Inside Market Data

April 2011

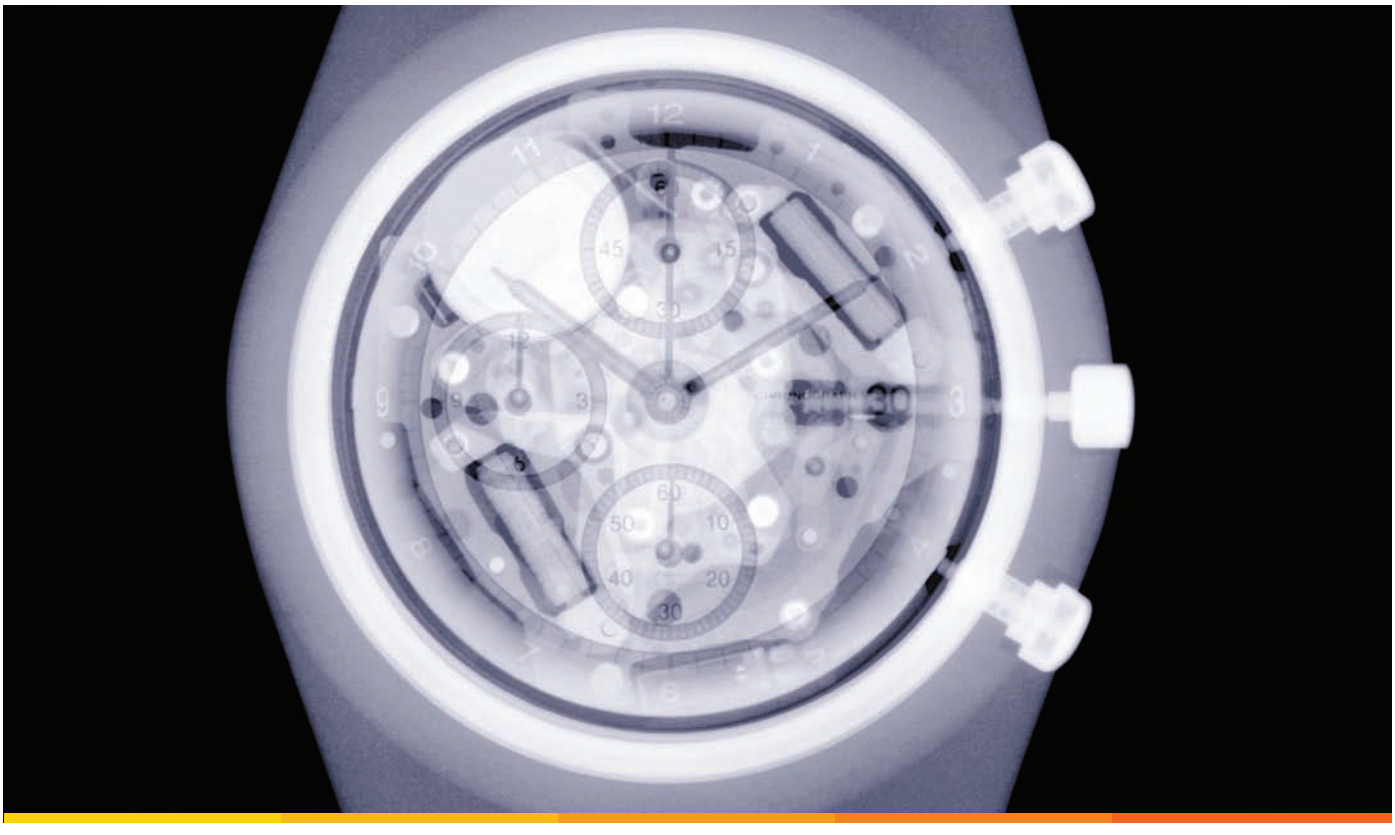
waterstechnology.com/imd

LATENCY

SPECIAL REPORT

Sponsored by:

citihub™ Corvil endace



Speed is good, transparency better

Get clear about latency management: Get CorvilClear™

SPEED – the holy grail of electronic trading. But searching endlessly for those elusive extra microseconds in the dark may not be providing you with the competitive advantage you need. In the complex maze of market data and order execution, a latency bump in the road could surface anytime, any place. Latency bumps can cost you bundles.

CorvilClear™ lets you see, and turn latency into a competitive advantage. For the first time ever, we're opening the doorway to inter-party latency transparency. Imagine a unique peer-to-peer technology for exchanging latency data with microsecond precision in a safe and scalable manner. Imagine minimizing risk and uncertainty without exposing proprietary information. Imagine precision latency transparency between trading partners, market centers, and market service providers — across all the parties at the core of your electronic trading operation.

Isn't it about time you were able to look beyond your own four walls? Now you can with CorvilClear™ from Corvil, the leading innovator in latency management technology.

For more information, please email us at: corvilclear@corvil.com.

Corvil

New York · London · Dublin **www.corvil.com**

LETTER FROM THE EDITOR

Speed Is Nothing Without Control

Since trading firms first realized there was an advantage to be gained by being able to capture market data and trade faster than their rivals, an arms race has been underway to see who could engineer the fastest and most efficient trading infrastructure with as little latency as possible.

