BMO Capital Markets
Quantitative Execution Services

Hitchhiker's Guide To Smart Order Routers

September 2, 2011

Quantitative Execution Services

(416) 359-5743
qes@bmo.com

Rizwan Awan, CFA
(416) 359-5195
rizwan.awan@bmo.com

Benjamin Chiu
(416) 359-4151
benjamin.chiu@bmo.com

Jeremy Dietrich
(416) 359-5692
jeremy.dietrich@bmo.com

Andrew Ng
(416) 359-8692
andrew.ng@bmo.com

Andrew Karsgaard
(416) 359-7670
andrew.karsgaard@bmo.com

Introduction

The Canadian trading landscape has come a long way from the days of a centralized exchange. In the last few years, a number of marketplaces have opened up in Canada to increase competition in the space. The transition from a centralized exchange to multiple marketplaces has meant a wholesale upheaval in legacy technology on many trading desks. New cottage industries have popped up to service niches that did not exist before. Smart Order Routers, consolidated quotes, dark pools have become part of our everyday lexicon. Participants, and even end investors, are forced to consider the impact of frequent market structure changes on child order placement logic, or parse the details of marketplace pricing strategies to take advantage of the buffet of pricing models available. High Frequency Traders are a permanent fixture in the landscape.

This paper explores the world of Smart Order Routers (SORs) and how this once obscure piece of technology is now a critical component of every trading desk that deals with multiple marketplaces. We will look at the various considerations that go into SOR technology and try to sift out fact from fiction as well as outline some best practices.

Introduction

The trading landscape is becoming more fragmented over time as multiple marketplaces pop up. As a result, Smart Order Routers (SORs) have become a vital component of trading technology in the modern marketplace. SORs try to simplify the details of market minutiae by allowing traders to trade multiple marketplaces as if they were a single order book.

SORs have three primary goals:

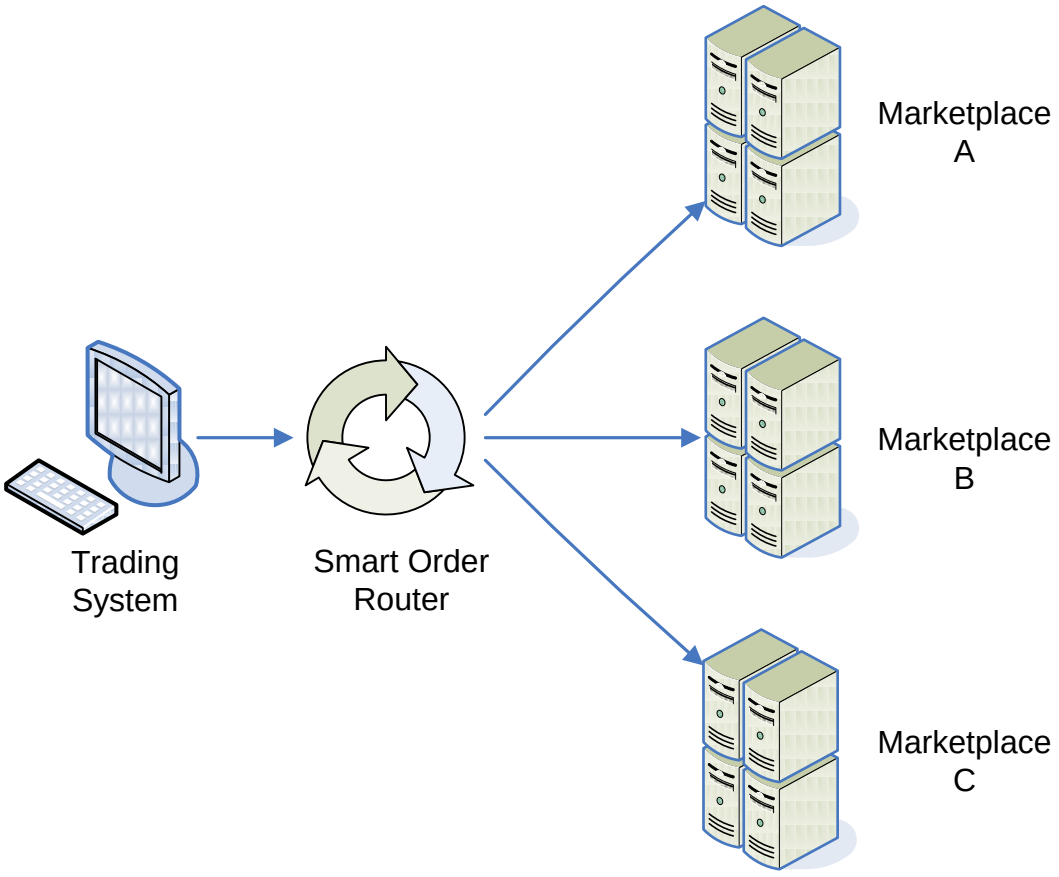
1) **Link** - Link multiple marketplaces in a fast, seamless and efficient manner

