and Frankfurt are around 640km apart. Even ignoring the latency of fibre or microwave transmitters, light (and therefore information) takes nearly 2 milliseconds to travel between the two. If orders appear on each market within 2ms of each other, there is no meaningful way to say which arrived first - that depends purely on which exchange you're closest to.

The further apart markets are the more pronounced this effect becomes. While already an issue in Europe, it's less of a problem in the US where equity exchanges are generally located at relatively close proximity to one another. However, with the growth in cross-border trading of fungible stocks in Canada and South America this looks set to change. Regardless of latency or strategy, two identically-programmed SORs located in different places could react very differently to the same market signal. This inescapable fact exerts itself more profoundly as markets spread more widely across the globe.

As a result of this, no single view of the market is 'real'. Deploying an SOR is not just about introducing more intelligence to the strategy or reducing the latency of comms lines, it's about understanding the objectives of the firm and appreciating that where an SOR is located makes a real difference in terms of the market that it sees.

## Myths and legends

Let's ignore this effect for now and pretend that all execution venues are in one place and all market participants in another. Now everyone has the same view across the markets and reacts to the same set of information and the only difference is how long it takes to get there.

Understanding latency and the impact it has on a trading strategy is essential in designing and operating an effective SOR. Some firms focus an enormous amount of effort on latency and base their business proposition on being the fastest. Others operate long-term strategies and have no interest at all in time-to-market. The majority of firms are somewhere in the middle. They understand latency and try to reduce it, but as part of a wider strategy rather than as an end in itself.

The question these firms tend to ask is, "how fast is fast enough and at what point does the investment required to reduce latency outweigh the benefits?". As important as it is to understand these points, they are secondary to the more fundamental questions around what latency you are reducing and why. To address these questions properly we first need to dispel a couple of persistent myths.

First, that tick-to-trade is only relevant for HFT firms. Tick-to-trade (the time between seeing a price appear on the market and being able to enter an order against it) is the only important latency

measurement regardless of your strategy. Being able to send an order to the market in a few microseconds is of little use if your view of the market is already out of date. As obvious as this may seem, it's often overlooked. A broker reducing his market access latency from 5ms to 1ms might see this as an 80% improvement, but if his market data, algo, SOR and risk checks take a combined 35ms, then this quickly drops to just 10%.

Second, that reducing my latency stops my orders being gamed by predatory HFTs. Although there is some substance to this, it's not the only solution, nor is it the best. Michael Lewis's 'Flash Boys' famously highlighted the issue with an order being split across two different venues and arriving at different times. If the time between the two orders arriving is long enough, then the signal from the first order can cause the market to move away from the second. While reducing latency does shorten this gap, thus reducing the chance of signalling, a suitably intelligent SOR can negate this effect regardless of latency.

So what's the real motivation for reducing latency? That's straightforward enough: to make sure that what you see is what you get; any strategy is useless without a true view of the market and the higher your latency the less accurate that view becomes. Of course, how accurate that view needs to be depends on your trading strategy.

## Need for speed

Consider a retail trader. The price he sees on a screen is typically guaranteed for around 30 seconds, so any time taken to make a decision and submit an order to the broker is insignificant, he will always get the price he saw. Looking at the other end of the scale things are very different. If a high-frequency trader spots a good price (or wants to cancel a bad one) then the chances are another HFT firm has spotted it as well. The result is a head-to-head race in which a few nanoseconds determine whether he makes the trade or misses it.

The rest of the market lies somewhere in the middle with most brokers generating orders throughout the day according to some strategy outlined by their client. As each order is produced, it is sent to an SOR which will try and trade it for the best available price at that time. Certain algos will be attempting to hit specific prices as they come up, others will be just trading according to a pre-defined volume curve. These actions are very similar to those of an end user, but make a big difference to the impact of latency.

The SEC's Market Information Data and Analytics System (MIDAS) offers a wealth of data on the lifetime of orders on US markets. It shows that over 18% of orders that fill are under 50 milliseconds old, so if your latency is 50ms and you're trying to hit a specific order, there's only an 82% chance of it being filled before your order arrives. Admittedly this is something of an over-simplification, but it gives you a good

idea of how quickly the reliability of the data drops as latency increases.

For more passive strategies, the impact is far less pronounced. In the above example, the algorithm was reacting to a signal and was racing to trade a specific price. If, instead, the strategy is generating orders according to a pre-defined schedule, the chance of the market moving drops to under 1% in the same time. Although small, this is still a measurable impact and can make the difference between beating and falling short of a key price benchmark.