First, firms turned to direct datafeeds, captured using low-latency ticker plants, encouraging exchanges to distribute their feeds even faster. Then firms implemented low-latency messaging layers to distribute that fast data throughout their organizations, which in turn placed pressure on vendors to streamline their technology and eliminate even the slightest trace of latency. This turned the industry's attention to hardware-based technologies that function faster than software, and eliminating network layers altogether by moving trading servers into proximity hosting datacenters close to exchanges, and ultimately into the same facilities as exchange matching engines, spawning a wave of mega-datecenters, such as those being built by NYSE Euronext, CME Group and the Singapore Exchange.

But as low latency becomes ever-more important, it also becomes harder to achieve. Saving milliseconds a few years ago was easier than saving microseconds or nanoseconds today, and firms are using sophisticated timestamping and measurement tools to monitor latency at the most granular level possible—in some cases, allowing them to generate microsecond improvements overall by eking out nanoseconds here and there. In addition, firms looking for end-to-end views of latency are also demanding transparency into the delays introduced by links in the chain outside of their control, such as trading venues themselves, prompting exchanges to roll out their own latency monitoring solutions, to provide members with statistics such as order-to-trade and trade-to-tape latency, and where a firm's overall roundtrip latency stands in relation to its peers.

Simply put, those who can't measure their performance can't perform, and will have to look elsewhere for trading advantage. But how long can it be before latency is exhausted as a competitive differentiator? Ultimately, this arms race is limited by light speed and the laws of physics. And perhaps in turn, those firms that can't compete on latency will have the upper hand when the rest of the industry gives up battling over picoseconds and femtoseconds. But according to the participants in this report, there's still a long way to go before we need to worry about that. ■

Max Bowie
Editor, *Inside Market Data*

Inside Market Data

Max Bowie, **Editor**
Tel: +1 212 457 7768
max.bowie@incisivemedia.com

Vicki Chan, **US Reporter**
Tel: +1 212 457 7758
vicki.chan@incisivemedia.com

Lee Hartt, **Publishing Director**
Tel: +44 (0)20 7484 9907
lee.hartt@incisivemedia.com

Jo Garvey, **Commercial Director**
Tel: +1 212 457 7745
jo.garvey@incisivemedia.com

Elina Patler, **Head of Editorial Operations**

Pedro Gastal, **Senior Marketing Manager**
pedro.gastal@incisivemedia.com

Lorna Graham, **Group Production Manager**
lorna.graham@incisivemedia.com

Incisive Media
120 Broadway, 5th Floor
New York, NY 10271
Tel: +1 212 457 9400
Fax: +1 646 417 7705

Incisive Media
32-34 Broadwick Street
London W1A 2HG
Tel: +44 (0)20 7316 9000
Fax: +44 (0)20 7316 9250
E-mail: customerservices@incisivemedia.com

Incisive Media
20th Floor, Admiralty Centre, Tower 2
18 Harcourt Road, Admiralty, Hong Kong
Tel: +852 3411 4888
Fax: +852 3411 4811

Subscription Sales
Gillian Harker Tel: +44 (0)20 7968 4618
Dominic Clifton Tel: +44 (0)20 7968 4634
waters.subscriptions@incisivemedia.com

Incisive Media Customer Services
c/o CDS Global, Tower House,
Sovereign Park, Market Harborough, LE16 9EF, UK
Tel: 0870 787 6822 (UK)
Tel: +44 (0) 1858 438 421 (overseas)
incisivehv@subscription.co.uk

NEWS ROUNDUP

CME Preps Aurora, Ill. Client Co-Location Datacenter

CME Group is finalizing the construction and client master agreements for a new co-location datacenter in Aurora, Ill., about 40 miles outside central Chicago, that will provide cabinet space for trading firms to co-locate their trading systems alongside the exchange's Globex trading engine and MDP market data platform.

The CME refitted a former grain warehouse as a datacenter to provide space for its high-density computer systems and trading engines—which are already running from the new facility—as well as a 428,000 square foot space for clients to co-locate their equipment, which will go live in early 2012.

"The concept is that bringing co-location facilities together within a facility where CME's trading engines are located allows the lowest latency connectivity for all products traded on CME," says CME Group chief operating officer Bryan Durkin. Previously, CME housed the trading engines at its datacenter in Lombard, Ill., and customers connected from proximity hosting space provided by third-party datacenter providers.

"We view this as an important offering in the overall enhancement of our business… and we made the strategic choice to build out our facility to accommodate demand," Durkin says. "As marketplaces become more automated… latency becomes a driver of demand across our user base of banks, proprietary trading firms, hedge funds and third-party providers," Durkin adds. These third-parties include network providers, who will also be able to co-locate network equipment in the co-lo center to provide low-latency connectivity to CME.

While the facility will "enhance performance and reduce latencies" in general, he says CME does not have specific metrics for latency within the co-lo center. ■

TMX Readies Micro-Message Binary Feed

Canadian exchange group TMX launched its new high-performance QuantumFeed datafeed on April 18, utilizing a new binary data protocol to reduce message sizes by a factor of 10, and reduce the required bandwidth capacity to one-third that of the exchange's existing feeds.