2) **Capture** - Maximize liquidity capture across all accessible pools of liquidity

3) **Hide** - Minimize information leakage

We will evaluate all three of these goals and the various practices used by SORs to help better achieve the stated goals.

Figure 1: Conceptual setup of a Smart Order Router connected to three marketplaces.



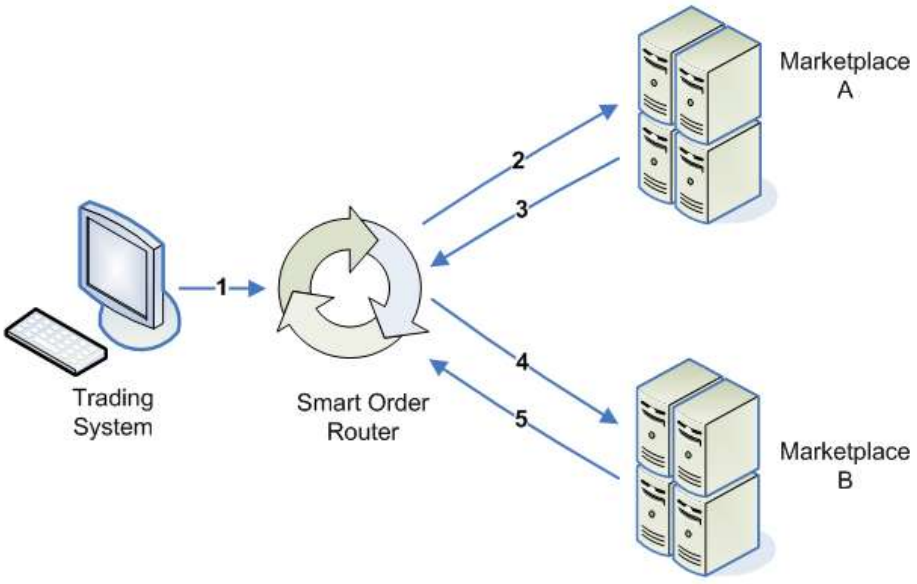
Link

One of the primary goals of SORs is to link multiple marketplaces in a fast, efficient, and seamless manner. In doing so, there are a few considerations that need to be taken into account.

Spray vs Serial Routing

In the early days of SORs, the first generation of routers typically accessed the markets in a serial fashion. Orders would hit the markets in a sequential or 'serial' manner, waiting for a fill from the first marketplace before sending an order to the second (see figure 2 below). While this was a simpler SOR to program, it had a major drawback in that High Frequency Traders (HFTs) could detect the fill on the first market and cancel their orders from the other marketplaces before the second order made it there.

Figure 2: Steps in a serial SOR route.