Measuring this impact is not always easy. Traditional benchmarks such as VWAP and Implementation Shortfall will be impacted by the performance of an SOR, but the effect is so overwhelmed by the impact of the volume curve that it's hard to distinguish. Even lower level metrics such as Spread Capture and Market Impact focus on the execution strategy rather than the SOR. While it's possible to directly measure the SOR by comparing the price, volume and time taken to execute each slice, the effort required to do this means it is often overlooked.

Even without huge geographical distances between exchanges, comms lines and system latency can have an immense impact on the reliability of market data and therefore the performance of an SOR. Not only is everyone looking at a different view of the market, but without a performant system that view is inaccurate a significant amount of the time.

## Dark matters

Even in an imagined world with all of the exchanges in a single location, as long as there's latency between the market and the traders, the prices seen are not always the prices you can hit. The next logical step is to ignore latency so now every market participant sees exactly the same set of prices. If they can see them, they can trade them. This is a step in the right direction, but there's still a large section of the market where volume remains hidden - dark pools, for example, where an entire order book is maintained without publishing any prices, or hidden orders on a lit order book.

Hidden orders may be iceberg orders, where only a small volume of a much larger order is shown, or completely hidden order types where nothing appears on the book. They may be native (operated by the exchange) or maintained on an external system. On top of this, different exchanges handle hidden orders in different ways, some giving visible orders priority while others preference hidden orders. These differences may be subtle, but they make an enormous difference to the way an SOR needs to interact with the exchange.

If there are only 100 shares visible at a certain price on the market and the SOR needs to trade 150, then it has a choice to make. If there's hidden volume at that price, the best strategy is to submit the full 150 immediately and take both the visible and the hidden volume. If there's no hidden volume, then it should take the 100 and

wait to see if any more volume appears at the price.

Of course there's no way to tell if there is hidden volume until it trades, by which time it's too late. The best an SOR can do is to try to predict hidden volume based on past behaviour, but even then measuring hidden volume after it's traded is not easy. Often trades against hidden orders are not specifically flagged so their presence needs to be implied from discrepancies between quoted and traded volumes. Our own analysis suggests that for liquid stocks this is around 25% of the visible volume. Back to the world of physics where similar discrepancies in astrophysics were used to imply the existence of 'dark matter' - also about 25%.

## A new approach

In financial markets, seeing isn't always believing. A single, consolidated view of the market doesn't exist and, even if it did, every millisecond of latency noticeably reduces its accuracy. On top of this, even the lowest latency systems can still only see 75% of the market with the rest concealed behind hidden orders.

This means that SORs can no longer rely solely on the information they see. A deterministic strategy that splits an order based on visible volume is likely to miss hidden volume or be gamed by other market participants. Instead, an SOR needs to take an analytical, predictive approach to look at how the market has been performing and calculate the probability

of the price moving and the amount of hidden volume on each market.

By understanding the reliability of the data, and by analysing market trends and microstructure, firms can add real value through their SOR. Not only will they avoid missed volume, but they will be able to leverage hidden volume and alternative liquidity pools to provide price improvement and more efficient alpha capture. All this comes at a cost, however. A flexible, analytical SOR may perform better when it's working, but unless it is part of a resilient global infrastructure it will be rendered useless.

A clever algo is no good without a decent SOR, but even the best SOR is no good without reliable, performant market access and market data.

## About Fidessa

Eighty-five per cent of the world's premier financial institutions trust Fidessa to provide them with their multi-asset trading and investment infrastructure, their market data and analysis, and their decision making and workflow technology. With around \$20 trillion worth of transactions flowing across our global network each year we offer unique access to the world's largest and most valuable trading community of buy-side and sell-side professionals, from global institutions and investment banks to boutique brokers and niche hedge funds. A global business with scale, resilience, ambition and expertise we have delivered around 25% compound growth since our stock market listing in 1997 and we're recognised as the thought leader in our space.

As markets continue to evolve towards electronic trading across asset classes and across geographies, this systematic movement means that technology must present a simplified view of these global markets. Fidessa's electronic execution capabilities – covering more than 200 markets and multiple asset classes – allow users to outsource the commoditised aspects of execution, such as exchange interfaces and connectivity, as well as taking advantage of execution tools including smart order routing, smart crossing and self-trade prevention. Underpinned by our latest next-generation technology and unrivalled market expertise, our electronic execution services provide brokers with consistent access to global markets through a normalised trading interface with simple integration to their own systems. Available as a complete service or as separate modules, brokers are empowered to offer a tailored, differentiated service to their own clients in the most cost-efficient manner.