André Craig, vice president of the exchange's TMX Datalinx market data arm, says TMX will offer Level 1 trade and quote and Level 2 full depth of order book versions of the feed, and will recommend bandwidth requirements of 8 Megabits per second (Mbps) and 35 Mbps, respectively—both around one-third of the exchange's existing Level 1 and Level 2 feeds, which is a big issue for many of the exchange's clients.

The new feeds also address jitter—another big issue for clients, Craig says—by making latency and throughput more predictable, since the smaller binary messages are more efficient than TMX's existing feeds, which utilize the exchange's STAMP tag-based, variable-length ASCII protocol that involves more client-side processing and can create message sizes of between 600 and 1,000 bytes.

By converting the fixed-length messages and alpha-numeric codes to binary format—in conjunction with the removal of additional information contained in the existing feeds that is not directly relevant to book-building, and development work to optimize the exchange's messaging layer and data gateway—TMX has been able to reduce order messages from 600 bytes to 60 bytes, he adds.

TMX is targeting latency in the low hundreds of microseconds from message generation to the edge of its network initially, but is aiming for double-digit microseconds long-term. Craig says the exchange does not publish latency statistics, but that this will represent a significant improvement over the latency of its existing feeds. ■

Tibco Unveils Sub-Microsecond Messaging

Tibco has released a new ultra-low-latency messaging system that leverages new multi-core servers to meet the nanosecond-level latency requirements of hedge funds, high-frequency trading firms and algorithmic traders co-located in exchange datacenters, as well as execution venues themselves.

The vendor has tightly integrated the new platform, dubbed Tibco FTL, with its Rendezvous middleware product, so firms with current RV deployments in their less latency-sensitive middle and back-office areas will be able to standardize on Tibco messaging across their enterprise by rolling out Tibco FTL to support the low-latency requirements of their front-office trading operations as well, says Rourke McNamara, senior director of global product marketing at Tibco.

For communication within the same physical hardware using shared memory, Tibco FTL delivers intra-host latency of 384 nanoseconds, while transmitting data between two servers using RDMA (remote direct memory access) over InfiniBand yields latency of 3.1 microseconds, he says.

Tibco worked closely with Intel to optimize FTL for Intel's 64-core processors, including its Xeon 5600 and 7500 series—for example, to allow the processing load to be spread over all 64 cores—which included writing a large portion of Tibco FTL's code in the Assembly programming language, which uses Intel syntax.

As a result, the vendor has been able to reduce latency without stripping out functionality such as metadata support and content-based addressing, which enables users to route messages based on information contained within the message, as opposed to subject-based addressing which relies on a single data field, McNamara says. ■

Spread Shaves Ethernet Latency

Low-latency network provider Spread Networks recently completed the final stage of connectivity to datacenters in New York and New Jersey with the extension of its Ethernet Wave service to Equinix's datacenter at 165 Halsey Street in Newark, NJ, while reducing the latency of the service overall.

The reduced-latency Ethernet route provides roundtrip latency between the facility and the 350 East Cermak Road datacenter in Chicago has cut latency from 15.75 milliseconds to as low as 14.6 ms, with a service-level agreement of 14.75 ms, compared to 13.33 ms of latency for its dark fiber.

Brennan Carley, senior vice president of product marketing at Spread, says that since the change came into effect, "our customers running latency-sensitive strategies tell us that they are getting better quality fills than they were before the latency reduction." The move follows the recent expansion of Spread's Ethernet wave service to Savvis's datacenter at 300 Boulevard East in Weehawken, NJ, in addition to its other end-points in Secaucus and Carteret. ■

Kyte Expands ITRS Latency Monitoring Tool

UK futures commission merchant Kyte Group is expanding its deployment of ITRS' Feed Latency Monitor solution, to measure relative latency between datafeeds and ensure its trading applications receive the most timely data.

Kyte rolled out FLM last year to compare the latency of data from Trading Technologies' trading platform against data from an unnamed data vendor, and now plans to expand FLM to monitor data captured from Object Trading—an Australian provider of price discovery and market connectivity—to which ITRS is developing an API connection.

Before, Kyte captured latency information from log files provided by software vendors. But many ISVs only provide double-digit millisecond timestamps, whereas FLM provides microsecond granularity, which Kyte needs to support sub-millisecond trading and quantify the timeliness of data.

The system subscribes to different feeds via data vendors' APIs to measure relative latency and detect internal distribution issues or whether a datafeed has slowed only for specific desks, to narrow down the cause of any problem.

FLM compares feed latency from different sources, or multiple feeds from the same source, enabling users to compare data latency from market data vendors, ISVs and its direct connections to exchanges, as well as from multiple installations of the same platform. ■

Greenline Adds New Protocol Monitoring

MarketAxess' Greenline Financial Technologies unit is expanding the range of data protocols supported by its message monitoring solutions to enable it to monitor low-latency datafeeds.

Jean-Cedric Jollant, head of European operations at Greenline, says the vendor is needing to develop support for new proprietary exchange protocols for distribution of low-latency data, to be able to monitor message flow and latency.

