A-TEAMINSIGHT presents

Electronic Trading

High Performance Technologies for Trading

Interactive Data

BLADE NETWORK TECHNOLOGIES

intel

www.a-teamgroup.com

A-TEAMGROUP
Business Intelligence for Financial Markets IT

Co-location, Performance Management and Technology Innovation for Ultra-low Latency Trading

By Scott Caudell, senior vice president global architecture, Interactive Data 7ticks

Over the past decade, electronic trading technology has undergone a major transformation. The speed of trading has moved from being measured in seconds in the pit to milliseconds on the screen and now to microseconds with the advent of co-location and automated trading. Ten years ago the word co-location was hardly used and now this and other terms, such as naked access or low latency, are almost commonplace. Today, exchange co-location is a given in the high-speed trading community, and the search is on for new ways to maintain the infrastructure edge.

> "This new landscape has created significant interest in cross-asset trading that links market centers in Chicago, New York, London, Frankfurt and Tokyo."

One way to maintain the infrastructure edge is to focus on network and systems latency. While this can be rather complicated, numerous improvements can be made with little or no change to actual trading applications. For instance, as newer generations of switches and their associated chipsets are released, reductions in latencies, in some cases from over 30 microseconds to now sub-microsecond times, are being seen. On the systems side, similar vendor and hardware advances have produced numerous kernel bypass and other Field Programmable Gate Array (FPGA) cards. These cards can easily produce improvements of over 100 microseconds within individual servers or even run trading programs themselves.

With designs on building even lower latency infrastructures, an influx of technology vendors has focused their resources on the financial space. Telecommunication carriers continue to build new fiber paths between major financial centers that are specifically tailored to financial customers. This new landscape has created significant interest in cross-asset trading that links market centers in Chicago, New York, London, Frankfurt and Tokyo.

For this type of trading, knowing the fastest route between two markets can be just as important as having co-located servers at either end. Several years ago when IntercontinentalExchange first began offering WTI crude oil futures in competition with the New York Mercantile Exchange, a number of high velocity trading firms invested in high-speed connectivity to the ICE match engine in Atlanta in order to arbitrage between the two markets. Because of the way that ICE managed its technology, firms in Chicago had to route market data and orders through Indiana, Ohio, Pennsylvania, Virginia, North Carolina and South Carolina prior to getting to the ICE match engine located in Atlanta. For a period of time those trading firms savvy enough to use a lesser-known path that went through Indiana, Kentucky and Tennessee could save 13 milliseconds on transit time. This was rendered moot when ICE chose to move its matching engine to Chicago, but while it lasted it was a sizable advantage for firms using the lesser-known path.

Many of the major exchanges—CME, Eurex, ICE, NYSE Euronext, NASDAQ, BATS and LSE among others—now offer high-speed access via co-location services. The Tokyo Stock Exchange recently launched the new Arrownet service, which allows for co-location with its matching engines, and a growing list of other markets plan to follow suit. The expansion of co-location services has opened up a number of interesting arbitrage opportunities between these markets and has led many firms to begin exploring multi-asset trading opportunities.

Network Latency

While there are numerous firms that successfully leverage existing co-location solutions and proximity access, the challenge today remains in successfully connecting physically diverse trading locations. The latency encountered between two or more co-location facilities can be the largest factor in determining the success of cross-market trading.

Direct fiber paths between two proximity sites are rarely available. Rather than simply drawing a straight line from A to B and then laying fiber along that path, a telecommunications carrier first must find right-of-way passage, which means utilizing railway or expressway routes to run fiber. A carrier also considers how best to maximize population and communication density along the line. Consequently, a connection from A to B can end up taking a very indirect path via the available transportation routes to hit as many cities and towns as possible.

The route between New York City and Chicago is a typical example, as until recently there was no direct fiber path between these two cities. With the continued focus on low latency infrastructure, numerous carriers, as well as dark fiber providers, have either pieced together existing fiber assets in more efficient ways or even trenched new fiber via more direct routes.

Distance traveled is not the only factor in determining the latency of any particular path. Some of the cables are older and are not capable of running newer and faster optical equipment cost effectively. The actual fiber itself can also be impacted by varying levels of attenuation depending on manufacturing standards, which means optical signals may need more amplification and regeneration on some routes than others.

As a result, there are significant variations in latency. On the New York to Chicago axis round-trip latency can range from upwards of 30 milliseconds to well below 17 milliseconds. On the New York to London axis latency can range from over 100 milliseconds to below 70 milliseconds. In a world where trading firms are spending hundreds of thousands of dollars to shave just a few microseconds off their transit time these variations are huge.

Although high-velocity trading is concentrated in the major financial centers where the largest liquidity pools reside, there is a growing amount of interest in emerging markets. This creates additional technological challenges. Finding the fastest path to these markets means grappling with vast differences in telecommunication standards and quality, and in some cases it requires dealing with a veritable rat's nest of submarine cable systems at varying degrees of age, performance and stability.