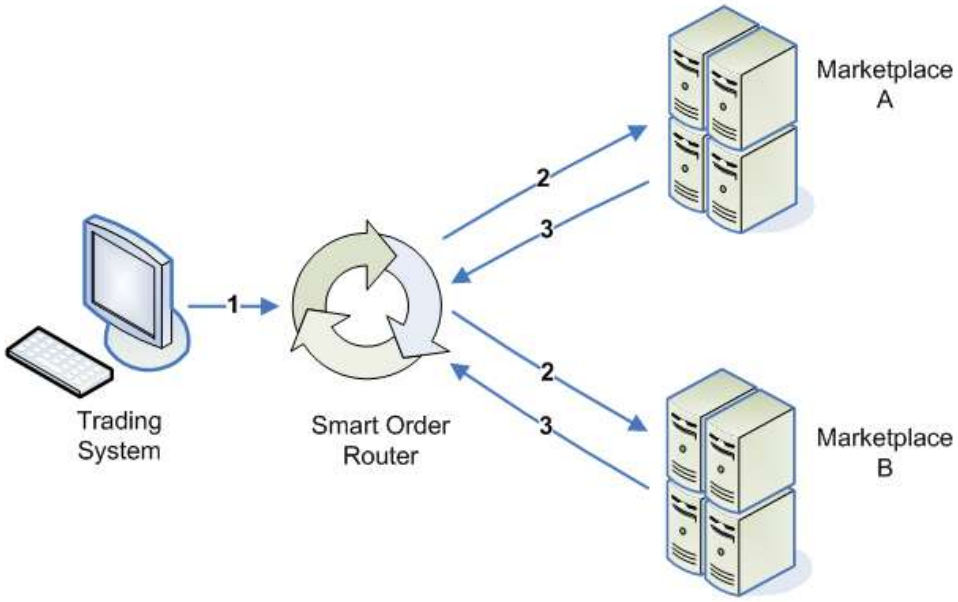


SORs initially tried to mitigate this problem by speeding up connections to the various marketplaces and writing smarter (more efficient) routing code. However, given their serial nature, this was akin to getting a more efficient horse buggy in the age of jet engines. Needless to say, the serial mode of routing went the way of the Dodo.

The newer generation of SORs all use 'spray' routing where orders are simultaneously sent to all the marketplaces at once (see figure 3 below). The nuances lie in the sizes sent to the various marketplaces (especially in orders that outsize the NBBO). SORs simultaneously try to hit all pools of liquidity at once and thus eliminate the ability of HFTs to beat the secondary orders to other marketplaces.

We will revisit some of these information leakage concepts under the 'Hide' section.

Figure 3: Steps in a spray SOR route.



Top of Book vs Depth of Book Routing

One of the common issues with spray routing is determining whether to spray to the top of book vs depth of book.

In a top of book route, a SOR can spray an oversized order to the first price level in the marketplaces, wait for the fills to see if there are any icebergs, and then carry on spraying down to the next level and so on. HFTs can fade their orders in between the various levels of route since waiting for fills at each level is essentially a serial route (Hence top of book routing is sometimes referred to as a serial spray). The price improvement offered by any potential hidden icebergs has to be weighed against the reduced certainty of execution for this style of routing.

In a depth of book route, the SOR foregoes any iceberg price improvement opportunities and routes directly to the maximum available depth (using bypass orders) to complete the order. In this scenario, there is greater certainty of execution albeit at an inferior price as hidden portion of icebergs are foregone.

There are routers that will allow the trader to customize top of book vs depth of book routes (and levels in between, ie hybrid combination to spray a certain number of levels before waiting for fills to spray again). This flexibility allows traders to choose between certainty of execution (depth of book) vs potential for price improvement (top of book).

Dynamic Routing

One of the holy grails of SORs is maximizing internalization of order flow. While holding up orders and automated order matching is currently not allowed in the Canadian marketplace, SORs can utilize broker-preferencing functionality on marketplaces to try and maximize intra-broker crossing. Dynamic SORs can scan the resting passive orders on the various marketplaces and dynamically change routing preferences to go to markets that have the same broker's passive contra order. For brokers that primarily represent natural flow, this avoids unnecessary intermediation by HFTs as orders get crossed with other natural flow.