Typically, message status data would be contained in log files for their network connection, but firms disable these to cut latency, Jollant says. "We implement MagniFIX on the client network between points A and B and 'eavesdrop' on the line without impacting the production flow—not for latency monitoring, but for monitoring message flows at low latencies," he says. The vendor already supports protocols from CME Group and the London Stock Exchange, and will roll out support for NYSE Euronext's UTP Direct and the latest version of Deutsche Börse's Values API by the end of Q2 to its MagniFIX solution for monitoring FIX messages, and its Latency Monitor tool.

After the initial exchanges, Greenline plans to add support for derivatives exchanges to MagniFIX, but will focus on equity exchanges in Europe and Asia for Latency Monitor, where more opportunity exists because low-latency protocols are just emerging. ■

SGX Taps Corvil for Reach, Co-Lo Latency

The Singapore Exchange is using Corvil's CorvilNet latency monitoring platform to measure data and trade message latency over its network and through its forthcoming new co-location datacenter and next-generation trading system, SGX Reach.

SGX engaged the vendor late last year, and Corvil has already installed servers running CorvilNet at SGX to monitor latency of the exchange's new platforms as it prepares to migrate to the new co-location center next month and to the new trading platform in August.

SGX has already installed the system into its production and test envronments in its current datacenter for monitoring the latency of its existing internal systems, and is in the middle of performance testing the new trading engine, and has also used it to monitor micro-bursts of data from SGX Reach to achieve sub-90 microsecond response times. The second phase of the rollout will involve deploying it in the new co-location facility to monitor and report on latency internally and to SGX's clients.

"This will provide our customers with transparent end-to-end latency information between their systems and the exchange's matching engine, allowing them to tune their systems and manage execution risk," says SGX chief technology officer Bob Caisley. He declines to identify SGX's former monitoring system, but says the previous tool did not provide sufficiently detailed transactional information for fine-tuning its infrastructure and trading engine, did not extend to monitoring latency between SGX and clients, and was limited to one-millisecond increments. ■

ROUNDTABLE

Low Latency's Theory of Relativity

As trading and data dissemination get faster and faster, latency has become a moving target, with firms constantly striving to achieve the lowest possible latency. Often, this isn't about achieving the fastest speeds physically possible, but merely being faster relative to one's peer group. But before you can fix the problem, you have to find it, and achieving this requires sophisticated tools to measure and analyze speed. In this roundtable, a group of latency experts share their views on the challenges currently faced by the industry, and those that lie ahead.

IMD: We've heard a lot in recent years about "the value of a millisecond." Since latency is now being measured in microseconds and nanoseconds, how should we be valuing latency's importance now and in future?

Tom Guinan, chief technology officer, Advantage Futures: Speed of execution and low-latency are extremely important to traders. The electronic financial markets become faster year after year, and electronic trading systems and strategies evolve to take advantage of faster access to market data. One of the beauties of trading is that the value of receiving market data and transmitting orders milliseconds or nanoseconds faster is evidenced by increases in traders' profits or lack of increases in traders' profits. The value of any improvement in technology can be determined by comparing the benefit to the trade from the new technology and increase in profits, to the cost of implementing the new technology.

Kevin Formby, vice president of business development, Endace: It's never really been about milliseconds, it's always been about being "faster than the other guys." When the others were all "slow," you just needed to be slightly faster, and latencies were measured in milliseconds. Now that many of the high-frequency-traders and prop traders are all very fast, the margin of who is fastest is measured in nanoseconds. The units of measurement have changed, but the basic benefit of being first has not changed—in fact, this has been a story going back to the advent of stock exchanges.

Donal Byrne, chief executive, Corvil: It is not the magnitude of absolute latency that is valuable; it is the magnitude of the relative latency—i.e. the difference in latency between the fastest and the rest of the chasing pack. To first order, if the magnitude of relative latency as a percentage of the total end-to-end latency remains the same, then it retains comparable levels of value. For example, a latency advantage of 1 millisecond over a total end-to-end trading loop latency of 10 milliseconds some years ago would be comparable to a 100 microsecond latency advantage today if the total end-to-end trading latency was reduced from 10 milliseconds to 1 millisecond. This property of relative latency advantage will continue, provided trading loop speeds continue to improve. The challenge today is that the difference in latency between the fastest and the rest of the chasing pack is shrinking on a relative basis. This reduces competitive advantage.

"Now that many of the high-frequency-traders and prop traders are all very fast, the margin of who is fastest is measured in nanoseconds. The units of measurement have changed, but the basic benefit of being first has not changed."

Kevin Formby, vice president of business development, Endace

Other factors also come into play. For example, the probability of success reduces as the number of competing participants increase. This is because latency is not a single number and can be better described using statistical distributions. Therefore, someone could be faster on average but can sometimes be slower. Understanding the specific latency distributions allows one to compute this risk quantitatively.

Ben Newton, associate partner, Citihub: We once heard it said, "I don't care about latency as long as I'm first," and we have found that the value of latency depends on a number of factors. These include how advanced you are towards your target latency, how that compares to your competitors, the assets being traded, and the geographic region you are in.