It is also important to distinguish between relatively closed co-location environments and neutral or open access environments. For instance, NYSE Euronext built and controls its own facilities, tightly restricting who can co-locate, space availability, and even how access is gained, while Eurex relies on Colt and Equinix, providing a much more open environment. Using third-party providers typically gives trading firms more control over the type of connections used and the equipment in the facility, and in some cases greater freedom in connecting to two exchanges through the same facility.

The CME Group, which is building its own facility, has recently announced it will take a different approach by allowing telecom providers open and fair access to its facility. This new facility has the added benefit of providing much greater power density than most commercial facilities but still provides a certain amount of choice to customers. The goal is to provide the lowest latency access to the exchange.

Managing Capacity

Another key factor in managing financial networks today is the importance of capacity. In many cases, high-velocity trading strategies are at their most profitable when the markets are the most volatile. But these periods are also the most difficult to manage in terms of capacity. Spikes of data at sub-second or even sub-millisecond intervals (known as micro-bursts) can overwhelm a system's capacity to process market data and add latency to the trading process.

This can be an infrastructure-wide management issue and tends to be ignored given the amount of time and effort usually required to analyze, manage and adapt to these events.

Some of the more advanced firms today are beginning to understand these factors and the large role they can play in performance management, but there is still a big gap between companies that can effectively integrate these tasks into their overall approach and those that are still in the dark. This becomes all the more important when a trading firm is building a system that requires high-speed processing of market data from several market centers that are physically remote from each other.

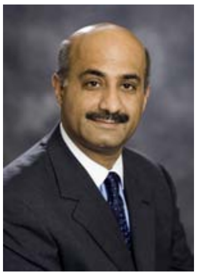
In response to these challenges, a host of companies have emerged to provide services to high-velocity trading firms. Some high-frequency trading firms choose to build their entire IT infrastructure themselves, but as this sector of the securities industry has grown, more and more companies have stepped in to supply infrastructure at lower costs and better time to market. This can have major implications as firms move into new markets and implement new strategies. In an industry of first mover advantages, it pays to get to the market more quickly than the next firm.

As high-velocity trading expands globally, only a limited number of firms will have the resources to maintain long term relationships with multiple co-location facilities around the world and build out international long-haul circuits to interconnect these facilities into one high-speed network. This is where managed service providers can deliver value, by interconnecting all of these facilities while enhancing service and reducing overall costs.

Scott Caudell is a co-founder and Sr. VP Global Architecture at Interactive Data 7ticks, a managed services firm that provides IT infrastructure services to trading firms. The firm designs, builds and manages networks for sophisticated trading firms, institutions and hedge funds.

Roundtable: High Performance Technologies

By Scott Caudell, senior vice president global architecture, Interactive Data 7ticks



Vikram Mehta, president and CEO, BLADE Network Technologies

Christian Sommer, director, Financial Services Marketing, Intel Americas Marketing Group

How much do business heads at Wall Street firms care about underlying technologies and technology choices?

Caudell: I believe business heads aren't concerned so much with the underlying technologies but with the business functions those technologies are enabling. They care about cost and performance, but not the technology choice itself. So, for instance, they might say "we'll build an ultra-low latency trading environment," but then the technology choices follow from in-house or paid expertise on how to deliver that objective.

Mehta: Wall Street firms realize that every element of their IT infrastructure plays a critical role in ensuring overall performance, particularly for the machine-to-machine interactions that characterize high-frequency trading. The cost of providing power, cooling and data center space are critical considerations when planning a next-generation financial services data center. Seemingly small differences in equipment footprint, power consumption and cooling efficiency quickly become major differences when multiplied by several hundred of several thousand switches, routers and servers.

Sommer: As we continue to emerge from the 2009 economic crisis and seeing IT budgets slowly increase, technology in 2010 will continue to play a critical role for financial services industry. IT will help banks manage risk, simplify infrastructure and access markets faster than ever before. 2010 will be a very important year with recovery, cost reduction and continued growth remaining front and center. New regulations and compliance rules are also coming into play.

How do you push performance with your application while also ensuring reliability and uptime?

Caudell: For me, proper instrumentation is a key factor. You can't improve upon what you can't measure. Not only do you need proper QA and testing environments to do rapid development and roll-out, but you also need objective and accurate methodologies in place to track and quantify how specific changes impacted an environment.

Mehta: Top performance combined with non-stop reliability and always-on uptime can be achieved with careful consideration of network design issues. For example, a low oversubscription ratio, or blocking ratio, is a key requirement for maintaining application performance. The oversubscription ratio is the amount of uplink bandwidth divided by the amount of server bandwidth. Acceptable oversubscription ratios vary depend on the application and the number of hops in the networks. To prevent application bottlenecks, a low oversubscription ratio can increase application performance and decrease latency for market data applications. Higher oversubscription ratios lead to network congestion and drive up latency. The financial services data center, a non-blocking, zero oversubscription ratio is the ideal, and can be designed into the network. However, the networking gear should also provide low latency and line-rate throughput in order for the non-blocking architecture to make a difference in application performance.