Internalization reduces both implicit and explicit costs of trading. The passive order doesn't get crowded out and as a result saves the spread more frequently. This saves both the implicit spread cost as well as the explicit active take fee since orders are less likely to cross spreads and trade actively.

SORs should also utilize the encrypted marketplace IOIs provided by dark pools to dynamically change routing to access dark pools that have contra liquidity. This can avoid costly routes that ping dark pools where there is no liquidity present.

Capture

In an ideal world, traders would like to capture the bid-ask spread as much as possible. It is thus important for SORs to represent the order intelligently in order to maximize the potential to get a passive fill. This is where order management becomes important not only on just local markets but also interlisted markets.

Passive Order Management

One of the best methods for a SOR to maximize liquidity capture is to represent an order on all marketplaces. When an active order comes to market, being represented maximizes the opportunity to get filled. This strategy does come with some issues that pertain to order management. It is extremely difficult to automate an efficient order management system since it requires knowledge about the state of all the child orders (most SORs typically just route an order and have a simple one-to-one mapping between the incoming and outgoing order). There are other practical considerations: what happens if an order keeps getting filled on one market and as a result the balance of the order is represented in dormant markets? What about the impact of placing multiple orders on the visible NBBO to represent orders in multiple marketplaces?

SORs should have the ability to manage the orders in the marketplace in an intelligent manner. At the very least the orders should be split and represented in marketplaces based on their historical share of trading volumes. Advanced SORs will look at real-time heatmaps and adjust participation on markets that have been active in the recent timeframe and avoid unnecessary over-advertisement of orders in dormant markets. In both scenarios, once an order is filled, the remaining orders have to be rebalanced across marketplaces.

Interlisted Routing

Interlisted routing is perhaps one of the most important aspects of SORs in the Canadian marketplaces. Canadian stocks that are interlisted can be traded on both sides of the border since they are fungible; the only quirk in doing so is managing currency risk. This is where interlisted SORs come in handy to efficiently automate realtime currency hedging for any order that gets routed across the border.

We want to highlight that well over half of all volume on interlisted stocks trades in the US markets and not having a SOR that can access that liquidity places a severe handicap on trading. In a prior paper, we attributed about 16% of all Canadian volume to interlisted arbitrage activity (see Interlisted Arbitrage: http://qes.bmo.com/papers/6_BMO_Interlisted.pdf). This volume only exists because an order is being arbed between the Canadian/US markets at a profit and there is unnecessary intermediation by the arbitrageur.