As ever, the Chicago and New York markets value latency most highly, closely followed by London. Interestingly, latency used to be a fairly level playing field in the Asian markets, but recent upgrades in the exchange space have acted as a catalyst for players to rapidly reduce latency.

Likewise, the asset classes most sensitive to latency are largely unchanged, with equities leading the pack. However, the leaders have started to realize that profits are possible in other asset classes, too—foreign exchange, for example, has a number of players executing natively in the single-digit microsecond space. One area that hasn't changed is in the advantage of agility, partly because proprietary trading shops often recognize the value of latency more than prime brokers.

History has shown that the latency evangelists were right—latency reduction can make trading strategies more profitable. Understanding the value of latency for each of us enables intelligent return-on-investment decisions. Our latency methodology has a framework that categorizes the business and IT requirements for latency against potential solutions. This ensures a cost-effective approach using the best-fit technology and process.

Kevin McPartland, senior analyst, Tabb Group: The unit of measure doesn't matter: it's about relative—not absolute—latency. You need to be faster than your competitor to capture the sought-after alpha. Furthermore, who exactly counts as your competition is not a function of trading in the same asset class or the same region, but of those trading a similar strategy, such as index arbitrage or latency arbitrage.

IMD: Is latency more or less important now than a year ago? If more, what's behind that change? And if less, what other issues are rising to the fore?

Newton: Latency has never been so important, but we're now much more focused on business benefit and competitive positioning. The competitive landscape has changed radically in some markets, forcing clients to better understand their comparative positioning, leading to much more industry analysis. Providing the ability to target investment in specific areas of the maturity curve translates into maintaining position or advancing to become best in class.

Just buying the latest technology isn't going to solve all your latency problems. The right cross-silo organization, process and controls are needed to manage and maintain these solutions. Of course, the scope of any platform work won't just be about latency—it will also about predictability, determinism, capacity, agility and cost, among other factors.

Byrne: Our experience strongly suggests that latency is more important today than a year ago. The reason is that the need for low-latency has penetrated the capital markets more broadly and has moved from a "must-have for the elite traders" to a "must-have for *all* traders." In a few years' time, we will not be discussing terms like "high-frequency trading" or "low-latency trading," because all trading will be "low latency." We are witnessing the re-birth of capital markets infrastructure and trading systems, where the lowest common denominator for trading speed is going to be set at a new threshold of operational acceptance.

While it is increasingly difficult and more expensive today to get a significant speed advantage over your competitor, it is very easy to find yourself slower and at a significant disadvantage. This characteristic of the market will tend to move all participants to a new higher-speed norm where the latency spread between the fastest and slowest will shrink. This is similar to what happened with the Internet, where normal access speed today is multiple megabits-per-second compared to dial-up connections with kilobit-per-second speeds not too long ago.

The idea of accessing the Internet today at such low speeds would seem ludicrous.

ROUNDTABLE

McPartland: It's no more important, but it's certainly much more sophisticated than it was a year ago. Ultra-low latency is now measured in tens of microseconds rather than single-digit milliseconds, and to achieve such low levels of latency, your end-to-end infrastructure must be perfect. When you're counting billionths of a second, there's not much room for error.

Tom Guinan
Advantage Futures

Guinan: The importance of latency to trading continues to depend on what strategies and algorithms traders are employing. For the most time-sensitive strategies, latency remains extremely important. New technology has provided traders with faster connections to exchanges. To stay competitive in this space, traders concentrating on high-frequency, low-latency trades must stay current on each component of their trading system and network infrastructure to ensure they adopt any technology that will provide a net benefit. The focus on latency will continue and perhaps intensify for this segment of the trading community.

Formby: Again, the benefits of being first are the same, but it's increasingly the cost of getting in front that is becoming higher. To shave a few microseconds off the entire trade latency is becoming a big-ticket item. Furthermore, once an investment is made, it may be only months, weeks or even days before competitors catch up, the arms race reaches another stage, and the investment benefit is wiped out. Clearly, two years ago it was easy to identify sources of latency—if you were 50km away from a datacenter, then this was clearly a source of latency. Today, with co-location centers, 10 Gigabit-per-second (Gbps) or even 100 Gbps networks, on-chip Application-Specific Integrated Circuit (ASIC) and Field-Programmable Gate Array (FPGA)-based devices and software and optimized code, it's a lot more difficult to understand the sources of latency—and that is why there is considerable investment going on in infrastructure to monitor latency throughout the entire trade latency chain. Every single nanosecond of latency has to be measured and the cost/benefit of eliminating the delay understood.

IMD: Where do the biggest opportunities still exist to reduce latency in the data distribution and trading process?

Formby: If we could only break the speed of light we would be fine! Seriously, at the moment there is likely to only be incremental adjustments in overall latency, but increasingly there will be more emphasis on the cost side of the equation. People say, "I know a project will reduce latency by 10 microseconds, but if it costs $10 million, it isn't practical. However, if I can do it for $100,000, then it becomes practical." So look out for cheaper ways to do the same thing. Sharing infrastructure is one obvious way to reduce costs. For example, don't deploy infrastructure to measure latency alone, use it to detect cyber-intruders in your trading fabric as well—it's all the same data.

