Enabling a Converged World™

Measuring Latency in Equity Transactions

915-4015-01 Rev A May, 2012

Introduction

The importance of minimizing latency in data center networks has been discussed intensely in the last year, mostly with respect to trading in financial markets – where time is literally money. Device and network latency is critically important in data centers of all types. Cloud data centers can benefit significantly from lower-latency devices.

Ultra-low latency (ULL) is a term used to refer to the cutting edge of technology. ULL networks (ULLN) are those used to implement ULL for specific or generalized applications. Some state-of-the-art latency values as of mid-2012 are shown in Table 1.

The importance of minimizing latency in data center networks has been discussed intensely in the last year, mostly with respect to trading in financial markets – where time is literally money.

Usage	State-of-the-Art Latency
10Gbps Ethernet Switches	10 nanoseconds
Core Internet Router	50 microseconds
Trade Latency	85 microseconds
Metro Area Network	20 milliseconds
Wide Area Network (Trans-Atlantic)	75 milliseconds

Table 1. State-of-the-Art Latency Values

Why is latency important? In financial transactions, latency affects the time between the initiation of a transaction and its execution. For example, when a brokerage's automatic trading system decides that the price is right for a buy or sell operation, it initiates a trade. Any latency between the initiation of a trade and the match of the buy/sell to an appropriate sell/buy order provides an opportunity for the price to make an unfavorable change. Additional latency can literally cost money.

High-frequency trading (HFT) is a common mechanism by which large numbers of small transactions are placed to capitalize on small variations in stock prices. Some estimates attribute up to 50% of all trades as originating from HFT systems. Highly-sophisticated computer programs are used to place the trades, often in markets around the world. Latency directly affects the number of trades that can be placed, along with calculations about where to place the trades.

According to the TABB Group "… estimates that if a broker's electronic trading platform is 5 milliseconds behind the competition, it could lose at least 1% of its flow; that's $4 million in revenues per millisecond."[1]

Likewise, latency affects Internet purchases. Amazon estimates that each 100 millisecond of latency can cost them 1% of sales.[2] Incidentally, Amazon.com had USD 48 billion in sales in 2011.

1 "The Value of a Millisecond: Finding the Optimal Speed of a Trading Infrastructure," The TABB Group report, April 8, 2008.
2 "Make Data Useful," slide presentation by Greg Linden. http://www.scribd.com/doc/4970486/Make-Data-Useful-by-Greg-Linden-Amazoncom

What is Latency?

The meaning of "latency" differs depending on the area of application. In electronic networks, it refers to the time between two events. For example:

- The time between the first bit of a packet entering a network switch at one port and the first bit appearing at another port. This time can be on the order of 10 nanoseconds.
- The time between the start of transmission of a market trade and receipt of the complete trade acknowledgement packet. This time can be on the order of 100 milliseconds.
- The time between a packet leaving New York City and arriving in London. This time can be on the order of several hundreds of milliseconds.

Sensitivity to latency and jitter varies by the type of data being transmitted.

A related term is "jitter," which is the amount of variation in latency. An exchange may have an average latency or 100 milliseconds, but if its jitter is 500 milliseconds or more, then it will be viewed as less dependable than systems with less jitter. Observed latency values tend to be distributed in a bell curve around a (close to) average value. The long tail of the solid line in Figure 1 reflects exceptional "latency outliers" caused by unusual events. System architects and engineers work to reduce these unusual events that cause extended latency.

