

## Exchanges in a Race to Zero Latency

Traders Magazine Online News, September 25, 2009

Peter Chapman

How low can you go?

Nasdaq OMX Group said earlier this month that upgrades to its technology have made it the fastest exchange in the world. That may or may not be true but regardless, the exchange operator's announcement highlights a drive this year by market centers to reduce the latency of their systems.

Behind the trend is the desire to appeal to such latency-sensitive traders as direct market access and algorithmic players. These folks which include high-frequency traders, bulge bracket prop desks and the electronic trading departments of large broker-dealers want their orders processed as quickly as possible.

"If we're not building a low latency competitive exchange, we're just not going to be in the game," Brian Hyndman, senior vice president for transaction services at Nasdaq, said at last week's Aite Group conference on high-frequency trading.

Nasdaq reported on September 9 that upgrades to its INET platform and other parts of the technology that underlie five of its trading venues have given Nasdaq an average latency of less than 250 microseconds. That's faster than BATS Exchange, which pioneered low latency trading. BATS announced in June it executes 80 percent of its orders in under 400 microseconds.

The upshot, according to Hyndman, is that latency has gone down, throughput has gone up and order acknowledgement times are more consistent. "We eliminated the outliers," Hyndman said. "That's very important to a lot of Nasdaq's customers."

This means, Hyndman explains, that a trader won't get an order acknowledgement in 250 microseconds on one trade and three milliseconds another time. "There will be no big spikes in standard deviation," the executive said.

(One millisecond equals one thousandth of a second. One microsecond equals one millionth of a second.)

Besides the changes to INET, whose main feature is the order-matching engine, Nasdaq also upgraded its network to 40 gigabits per second from a 10-gigabit connection; upgraded its hardware; and made a variety of other changes.

Nasdaq's announcement was just the latest. Ever since June, all five of the major U.S. trading venues as well as one in Canada have put out notices describing the steps they have taken to cut their processing times.

On the same day Nasdaq made its announcement, Chi-X Canada ATS, owned by Instinet, claimed it was the fastest market center in Canada.

The ECN said it boosted its capacity to be able to handle 175,000 messages per second, a 500 percent increase from previous capacity of 30,000 messages per second. Chi-X Canada has benchmarked its average response time for marketable immediate-or-cancel orders at about 350 microseconds. That's at least 10 times faster than any other major Canadian market center, it contends. Previously, Chi-X Canada's internal latency was pegged at 890 milliseconds.

"We need to keep up with our customers," Tal Cohen, Chi-X Canada's chief executive, said. "The drive for latency is not slowing down anytime soon."

Chi-X Canada, based on the same technology as Chi-X Europe, launched in February 2008. Cohen said part of Chi-X's strategy to drive latency lower is to run the system on "commodity" hardware. That way, the ATS can simply plug in the newer and faster boxes once they become available.

Both BATS and NYSE Arca also announced latency reductions this summer. BATS cut its average latency, or the amount of time it takes to execute an order, by 50 microseconds to 395 microseconds. It also announced that it can now convert an order into a quote for transmission on its market data feed in 631 microseconds.

NYSE Arca, a unit of NYSE Euronext, announced that its order acknowledgement time, the time it takes to confirm receipt of an order, had been reduced to under one millisecond for Tape A and Tape B names and under 650 microseconds for Tape C issues.

Perhaps the summer's most symbolic announcement came from NYSE Arca's sister exchange. The New York Stock Exchange said in July it scrapped its 33-year-old SuperDOT platform order delivery and processing system, as well as an internal routing system called Post Support System. In its place, the NYSE installed its Super Display Book system, technology based on NYSE Arca's trading engine.

The move cut the time it takes to execute an order from 105 milliseconds to five milliseconds, according to the exchange. That's down from 350 milliseconds in 2007. NYSE customers now get order and cancellation acknowledgements in two milliseconds, the NYSE added.

Five milliseconds is a far cry from Nasdaq's 250 microseconds, and NYSE executives acknowledge they still have work to do. Still, the rollout this summer of the Super Display Book system was the culmination of an 18-month project that saw the Big Board completely reengineer its underlying hardware and software architecture using technology it acquired with the purchases of Euronext, Wombat Financial Software and Archipelago. The NYSE completely replaced its order entry, order database and routing systems, market data systems and pieces of its post-trade system.

The move to revamp its dated infrastructure potentially opens doors that were previously shut to the NYSE. At least one bulge shop refused to send any of its algorithmic flow to the NYSE because it was too slow, an NYSE exec told Traders Magazine.

On the other side of the Hudson River in Jersey City, N.J., technicians at DirectEdge ECN have also been ripping out old and installing new technology.

DirectEdge has seen its share of trading volume shoot from about 5 percent a year ago to about 12 percent today. To deal with the increase in messages flowing through its systems and prepare for the future, the ECN chose to replace its messaging middleware, TIBCO. It chose faster technology from 29West, a relative newcomer to the business. DirectEdge says installation of 29West's technology has produced a "dramatic reduction in overall system latency" and increased its throughput.

"It allows us to use persistent messaging without the penalty," said Steve Bonanno, DirectEdge's chief technology officer. Persistent messaging is typically slower than non-persistent messaging, but offers guaranteed delivery. No messages are lost as they can be with non-persistent messaging. 29West's technology eliminates that latency "penalty," Bonanno explained.

DirectEdge is now able to guarantee its customers an order response time of 300 to 500 microseconds. That's measured from the time the order enters DirectEdge's gateway to the time the acknowledgment hits the gateway. Previously, response times of DirectEdge's three trading platforms were in the 1.2- to 1.5-millisecond range.

Although DirectEdge needed the changeover to 29West primarily for throughput, it couldn't ignore latency. "Speed is the toll you have to pay to be in this business," Bonanno said. "Being fast is a given. That has to be there."

All things are relative, of course, and speed is no different. Market centers are forever engaged in a ferocious battle with each other to win market share. In their desire to impress traders, they will often put out overly rosy latency numbers, critics charge. "There is a lot of confusion, and, in some cases, obfuscation about the actual latency at the venues," said Donal Byrne, chief executive of Corvil, a Dublin, Ireland-based maker of latency monitoring and management tools. "It is impossible to make apples-to-apples comparisons."

That applies to both the speed at which the venues disseminate market data as well as the speed at which they convert orders into trades, Byrne added. The problem is that the venues typically offer up an average number, which may be of little use. "If you measure it at one time, you get one number," Byrne said. "If you measure it a few seconds later, you get another." He believes trading venues should publish a schedule of numbers for all times during the trading day.

Doug Kittelsen, chief technology officer of execution management vendor FTEN, is also concerned. He believes exchanges should quote their latency data in the 95th percentile, or at the slow end of the spectrum, rather than the median or average, which is standard. "You want to see what the curve is like all the way through to the end," Kittelsen said, "not just the middle."