Mark Dodds, associate partner, Citihub: In our experience, we're still a long way from the end-game, and we have yet to see an environment that couldn't be significantly improved. We categorize opportunities into four areas—venue and connectivity choices, application and architecture, network and middleware, and hardware and operating system. All of these are underpinned by effective monitoring, and all areas are still yielding results—in many cases without additional capital expenditure. Venue and connectivity choices present an opportunity for massive amounts of optimization, depending on your market. The idea of getting the data to you faster or moving closer to the source isn't new, but detailed analysis of an exchange's connectivity options often yields interesting results. Subjects like connection, location and protocol optimization can all deliver significant enhancements.

Guinan: Some exchanges are offering—or have announced plans to offer—co-location in new datacenters that host the exchange matching engines. This eliminates any latency due to geography and demonstrates the exchanges' commitment to controlling access to market data for participants seeking the fastest available access to market data. These proximity solutions are a key factor in reducing latency as they eliminate all geographical constraints.

> **"Most people have moved to multi-venue co-location trading and are running arbitrage strategies involving multiple instruments at each venue. Latency between these sites can be in the multiple milliseconds range, so shaving microseconds off the total latency can have a dramatic effect."**
>
> Donal Byrne, CEO, Corvil

Byrne: Reducing latency in the trading loop is a moving target. One has to first detect and characterize accurately where the current latency bottleneck lies. This could be in the trading engine, the matching engine, the network, or a myriad of other potential sources. Once the bottleneck is identified and fixed (improved), then some other part of the trading loop becomes the new bottleneck. Thus the process begins again, and with further iterations the overall latency is reduced. In the past, a lot of focus moved to reducing the speed of the trading engine and the matching engine. With technologies like FPGAs, the response times of algo trading engines have been reduced below 10 microseconds, and some DMA client gateway engines have been reported in the range of 1 to 2 microseconds. New-generation matching engines have reported achieving latencies in the range of 10 to 20 microseconds. As a result, attention has now started to shift back to the network where the propagation delay is roughly 1 microsecond

Only an Endace Monitoring and Recording Fabric gives you the power to see *everything* on your network.

Guaranteed 100% packet capture and scalable to 100Gbps, Endace provides the most accurate, open and predictable platform for running your choice of network security and monitoring applications.

power to see all

endace.com

endace

per 200 meters of fiber. However the biggest latency contribution in the network comes from active elements like firewalls, routers and switches. New-generation switches are being delivered to reduce transit latencies below 1 microsecond.

The biggest focus today on latency reduction seems to lie with inter-co-location network latency—i.e. the latency between co-location centers. Most people have moved to multi-venue co-location trading and are running arbitrage strategies involving multiple instruments at each venue. Latency between these sites can be in the multiple milliseconds range, so shaving microseconds off the total latency can have a dramatic effect. This is the reason why specialty telecom providers are digging new trenches and laying new fiber just to achieve such latency reductions along major network routes, such as between New York and Chicago.

Kevin McPartland
Tabb Group

McPartland: Latency reduction goes in cycles. For a while, we spend all of our time on upgrading hardware and improving the network, then when little additional savings can be found there, we move on to optimizing software, and then back to hardware, and so on. We seem to currently be in the software phase. The last year or two saw networks flattened, fiber routes shortened and the implementation of inter-process communication, so now the quants are working to make their trading algorithms more efficient by cutting lines of code and making calculations more quickly.

IMD: What are the main challenges to reducing latency, and are these becoming even harder to address as the industry picks off the low-hanging fruit from the latency chain? With such small gains on offer, is low-latency still worth the amounts spend on it, or is it more productive to invest in other areas?

McPartland: Becoming harder is a huge understatement. It was much easier to cut a millisecond in 2005 than it is to cut a microsecond in 2011. For so long, milliseconds were a mere rounding error when it came to measuring execution latency, so when it was decided that trading platforms needed to be more efficient, there were some obvious steps that could be taken. But after years of taking these obvious steps, the remaining steps are no longer so obvious and require the best brains (and hardware) in the business to carry out.

Formby: There are two main challenges: First, it's a zero-sum game. If you want to play in this space, you need big bucks. If you don't have it, then don't play. There are no prizes for second place. Second, the costs for each incremental improvement are going up. Both mean we will see consolidation of low-latency traders or their withdrawal from the market. Many of those who decide not to play will put their intellectual capital into more sophisticated trading strategies than just timing arbitrage. Data mining technology, pattern recognition, complex event processing and anomaly detection will increasingly be the new buzzwords in this space as traders replace raw dollar spend with intellectual capital. All of these techniques will require more comprehensive and time-synched datasets and greater understanding of the relationships between market participants and their trading strategies. Look out for more applications that reverse-engineer competitors' trading strategies, and the development of countermeasures to combat these.

citihub™

Ben Newton
Associate Partner (above)

Mark Dodds
Associate Partner (below)

Citihub
Tel: +44 0800 028 1901
www.citihub.com

Dodds: Organizations face different challenges. For some, the challenge is in introducing new technology platforms and solutions into established teams, while others strive for highest levels of production quality. More nimble players want the latest technology as fast as possible. We've got toolkits to help each of them reach their goals quickly, but the key is to take a systematic approach and prioritize all possible avenues of improvement. The demand for latency can be grouped into tiers, from seconds down to nanoseconds. Goals will be different for each tier, and different depending on how far down the latency path a firm is already, and how far it needs to get. People can sometimes be off put by small gains, but really it's not about that—it's about your competitive position. The mantra is to target just enough investment to sustain the advantage in your tier.

