Is the National Best Bid or Offer (NBBO) being Ignored?

Publication Date: 07/13/2011
http://www.nanex.net

Executive Summary

The NBBO lies at the heart of Regulation NMS (Reg. NMS) and is the key concept that assures investors are getting the best price when buying or selling stocks. However, due to the recent industry trend that emphasizes speed at all costs, the NBBO, in practical terms, no longer exists. There is no audit trail that can show definitively whether an investor received the best price on their trade. The regulators do not appear to understand the root of the problem because they continue to promote new regulation, when a simple and effective solution exists: enforce Reg NMS. Likewise, getting rid of quotes that only serve to manipulate others is as easy as enforcing Section 9 of the Securities Exchange Act of 1934.

If any new regulations are needed after enforcing existing ones, we think a minimum quote life of 50 ms makes the most sense. Just the discussion of implementing such a rule would expose how deeply flawed the system is today and would be sure to raise a lot of eyebrows. But to be clear, we do not advocate new regulation.

The Death of the NBBO

As the lifetime of a quote approaches zero, the arguments for and against a minimum-quote-life rule become more interesting. Let’s suppose, for example, that the day has arrived where a few of the top HFT systems are able to send and cancel quotes in 1 nanosecond (ns). For reference, light travels 30 cm (1 foot) in 1 ns. At this rate, we could see 1 billion quotes per second per stock (qpss). A thousand active stocks trading at this rate would generate 1 trillion quotes per second (qps).

We admit that was extreme, but given the hyperbole from some HFT marketing groups, we couldn't resist. So let’s slow things down by a factor of 1,000 and imagine a world where HFT systems can send and cancel a quote in 1 microsecond (us). At this speed we could expect 1 million qpss and a thousand stocks would generate 1 billion qps. Put another way, an active market with a speed limit of 1 microsecond would generate 1 billion quotes per second, requiring everyone receiving CQS to spend about 1,000 times more for telco and equipment.

Still extreme? Let's make things 1,000 times slower again and imagine a world where HFT systems can send and cancel a quote in 1 millisecond (ms). At this speed, we could expect 1,000 qpss and a thousand stocks would generate 1 million qps. Even at this rate, many quotes will expire before leaving exchange networks. Today we frequently see qpss rates of 2,000, with peaks in the 5,000 - 30,000 range. These numbers are growing at alarming rates, and within a year, if left unchecked, recipients of CQS will need to upgrade their telco to 10 gigabit.

It's helpful to keep in mind that section I.C.4 of Reg NMS (page 30) states:

> *Accordingly, one of the Commission's most important responsibilities is to preserve the integrity and affordability of the consolidated data stream.*

And from the same document, page 410:

> *But in those limited contexts where the interests of long-term investors conflict with short-term trading strategies, the conflict cannot be reconciled by stating that the NMS should benefit all investors. In particular, failing to adopt a price protection rule because short-term trading strategies can be dependent on millisecond response times would be unreasonable in that it would elevate such strategies over the interests of millions of long-term investors – a result that would be directly contrary to the purposes of the Exchange Act.*

Perhaps we need to modernize the definition of a quote, or maybe we need a new name for what used to be called a quote. Not long ago, when a trader (or auto-trading software) received a quote marked auto-execute, there was a reasonable expectation of hitting (trading at) that quote — the only real exception being that another trader might beat them to it. Today, under the same circumstances, there is a significant chance that the same auto-execute quote would have already expired in transit and not be honored; the speed-of-light just isn't fast enough.

"No one uses the SIP for the NBBO anymore"

The change in semantics becomes most important when we look at the definition of the NBBO. Per Reg NMS, the NBBO for a stock is defined to mean the best bid/offer sent by a market center to the Security Information Processor known as the SIP (CQS, UQDF). But no one uses the SIP for that anymore, we are often told, which would seem to violate the letter and spirit of Reg NMS. We'd like to note that last year, 2.5 million subscribers spent over $450 million to receive and process CQS.

So what do they use for the NBBO if not from the SIP?

Each exchange computes the NBBO internally from direct connections to other exchanges. As the speed of trading increases, the likelihood of any two exchanges having the same NBBO decreases. Most of this is because of the pesky speed-of-light limitation.

So how does a trader know whether a trade was routed properly to the exchange with the best price?

He doesn't. It is impossible.