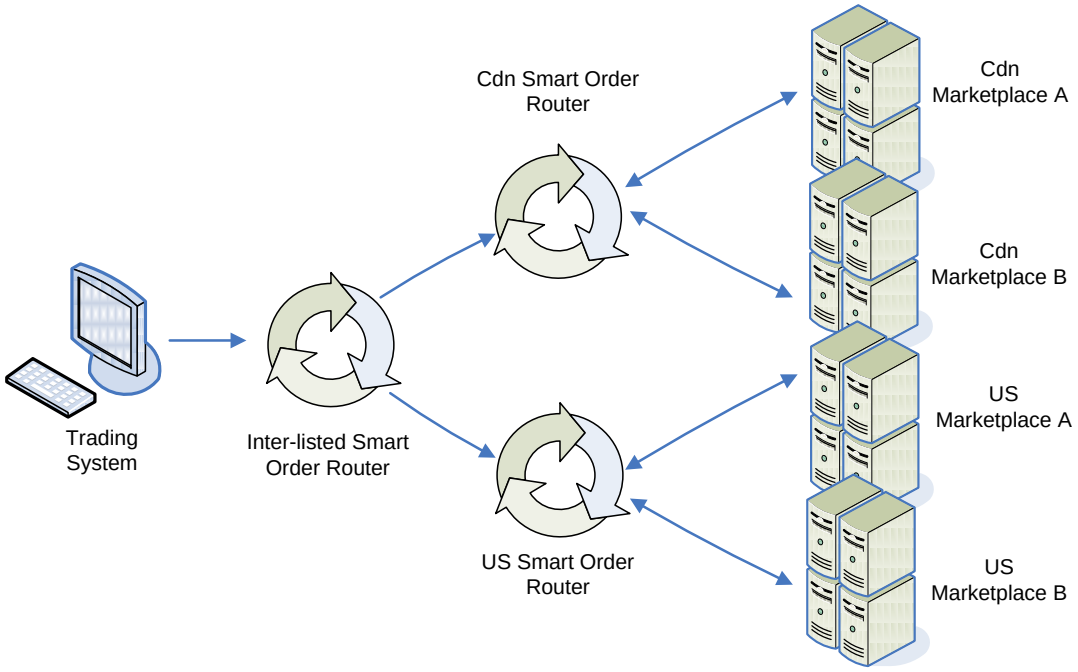
It is interesting to note that interlisted SORs were the very first SORs in the Canadian marketplace and existed even when we only had one centralized marketplace in Canada. However, with multiple marketplaces on both sides of the border, it is important to have an interlisted SOR that works orders in both markets together by building a consolidated order book across North America.

There are two major types of interlisted SORs:

- Interlisted SORs that employ specialized local SORs behind the scenes (see figure 5)
- Interlisted SORs that manage both markets (see figure 6)

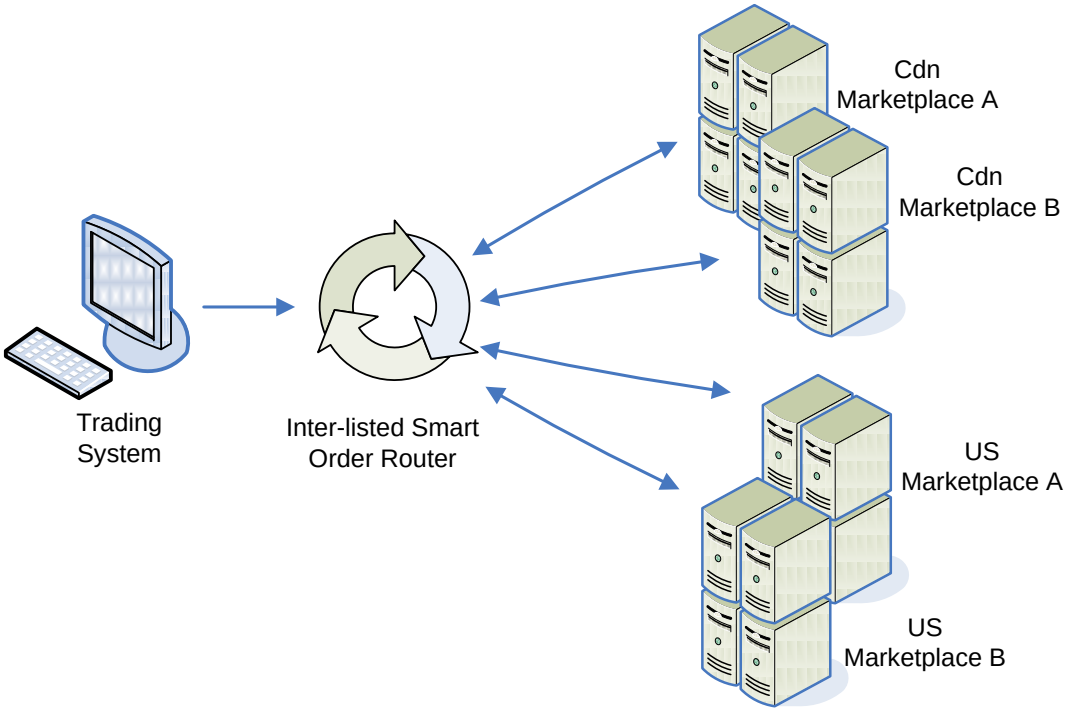
Interlisted SORs that use specialized SORs behind the scenes have the advantage of getting highly sophisticated country specific SORs optimized to work with the nuances of the markets they operate in. However, the downside is that the parent SOR can not coordinate end child orders between the two SORs (since it can only control the sub-SORs) and it is tricky to manage multiple orders dynamically in both markets due to the lack of control at the parent SOR level.

