The Greed for Speed
Waters Magazine
October, 2005

Since the creation of auction-based financial markets, a competitive advantage has developed for those with access to the lowest latency, most accurate information and the ability to act on that information immediately. The advantages of speed are being seriously examined in today's market due to the incredible growth of algorithmic and black box trading operations, fueled largely by the recent hedge fund boom.

Analysts estimate that 75 percent of US buy-side firms currently use algorithmic trading, and this number is expected to increase to 77 percent in 2007. (Tabb Group, 2005).

While the number of actual traders executing "high-octane" trading strategies – index arbitrage, program trading, high-volume equities etc. – is small, the money invested in capital markets represents a high percentage of the world's wealth.

One of the primary impediments to high-speed trading is latency. Latency stems from delays occurring at multiple points along the trading cycle. Some solutions to the problem are hugely cost-prohibitive to all but the largest financial and investment institutions. Other solutions include technology upgrades and system tweaks that small- and mid-sized firms can afford.

Recent Advancements

There have been periods throughout history when technological advances have produced competitive advantages for certain market participants. For example, getting market data in "real-time" from ticker tapes was a quantum leap in its day. Similarly, the advent of SOES (NASDAQ's Small Order Execution System) produced a significant advantage for SOES traders, as they were able to get Level II data faster and more reliably than other market participants.

More recently, the advent of widely-available broadband and advanced system architecture has somewhat leveled the speed playing field and caused an attention shift to other important issues such as platform reliability and execution quality.

Where Can Latency Occur?

The initial source of latency in the trading cycle is the data feed. For example, certain data vendors have an inherent latency in their products, because often time there is a high processing overhead prior to the data being delivered to the clients. Pressure to reduce data feed latency has forced many vendors to offer new low-latency data services that connect clients directly to each trading venue,

bypassing the vendor's own data center, and reducing latency by several tenths of a second.

The next type of latency occurs as a trader waits for an automated trading system to implement a trading signal. The latency created here can be reduced with better programming and the design of more sophisticated algorithms.

The third, and perhaps most currently scrutinized form of latency, occurs after the order has been placed but before it is executed and entered into the order book. One UK-based market maker who trades heavily in arbitrage, placing up to100 orders per second states, "The average latency between the placement of each of my trades and the time they are executed is about 60-80 milliseconds. Since a couple of milliseconds can mean thousands of dollars, I needed a way to cut down that time," the trader said. "To reduce latency I've chosen to co-locate my server with my broker to stay on the low end of that range."

Other Factors

A Zurich-based financial technologist points out that, "Part of the responsibility to keep latency low resides with the broker. While end users can do things for themselves like co-locating their servers to eliminate transatlantic or even cross-country lags, brokers need to examine other parts of their trade execution systems to make the entire process run smoothly."

David Schehr, a research director for investment services at Gartner, Inc., comments, "For active traders buying and selling on a tick-by-tick basis, trade execution speed is paramount, but other kinds of latency such as the amount of time required for their models to generate new trading strategies, should not be ignored."

"What's more," Schehr continues, "there are thousands of longer-term investors and asset managers to whom micro-seconds don't matter. What is far more critical to them is cost-effective, anonymous and reliable trade execution. They would be better served by evaluating other ways to improve this process through better research, easy to use and flexible tools for pre- and post- trade analysis and better order management systems."

Examining the Solutions

One extreme (and expensive) example of how to reduce latency in the trade execution process is the billion-dollar hedge fund that buys a building adjoining an exchange. It relocates servers, programmers, traders and desktops and drops them all on top of a dedicated, high-speed, fiber-optic connection feeding right into the exchange floor's control room. This works well if you have bottomless pockets, but what if you're a mid-sized asset manager who relies on a broker for your market access?

There are some alternatives to relocating an entire operation and/or spending $70,000 a month to maintain dedicated data lines:

- Work with your broker to induce them to improve their own technology. One hedge fund manager stated that, “At one point I was placing about 40 more trades per second than my broker could handle. Sometimes that was good enough, and sometimes it wasn’t.” In his case he had a strong, long-term relationship with the broker and he helped the firm understand the kind of performance he required. The brokerage firm decided to dedicate the time and resources to improving their processing speeds and capacity issues and gained additional clients from the effort.

- Co-locate your trading model server, not your entire business. When moving an entire base of operations is cost-prohibitive, a more reasonable option that will shave crucial time from order execution is to co-locate your server. One London-based trader co-located his server in his brokerage firm’s US server site to avoid the transatlantic lag. “Many traders are wary of this solution because they’re afraid that their proprietary algorithms or their equipment will be tampered with,” the trader said. “It was worth it to me to take that risk. I’ve shaved 20-30 milliseconds off of each order and I don’t have the expenses of maintaining a full satellite office.”

- Choose your execution venue wisely. Many traders prefer to deal directly with ECNs because they are fully automated and often less expensive than dealing directly with an exchange. One California-based trader is, “…thrilled with the alliance of NYSE and Archipelago, because it means access to all of NYSE’s equities with Archipelago’s technology and pricing.”

A side note about exchanges: they are not equal in their ability to handle the demands of black box and algorithmic trading. For example, some exchanges can only process 200 trades per second in total. This can be the kiss of death for high-volume traders. Other exchanges – such as the TSX, which has a central limit order book – behave more like ECNs with regard to technology and automation.

- Rethink how to use FIX. FIX protocol has made order execution easy and standard, and the messaging specifications are relatively simple to implement. Rather than code directly to an execution venue’s API, consider using a FIX hub. A FIX hub manages multiple destinations through a single FIX API and connection. This requires a small upfront time and programming investment but will cut down on the overall latency and cost.

What’s Next?

Now that we have several ways to approach the problem of latency, the question has become: what comes after speed? If hundreds of black box systems are all competing to be first to the market with their triggered order flow and traders are actively reducing their own latency, then logic dictates that a point will be reached when the returns diminish to a level where shaving additional milliseconds is no longer sufficiently rewarded.

As the “arms race” for the least latency continues, we have to watch and analyze the adjunct trends and benefits that result from technological improvements – new trading opportunities, more efficient trading tools, truly real-time 24/7 global trading. Inevitably, another controllable factor will emerge within the process, and we’ll be off on another “race to the bottom.”

"R. Bruce Pitt is vice president of system engineering and one of the founders of Nexa Technologies, a provider of online and direct access trading solutions."