You see, each exchange’s view of the other exchange prices only exists in memory on that exchange's machines. It is not recorded. There is no audit trail. Sure, each exchange provides book-level data, but that only includes prices for that one exchange — not the prices that existed on the other exchanges at the time of each order.

"It is impossible to verify that a trade received the best price"

If we are going to allow machines to trade faster and faster, then at the very minimum, there must be audit trail data that includes each exchange's view of top-of-book quotes for every linked exchange trading that stock. In other words, what now only exists in an exchange routing computer's RAM, needs to be captured and made available. The exchange routing simulation video below may help you visualize this. In the video, each box represents one exchange and the price information it has from the other exchanges -- that is the information which needs to be recorded to assure trade through price protection. Essentially, each interlinked exchange would end up recording its view of every exchange's top-of-the-book prices -- exactly what the SIP (the box at the bottom) does now.

Reg NMS has already addressed this issue: page 32:

> *If the benefits of a fully consolidated data stream are to be preserved, each consolidator would need to purchase the data of each SRO to assure that the consolidator's data stream in fact included the best quotations and most recent trade report in an NMS stock.*

"The claim that aggregation causes the SIP to run slower is absurd"

So why not use just use the SIP? Auditing, maintaining, and improving one system is much better than 14 (one for each exchange). Well, because as some, including the SEC, have said: The SIP is not as fast because it has to aggregate information from the other exchanges.

Wait a minute.

How can an exchange ensure every order receives the best prices if it doesn't aggregate information from all the other exchanges? Take a close look at the simulation video and note that every exchange must essentially replicate the SIP's aggregation function. There is no way around this. Which means the claim that aggregation causes the SIP to run slower is absurd. Quote rate overload is the primary cause of latency in the SIP and direct feeds.

Besides, there are numerous places where Reg NMS language is expecting that order routing would reference the SIP such as this (emphasis ours) on page 314:

> *For instance, modifications to **order routing and execution systems will need to** be made to route and execute orders in compliance with the requirements of the Rule to prevent trade-throughs of protected quotations (which include, for instance, the ability to **recognize quotations identified in the consolidated quotation system** as manual quotations on a quotation-by-quotation basis).*

And on page 423, the use of the adjective displayed referring to prices in the SIP:

> *Intermarket sweep orders must, by definition, be routed to execute against the full displayed size of protected quotations...*

There is more.

A computer that aggregates information from other computers uses an aggregation policy that details how it selects between messages coming in at the same time from multiple computers (things like time-slice period, scheduling quantum, size of input queues, overflow behavior, etc.). Let's just say it's very complicated and has many dependencies: it's easy to make mistakes (or hide bias). The aggregation policy is important because it affects the 3rd criterion in NBBO selection: price, size, time. The SIP's aggregation policy is something that is fairly easy to verify. How easy is it to obtain and verify the 14 aggregation policies used by the 14 other exchanges? How do you know if exchange X's aggregation policy treats other exchanges equally?

It gets worse.

We have discovered a few anomalies when analyzing aggregation characteristics of CQS, which is the SIP for stocks listed on NYSE, AMEX, and NYSE Arca (home of many ETFs). One disturbing anomaly is that the SIP appears to be ignoring the timestamp in the quotes it receives from an exchange, using instead the actual time the SIP receives the quote, even if the timestamp in the quote is over 5 minutes late. That is, we found an example of the SIP treating 5 minute old data as real-time, affecting the NBBO in thousands of stocks. We will be publishing the results of our analysis shortly. The main point, is that this type of detection is impossible to carry out for individual exchange aggregation behavior; and if it was, it would take 14 times more work.

"We found an example of the SIP treating 5 minute old data

as real-time, affecting the NBBO in thousands of stocks"

Even more disturbing, the SIP applies the timestamp to quotes and trades after these messages have been throttled and queued, so it is impossible to detect or measure latencies occurring within the system. We have published numerous studies on this problem because it was one of the prime factors in the Flash Crash on May 6, 2010. Recipients couldn't detect significant delays until the system suddenly became overloaded. If timestamps reflected the actual time a quote was generated, we would have known how dangerously close to saturation the SIPs had become weeks earlier. Sophisticated trading firms with direct exchange feeds had to be aware of this.

We understand how difficult it can be to grasp these problems. Understanding complex networked systems where the speed-of-light is the dominant source of latency is hard. When the fate of your nation's financial infrastructure is at stake, we think regulators should have a solid understanding of these issues.

Inquiries: **pr@nanex.net**

Publication Date: 07/13/2011
http://www.nanex.net