Figure 4: Interlisted router with sub-routers.



Interlisted SORs that operate in both markets directly have the advantage of being able to control all child orders and as a result have precise control over order management. These SORs can coordinate orders to avoid some of the pitfalls (such as fading interlisted arb quotes) associated with interlisted SORs (we will discuss these below). The added flexibility and control makes this architecture far more powerful but does add complexity to the design.

Figure 5: Interlisted router with direct connections.



There are definitely some quirks that need to be accounted for in interlisted SORs since price differences are not always in $0.01 increments (they are typically fractions of that given the FX rate fluctuations). Consider the case where an interlisted SOR routes to a better price in the US market for 2000 shares at a fraction of a penny price improvement only to realize that the 10,000 shares in the Canadian market disappears as a result of interlisted arbs removing their orders. The serial nature of top of book routing, as discussed previously, makes it almost impossible to use an interlisted SOR and we suggest a full depth route for interlisted SORs to avoid fading quotes from interlisted arb machines.

Another common issue plaguing interlisted SORs is the hold-up of the order while an FX transaction takes place. This introduces latency into the route which can have adverse impact as it gives HFTs a chance to fade their quotes. In a fast moving market, the split second hold up can be a severe handicap.

Hide

One of the biggest obstacles that SORs face are fleeting quotes from nimble High Frequency Traders (HFTs) as they look to collect rebates and minimize trading risk. Once they buy (sell) shares on one marketplace, HFTs typically race to cancel all their orders and send contra sell (buy) orders instead to hedge their risk immediately. The lightning fast speed at which this takes place is now measured in sub-milliseconds and often appears as 'electronic front-running' since the contra orders appear ahead of the next piece of the SOR slice. This is a clear sign of information leakage during the routing process.

The bad news is that HFTs will keep engaging in this behaviour; however, the good news is that it can be avoided. 'Smarter' SORs can mitigate some of this behaviour and help improve liquidity capture.

