

new york
london
paris
mumbai
singapore

SECURITIES INDUSTRY | news

Visitors Per Month: 53,883
January 11, 2010

Budgeting for Latency: If I Shave a Microsecond, Will I See a 10X Profit?

[Katherine Heires](#)

As 2010 starts, it's no longer enough to be able to evaluate risks, check market conditions and adjust trading strategies fast enough to send out thousands of buy and sell orders a second.

If you want to beat the next outfit to the punch by a millisecond or a microsecond-as many high frequency traders aim to do-you're going to have to constantly wring out bottlenecks and delays in the way your instructions get sent to and executed by the various exchanges.

Enter the demanding exercise now known as latency budgeting.

High-frequency, low-latency trading now accounts for 70 percent of all trading in equities, according to Aite Group. The high-frequency trading unit of the Chicago hedge fund Citadel Investment Group raked in almost \$1 billion in trading profits in 2008 from the practice. And its 2009 numbers aren't in yet.

But the message is clear: Numbers like these mean that for those traders whose strategies require the highest speed of performance, every millisecond matters. And unless you manage by the numbers, lowering the latency in your trading systems every year, every month and every day, you will not be able to compete with better, more vigilant managers of latency "budgets."

It's not just time that must be budgeted. Time also is money. Buy side and sell side firms each must keep careful watch on IT costs. Because each microsecond saved has a calculable cost. And, in the words of the rock standard, it keeps getting higher and higher.

"The cost of lowering one's latency on any given component of the trading operation is rising. In particular, we see this trend in co-location costs and, for that reason, there is greater focus on the total latency management process," explained George Hessler, executive vice president at New York-based Lime Brokerage, an agency broker that works with many high-frequency trading firms.

Co-location has installation, maintenance and rental costs, as firms try to put their execution systems as near as possible to venues' matching engines. But every aspect of high-frequency trading matters.

"Latency budgets only truly came into vogue in the last four or five years. That's when we saw purely electronic trading or trading without a person in the mix," says Matt Meinel, global director of business development at 29 West, a Chicago-based provider of ultra low latency messaging platforms and solutions. Once automated trading systems got combined with electronic exchanges, the performance limits associated with human interaction were removed, he said. And the bar inevitably got raised on trading performance.

TIME IS TIGHT

Simply defined, a latency budget is kept in order to assess, monitor and then manage how much time a firm spends on each and every step of the trading process, in both seconds and dollars.

For a broker-dealer or hedge fund measuring equity trading, what gets watched is the round-trip time of an order-how long it takes for trading data to travel from your servers to an exchange's matching engine for execution and then back again.

This is sometimes referred to as end-to-end latency. The less latency, the faster trades are executed and the more trade fills you will see.

Alternatively, if you are an exchange, you are analyzing latency from the time you receive an order to the time you send a filled order back. This is door-to-door latency.

Managing a latency budget means determining precisely how much time in milliseconds or less is taken up by each component in the trade process and then determining if one wants to or can reduce latency at any step along the way.

For instance, says Meinel of 29West, a hedge fund may determine that it is not seeing as many trade fills as it would like on a particular high-speed trading strategy. The fund might want to lower latency in the hopes of filling more trades or completing more trades at better prices and thus, in either case, upping firm profits.

Technicians then get to set up the steps to track in the latency budget, arrive at averages for the current times each step takes to complete and help select stages where they judge if there is room for improvement. These can be, for instance, the time it takes data to travel from an exchange's matching engine to the servers on a fund's premises or the time it takes for a packet of market data to be received by a feed handler to the time it takes for that data to be acknowledged by the firm's messaging software.

According to John Panzica, vice president of facilities manager Switch and Data's financial services practice, this establishes a firm's "latency value chain."

Every step involves a combination of hardware and software, from time spent running through middleware, applications, servers, other processing equipment, network routers, and cables, right down to such details as the types: dark or lit fiber, Ethernet 10G or Infiniband. Then, there's the number of network points involved, the location of venues, the exchanges selected, etc.

There are customized approaches. Hessler of Lime Brokerage breaks down latency analysis into six groupings-communications-related latency, feed handler latency, latency related to complex event processing performance, risk management-related latency, market center processing and reporting latency.

And almost all budgets are custom-built and custom-managed. More than one source interviewed for this story asserted that given the broad range of trading strategies and technology configurations employed by various traders, it is nigh impossible to assign average latency ranges to each and every component of a trade execution. Hessler of Lime Brokerage, however, estimates that for those trading firms whose trading strategies require low latency, most of the steps have latencies ranging from microseconds to multiple milliseconds. Alternatively, U.S. market center latencies range from hundreds of microseconds to multiple milliseconds.

A microsecond is a millionth of a second. A millisecond is a thousandth of a second.

TIME IS MONEY

Once the performance numbers are obtained, analysts say, the financial part of the latency budget exercise comes into play: Assessing the costs associated with any latency improvements.

"Latency budgeting is what you derive from what you invest in your technology-a kind of cost benefit analysis-but it can mean slightly different things to different people, depending on their trading strategies," asserts Adam Honore, senior analyst at Aite Group. "If you're trading in microseconds as many high frequency traders do, you'll be watching it far more closely than someone who is trading at a slower rate."

The trick is to make sure the money being spent to remove delays will be more than reimbursed by the profits achieved in the trading strategy that is improved by it, says Kevin McPartland, senior analyst at Tabb Group.