Does high performance mean proprietary architectures and code, or can open standards co-exist with high

performance applications?

Caudell: "High performance" is a broad term. Does it mean "as fast as or faster than the other guy?" Or "high throughput" or "low latency?" In my opinion, there is no single formula for achieving it--it can be realized in any number of ways--with proprietary architectures, open standards or through a third-party vendor.

Mehta: Open standards and high performance applications do go together. A case in point is Ethernet, which is based on universally accepted standards and allows for broad interoperability across applications, co-location facilities, data centers and networking equipment. With the emergence of 10 Gigabit Ethernet, high performance applications can benefit from Ethernet economics while delivering the fast throughput formerly associated with InfiniBand and proprietary network fabrics.

Sommer: Intel supports open standards by providing technology that allows for greater innovations to support the need for performance, cost efficiency, volatility and agility.

Is the lowest latency still the biggest "performance" issue in trading? What else matters?

Caudell: I feel performance issues are interrelated. Low latency is a big component, as are scale and consistency. If your system is not consistently delivering low latency, or if there's not enough capacity during market peaks, then you're severely limiting how successful you can be.

Mehta: Latency remains critical. Low network latency can be achieved with a combination of the right software and hardware, along with a network architecture that minimizes the number of tiers, or hops, for data to traverse. A low latency network should also support technologies such as Remote Direct Memory Access (RDMA) and TCP offload can help reduce server processing overhead. Cut-through switching and a latency-optimized data path can ensure that the switches' latency contribution to overall latency is negligible. Switching and routing equipment should also demonstrate line-rate throughput. Also important are support for multicast traffic flows so that traffic streams from multiple exchanges can be efficiently deliver to many users without overloading the network and dynamic multi-path routing to lower overall network latency.

Sommer: Intel perspective: Intel understands the demand that financial markets have for performance and low latency technologies. We are committed to designing and delivering products and platforms that provide industry leading performance and mission critical features to support the continuous performance demand. The Intel Xeon 7500 Processor for expandable server segment and Intel Xeon 5600 Processor series for 2-socket servers deliver breakthrough performance.

How have you leveraged recent technology advances in your applications?

Caudell: Interactive Data 7ticks has leveraged two recent technology advances. There have been major improvements in kernel bypass technologies--we've taken existing client systems and reduced latency by simply plugging in a new network card. We're also taking advantage of network hardware vendors utilizing the latest evolution of ASIC design. The improvements in these ASICS adopted by network hardware vendors have aided us in the creation of significant reductions in latency within our managed service environments.

Mehta: BLADE Network Technologies is at the forefront of next-generation network technologies. For example, stock exchanges are looking to reduce latency by reducing the number of tiers in the network. While traditional networks are based on three tiers – at the access, aggregation and core layers – some new deployments are removing the aggregation layer and connecting access layer switches directly to a 10 Gigabit Ethernet switch core. There are several proposals for how to achieve a large flat Layer 2 network. BLADE is a proponent of the Transparent Interconnection of Lots of Links or "TRILL" charter that proposes a hybrid Layer 2 routing protocol that would enable all links in a Layer 2 domain to be used, effectively increasing available bandwidth.

Where do you think the next big advance with underlying technology will occur – and why is it important to your applications?

Caudell: I think the biggest advance will be further adoption of existing technologies. I don't think we'll see a lot of big technology advances. Looking at the 80/20 rule, we're out of the 80 and into the 20. These minor improvements are important and add up, but you won't get this one thing that's a silver bullet unless you're significantly behind the curve in technology adoption.

Mehta: Ethernet will evolve to 40/100 Gigabit speeds. BLADE expects to see 40 Gigabit Ethernet in 2010 and 100 Gigabit Ethernet soon after. These high-speed networks will enable fabric-based computing that aggregates thousands of servers, storage and network devices in the never-ending quest for "zero latency."

Sommer: Intel continues to take a holistic platform strategy to address the demands for capital markets across trading and risk management. After working closely with our customers to understand their most critical needs and working to design solutions that go well beyond simply building faster processors for servers we went to the drawing board and introduced "Nehalem EP" in 2009 with Intel Xeon 5500 Processor Series. It started with the economic benefit Xeon 5500 delivered to IT end-users for those looking to refresh their old servers. This trend will continue to improve in the EP and EX server segments. Intel recently introduced the Intel Xeon 5600 Processor Series (codenamed "Westmere") and Intel Xeon 7500 Process Series (Nehalem EX) delivering the biggest performance leap in the history of Xeon product line with up to 3x increase vs. Previous generation and a 20x increase vs. Single core servers just a few years ago.