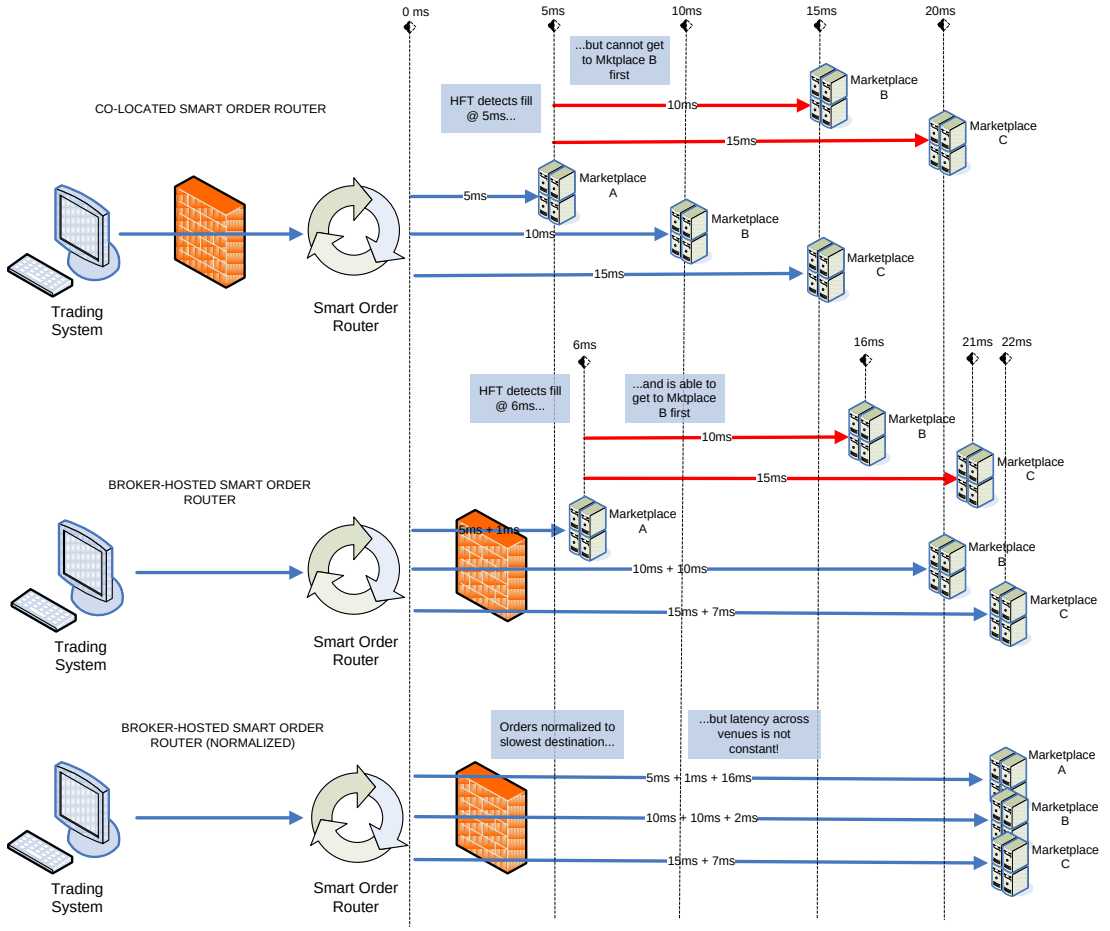
Latency Normalization

Due to the nature of connectivity to the various marketplaces, there is a discrepancy between the time it takes for an order to reach the different marketplaces. In some cases, this latency discrepancy can be extreme. Even though there are different latencies to the marketplaces, this is not an issue for SORs as they can beat HFTs to all marketplaces before any information is leaked (see 'Co-Located SORs' in figure 6 below).

Where the issue comes in is when there is *variable* latency introduced into the connection due to infrastructure delays. This is typically inherent in SORs that are located in-house (as opposed to co-located) and can as a result leak information like a sieve (even with spray routes). The 'Broker Hosted SOR' section of figure 6 below highlights the same setup as before with the exception of variable latency of 1ms, 10ms, 7ms introduced to marketplaces A, B, C respectively by internal infrastructure and shows potential information leakage to HFTs.

One of the strategies employed by SORs to mitigate any arbitrage issues is latency normalization. SORs that employ this strategy hold up orders on faster connections to try and match the time it takes for the slowest order to reach its destination (see the 'Broker Hosted SOR (Normalized)' in figure 6 below where the latency is normalized by adding 16ms, 2ms, 7ms to marketplaces A, B, C respectively). While in theory this sounds like a great idea, there is an inherent problem with this solution… infrastructure latency to marketplaces is not a static number and can change as network traffic conditions change (hence variable latency). As a result of this variable latency, SORs combat it by using dynamically changing delays to reflect current network conditions… a better solution but still not ideal since it is difficult to accurately predict network traffic conditions.

Figure 6: Colocated vs broker hosted vs broker hosted (normalized) SORs.