Guinan: All the low-hanging fruit from the latency chain has been removed. The reductions now are in smaller and smaller time increments, and the discussion has moved from milliseconds to microseconds and nanoseconds. For high-frequency, low-latency traders, whose algorithms are speed-dependant, the process of evaluating the cost/benefit of new technology and co-location offerings will continue. We expect traders to devote more resources to developing algorithms and trading new products as the incremental return on technology decreases.

Byrne: The main challenge we see is latency measurement and transparency. Albert Einstein has this wonderful quote about measurement—"Not everything that can be counted counts, and not everything that counts can be counted."

In the world of high-frequency trading, it is critical to have visibility of latency throughout the complete trading loop, including every element in the path. Unfortunately, this is not always possible, as many exchanges do not provide this level of latency transparency to members. In addition, it is often impossible to measure the latency for a member's trading session relative to all other members. This is something uniquely offered by Deutsche Börse, where a member has access to an anonymous speed ranking table and can determine their position relative to all other competing members.

Latency measurement and visibility is also critical to determining the likely return on investment from a reduction in latency. It might be the case that a microsecond reduction in latency at the critical bottleneck point could be worth a major investment. Likewise, if the investment is made in the wrong part of the trading loop, it might be a complete waste of money. Having accurate latency measurement and good latency analytics will de-risk this decision and maximize return on investment.

IMD: What do you predict will be the next big step in reducing latency, and how much of a difference will it achieve? Or will the next big thing be something that delivers an advantage not related to latency?

Guinan: Technology is improving at an incredibly fast rate, which challenges the most time-sensitive traders to continually assess the impact of faster devices constantly being introduced. There is not a lot of latency left to reduce, so changes will probably be incremental as processing devices and network equipment improve. As mentioned before, as improvements in technology offer smaller benefits, traders are likely to allocate resources to new strategies and products.

endace

Kevin Formby
Vice President
Business Development
Endace
Tel: +1 770 362 2226
Kevin.formby@endace.com
www.endace.com

Formby: The next big step will be exchanges and multilateral trading facilities offering the ability for traders to run their applications within a virtual machine located on the same physical hardware as the matching engine. There are still lots of issues around security and "fairness," but this is the next big step. Looking three years ahead, I expect end-to-end trade latencies to drop to sub-microsecond levels.

Newton: Apart from quantum computing and entanglement, a true revolution seems unlikely in the near future. Evolution of existing tools continues a drive ever-closer to the zero target. 2011 should deliver some exciting developments, such as hybrid systems with GP-GPUs (general-purpose computing on graphics processing units) and offloading core functionality to FPGAs (field-programmable gate arrays) both poised for mass deployment. The wider adoption of FPGAs throughout the platform is imminent, and hardware-only will soon be a choice.

The latest latency-reducing technology alone can't buy success. It needs to be supported by monitoring and control, as well as the right people and processes. We get excited by implementing hardware monitoring as well as PTPv2 (precision time protocol) and the nanosecond-level precision that it drives, but we're equally motivated to introduce the concept of a cross-silo "chief latency officer" to ensure this topic gains the same focus as availability, capacity and performance management.

Beyond pure latency, gains will be made through powerful analytical tools correlating events within the platform at the highest level of granularity, to sustain performance and drive business objectives.

Byrne: We predict that the next big step will be the provision of information and analytics about latency as opposed to absolute latency reduction. The relationship between relative latency reduction and cost is typically exponential—i.e., to halve your current latency typically requires spending more than twice as much. Of course, if you were prepared to wait long enough, you could achieve a major latency reduction in 12 to 24 months for the same price as today. However, this is not a reduction in relative latency, as everyone would be able to access the same faster technology at the same time.

We believe that in the coming years, investment in high-quality information about latency of your trading environment will ultimately deliver a higher return than investing in absolute reductions in latency. As trading environments become more complex, involving multiple venues and asset classes, the trader with the best quality latency information will be able to detect and capture more trading opportunities for a lower total investment compared to someone who does not have this information.

Think of latency information as a measure of the risk that we will not receive the price advertized on the market data feed. By leveraging latency information in this manner, one can make a lower-risk decision on a trade that attempts to hit a specific price.