"If I'm spending \$1M to get one microsecond faster, will I make ten times what I was making before?," he asks. "Another way of putting it is to say that latency budgeting involves making sure, when you spend so much money over so many weeks on an upgrade effort, that you see multiple times the value spent in additional revenue."

According to Meinel of 29West, the trading firms he works with typically see budgets of 50 microseconds or less for processing messages. Alternatively, when it comes to moving messages within a server to core trading applications in a black box format the latency budget can come down to less than one microsecond.

This is in a co-location setting, utilizing multiple core processors in servers and incorporating the dedicated feed handlers, streamlined software for distributing messages and execution algorithms all at the same server. In doing so, this eliminates the need for the operations to make unnecessary network hops and can dramatically reduce latency. Meinel says such operations will often also entail the use of either Infiniband or 10G Ethernet network connections, for further speed.

Meinel also says that every time you add a new functionality to a component, such as an algorithm, there's always the possibility of adding latency.

This results in a trade-off that firms need to assess. "Is it worth it to add more intelligence to an algorithm if it means increasing the latency in your operations? And what are the cost factors that come into play?" Meinel says. "When assessing your budget, those are the questions you have to ask."

CHECK EVERYTHING

Managing a latency budget at a trading firm is akin to the way that Lance Armstrong, the champion cyclist, cares for his bike and professional racing career, according to Dariush Nazem, vice president and head of business development for low-latency solutions at Goldman Sachs.

He constantly checks and evaluates the performance of every component-the gears, the pedals and frame, etc.-to see what parts might need to be upgraded, replaced or refined, to ensure the highest speed and optimal performance.

"Clients should think about evaluating latency in the same way that Lance Armstrong dissects cycling. They need to look at every little piece or component within their business operations on a regular basis to see how you can decrease latency and maximize trading performance," Nazem said. "From eliminating a hop in the order path to rewriting code more efficiently, each component has to be assessed with the goal of improving its performance. Once a bottleneck is identified, it's important to fully understand which improvements can be made."

Feargal O'Sullivan, managing director of high performance messaging at the NYSE Euronext, says this involves looking at "the combined amount of time it takes for your trading applications to process an order." He notes that many clients will measure their end-to-end latency as well as a given exchange's and that the resulting number is what one should officially call one's "latency budget."

"They then might say, we're at 10 milliseconds, we need to get to 5 and will analyze every step of the way to see where they can reduce the latency," he says.

While the idea of managing a latency budget sounds rather straightforward, industry experts insist it is anything but. In fact, the process can be quite complex if it involves trading operations spanning stocks, bonds, futures and other assets. There can also be lots of measuring of latency inside a large sell-side firm where many trading operations exist involving multiple assets being traded in multiple parts of the world. In such environments, the vast quantity of variables make it ever more difficult to establish latency budgets and keep a close watch on them.

"Measuring latency becomes a very difficult task when you are looking inside the many trading components that exist inside a large sell side firm," according to Barry Thompson, founder and CTO of Tervela, providers of high performance and hardware-based messaging platform for traders. "There are so many variables-how the market data comes in, the market data feed itself, network issues, the performance of specific matching engines and algorithms and dealing with what is cached and what is not-that it becomes a true science project to adequately measure the latency of each and every component," Thompson said.

Carl Carrie, a consultant to trading firms and formerly a senior trading executive at JP Morgan, confirms that managing a latency budget inside a large trading organization versus a small buy side firm can be a major challenge. "I had multiple businesses and trading

operations to manage-including a high frequency desk-and so, where do you put your capital? Do you speed up the U.S. rather than operations in Korea? Do you apply your low latency efforts to market data, order processing, analytics or all three? There's a whole panoply of choices to consider," and no set system or formula in use.

Then there is the speed of change, which causes delay in improving delay-tracking systems. "You build a system from scratch, you are proud of it but then it becomes hard to ask, is this really state-of-the-art anymore?" points out Mark Mahowald, president and founder of 29West. "Some firms are being passed by in the latency game because they are not allowing their technology to evolve."

Examples might be falling behind in keeping their market data or server systems up-to-date or failing to utilize equipment that can scale up to the volume that their trading strategies might require.

There are no simple yardsticks. Execution speeds vary by trading venue, for example. "People tend to forget the fact that a firm will not trade on just one exchange but will connect to several and perhaps trade in different asset classes," explained Stewart Orrell, senior manager, financial services at global data center services provider, Equinix. His firm provides a "point of interest" mapping service to determine what paths to and from exchanges are the fastest.

FINE TUNING

Experts such as Carrie and Matt Bretan, senior technology consultant with Eze Castle Integration point out that there are latency testing products from firms such as Corvil, Correlix and Seanet Technologies that can assist in latency management, but these tools have their limitations. "While there is a great place for latency testing firms-they can help in managing latency from your data center to the exchange-but they are only one part of the latency management solution," insists Matt Bretan, senior technology consultant with Eze Castle Integration. He points out that when it comes to managing latency inside the enterprise, most firms will develop their own internal monitoring systems or trade simulations to keep latency measurement information close to home.

"It's not to the advantage of a firm to let someone come in, fine tune their latency and perhaps leave with intellectual capital; Everyone tends to conduct internal latency tests in their own way," says researcher Honore of Aite Group, noting that some build their own or mix their internal systems with various vendors' monitoring capabilities.

According to Nazem at Goldman Sachs, however, it's best when latency budgeting becomes a cross-functional, team effort. "At Goldman, everyone from senior executives to sales guys on downwards is an active stakeholder in the latency management process and is encouraged to offer their ideas regarding improvement or change in or trading processes," he said, including the firm's customers.