A better solution to this problem is to avoid placing the SOR within the firewalls altogether and instead co-locate the SOR outside the latency prone IT infrastructure. Direct connections do not have variable latency and as a result are not prone to information leakage during the route (see 'Co-located SOR' in the figure 6 above). The variable latency portion is on the incoming order to the SOR and makes no difference to the outcome of the route (orders always reach their destination in a determined timeframe after being routed out from the SOR).

Co-location

One of the common misconceptions on co-location is that it is used simply to be closer to a market-place and as a result be faster. While it is certainly true that a co-located router will have faster direct connections, it is a negligible delay for firms that have their own servers within a walking distance of the co-location site. A bigger reason for co-location in these conditions is to place trading engines outside of latency prone in-house network infrastructure.

To highlight this point, we use the Hubble space telescope as an example. The reason why Hubble produces clearer images than ground based telescopes is not because it's closer to the stars (it is, but in relation to the distance to the stars it's a rounding error) but instead because it sits outside of the earth's atmosphere that distorts light and as a result produces inferior imagery. The internal IT network latency is the atmosphere in our example while co-locating the SOR (the telescope) is akin to placing it outside the atmosphere allowing for direct connections to the marketplaces. Normalization in this example is the equivalent of image correction for ground based telescopes to account for the atmosphere… better than no correction but not the same as placing it outside the atmosphere altogether.

Thus co-location becomes an important part of any SORs ability to combat information leakage during the routing process.

Other Considerations

Optimizing Trading Fees vs Best Execution

One of the common dilemmas when configuring routing preferences is optimizing execution fees vs best execution obligations. It should be clear that best execution duties trump any optimization of trading fees and fees should only be used as a tie-breaker.

However, given the subjective nature of 'best execution', it is important to have clear guidelines on routing protocols. Does size trump price? Should dark pools be heavily employed both passively and actively? What about de-preferencing trading venues that primarily cater to HFTs? These are some of the considerations to keep in mind when setting routing preferences.

Exchange vs In-House vs 3rd party SORs

There are three categories of SORs in use today; exchange based, in-house, and 3rd party. Each category has its pros and cons.

Exchange based SORs are relatively straightforward to set up and don't require additional infrastructure or technology. Exchange based SORs also take away latency from themselves as they are built right into the trading engine. Exchange based SORs can also be cognisant of hidden icebergs and take advantage of them although there is debate on whether this creates an unlevel playing field and if the marketplace has an obligation to broadcast the iceberg information to all SORs. The major downside of exchange based SORs is the limited routing capabilities and lack of advanced functionality such as interlisted routing.

In-house SORs offer complete flexibility but come with the downside of time and costs associated with constant maintenance and upkeep. Also, in-house SORs require specialized skill-sets that might not be readily available for development. This option is only feasible for firms that are committed to building and maintaining a top tier SOR.

3rd party SORs are licensed out from vendors and brokers. This model results in a commoditized SOR but is typically best of breed (most functionality and easiest to maintain). The lack of flexibility is offset by quicker time to market and better maintenance due to specialized dedicated teams maintaining the SOR.

Each category has common functionality and vendors often differentiate their SOR by adding functionality above and beyond what is the typical role of a SOR. Examples include the incorporating risk management features such as position limits and fat finger detection.

Conclusion

The topic of SORs is deep enough to have dedicated books written on it. We have attempted to touch on some of the salient points on what is involved in SOR technology and covered the basics of some of the issues at hand today.

We hope that this paper sparks discussion on some of the elements of SORs and how they can be used to streamline and optimize trading. While we have come a long way from the early days of the first generation SORs, there is still a lot of work to be done to get to the moving target that is the perfect SOR. No one SOR in the market is perfect and each has its own strength and weaknesses.

One thing is clear, the ever evolving nature of the marketplaces is going to mean continuous development of newer, better, and smarter SORs in the coming future.

As always, questions/comments are welcome!

Rizwan Awan, CFA
Benjamin Chiu
Jeremy Dietrich
Andrew Ng
Andrew Karsgaard
Nikhil Kanwar