McPartland: I foresee that eventually [glass and fiber-optic manufacturer] Corning will figure out how to make light travel faster through fiber, as it does not yet come even close to the speed of light in a vacuum. I think Moore's Law also still comes into play here, as the faster the processor, the more operations it can perform per microsecond. These are not small problems. These are the problems that get solved in labs in Palo Alto and Burlington, VT and take years—if not decades—to solve. ■

SPONSOR'S STATEMENT

Extreme Measures: Achieving Nanosecond Visibility

Reducing latency across firms' infrastructures remains a priority. But achieving latency gains—and more importantly, understanding them in the context of the market as a whole—requires technologies that can measure latency at a more granular level than the speed of the systems they monitor. Donal Byrne, chief executive of Corvil, outlines architectures that allow systems to achieve this performance


Donal Byrne
Corvil

As high-performance trading drives infrastructure to deliver ever-lower latencies, organizations need visibility into that infrastructure at correspondingly shorter timescales. This visibility must be based on timestamps with a precision of at least an order of magnitude or two finer than the latencies being measured.

Much of the core functionality of these systems is implemented in software on general-purpose CPUs from Intel or AMD. Software latency is lower than ever because of hardware improvements but also because of improvements in how software uses the hardware, such as through the use of cache-friendly algorithms that achieve lower latency through lower cache-miss rates, and lock-free structures that minimize contention between different threads in multi-threaded designs.

Applications cohabit with any number of other applications and processes, but the most effective way of ensuring application performance is to keep it as isolated as possible from the operating system and other processes. There are two principal approaches to lowering the latency of network I/O (input/output) through the operating system—kernel bypass, which moves the network stack out of the kernel and into user-space, and stack bypass, which avoids the IP stack entirely, using a much more direct mechanism to transfer data from host to host. This almost always means some form of Remote Direct Memory Access (RDMA).

Field-programmable gate array (FPGA) instruction sets are implemented in silicon but are programmable and can be rewired on demand. This is a huge advantage for trading algorithms that need to be regularly updated. FPGAs typically have a very deterministic latency profile—bandwidth and latency of the processor is known and doesn't change under load or other conditions.

The main causes of network latency are propagation delay, caused by the finite speed of signals in fiber or electrical cables; serialization delay, caused by large data packets being written one bit at a time; and queuing delay, caused by aggregation, which is necessary to make networking scalable and economically viable.

The simplest way to tackle network latency is to increase bandwidth. This reduces serialization delay, the likelihood of congestion, and queuing delay.

Interestingly, the simplest component of latency—propagation delay—has received the most attention of late. By taking space in an exchange co-location facility or proximity hosting site, traders can eliminate nearly all of the propagation delay between their algorithms and matching engines.

InfiniBand, a high-bandwidth, low-latency host-interconnect technology, is most often used in high-performance computing clusters. InfiniBand fabrics provide deterministic latency guarantees, and its architecture allows for very low-latency messaging. However, it also pushes extra complexity into the end systems, resulting in a network interface that does not map cleanly to the standard socket model.

The low-levels of latency that can be achieved by the technologies discussed here are impressive, but there are important considerations to take into account.

It is important to define exactly what is being compared and how is it measured. For example, one might describe the time taken to send a message across the network as "four microseconds" or "under two microseconds," depending on how it is measured: the latter figure is the latency to get a message from the sending application out onto the wire, but to get that message into memory for use takes the same length of time again on the receiving host, as well as the network switching time.

Another important consideration is that the lowest latencies are achieved under laboratory conditions. Production systems may achieve close to these latencies, but it is certainly not a given. There are sometimes trade-offs to be made between latency and throughput.

Most attention is paid to minimum or average latencies, but maximum latency or the high percentiles are usually the most important. For example, microbursts in trading networks can drive significant queuing in the buffers that protect aggregation links against packet loss.

Low-latency trading has driven the adoption and development of a set of technologies that enable trade execution and the handling and delivery of market data at microsecond timescales, which means that precision latency management must be capable of delivering measurements accurate to hundreds or tens of nanoseconds.

At the same time, total system latencies can vary by orders of magnitude because of the effects of dynamic congestion. Effective latency management requires capturing the complete distribution of latency, and analysis of the causes of latency. It is not sufficient to just measure spikes in trading system latency: you must also capture microbursts and other infrastructure behavior that drives latency spikes. ■

Inside Market Data | Inside Reference Data | Buy-Side Technology | Sell-Side Technology

waterstechnology

Premium pickings

Do you want to cherry-pick the most valuable content and tools for your company in minimal time?

Our brands have long been the choice for professionals within financial-market technologies. Our premium content from **Waters**, **Inside Market Data**, **Inside Reference Data**, **Buy-Side Technology** and **Sell-Side Technology** is now available to you via a multi-channel business intelligence platform with daily analysis and news to enable businesses to deliver better strategies with more efficiency than before.

By integrating our five market-leading brands you will get:

- the industry's most respected editorial teams under one web interface
- up-to-the-minute analysis and news – uploaded directly from our editorial desks
- easy navigation – articles linked by topic across all five brands
- free content from our extensive selection of special reports and webinars
- comprehensive news alerts delivering premium business content
- consolidated events and training calendar
- one community to network and interact with
- a multi-level premium subscription service to serve your company's needs – individually, departmentally or globally

waterstechnology.com
Premium content for financial-market professionals
Trial today – visit waterstechnology.com/trial

For details about individual access and corporate site licences to the entire WatersTechnology platform, contact waters.subscriptions@incisivemedia.com

An incisivemedia website
incisivemedia.com


An
incisivemedia
publication