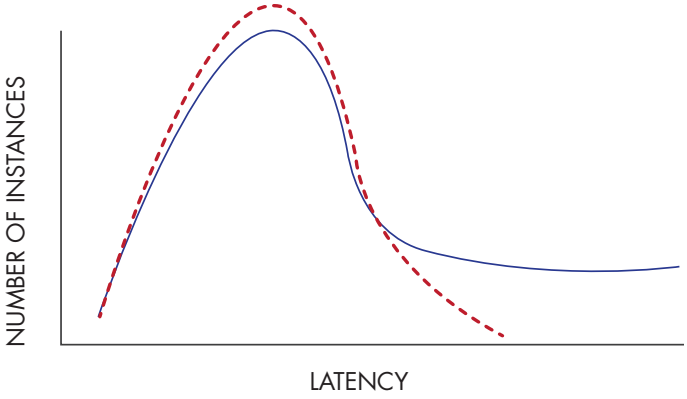


Figure 1. Latency Distribution

Sensitivity to latency and jitter varies by the type of data being transmitted. Voice over IP (VoIP) traffic, for example, is very low bandwidth (less than 56k bps) but very sensitive to latency. If more than 50 milliseconds of latency occur, the human ear immediately hears the delay. Streaming video applications have a larger bandwidth requirement (about 500k bps), but are far less sensitive to latency and jitter due to various buffering techniques in the transmission and reception networks and devices. Streaming video can tolerate up to 5% loss and up to 5 seconds of latency. In general, web-based applications have widely-varying bandwidth requirements and are relatively insensitive to latency and jitter.

What Causes Latency?

Latency is a fact of life, due mostly to the number and types of devices between the parties involved in a transaction. Data center, metro, and wide area networks all generate delay due to the time required to send data over a distance, and through networking devices such as routers and switches. Additional latency is caused by the processing of requests by computers and their associated storage media.

In this white paper, we'll explore these latency sources in the context of equity trading, what steps are being taken to reduce latency, and how to test networks and systems for latency.

Latency in Financial Networks

Electronic trading has grown rapidly in the last 41 years, since starting with NASDAQ in 1971. In the U.S. market alone there are more than 12 regulated security exchanges, approximately 40 alternate trading systems (ATSes), 9 equity options markets, and 13 futures exchanges. Cash equity trading is largely handled by four parties:

In this white paper, we'll explore these latency sources in the context of equity trading, what steps are being taken to reduce latency, and how to test networks and systems for latency.

- NYSE – 35%
- NASDAQ OMX – 20%
- BATS – 10%
- Direct Edge – 10%

The situation is much the same in Europe and Asia, with substantial numbers of trading entities. Asia has 48 exchanges and Europe has 7 major exchanges, plus another 80 lesser exchanges. Figure 2 is a simplified picture of the electronic trading chain.

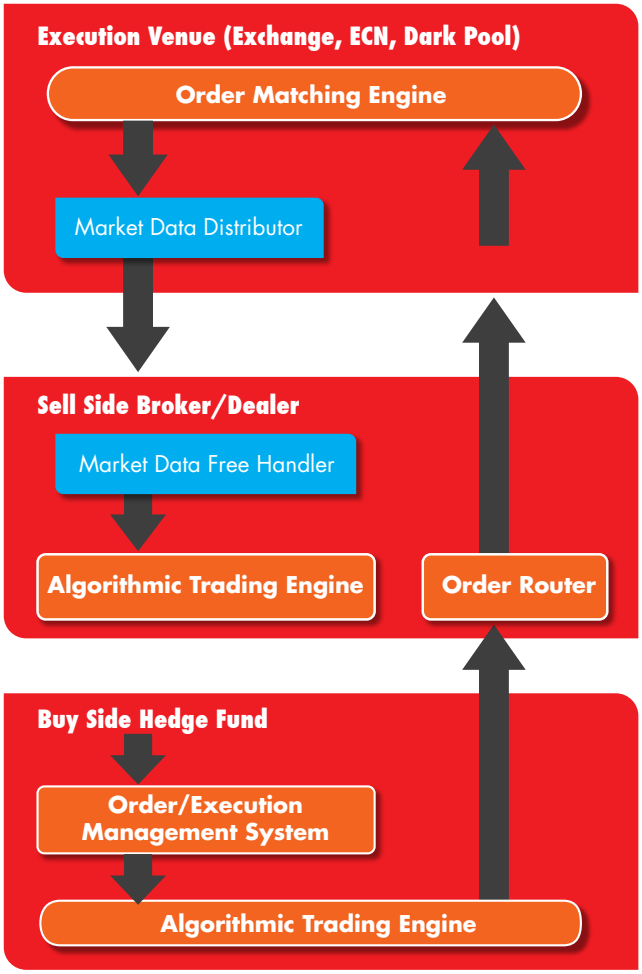


Figure 2. Simplified Electronic Trading Chain

The key players involved in trading and their roles are:

- **Execution Venue**. Execution venues are where buyers are matched to sellers through the operation of an order matching engine. These are operated by major exchanges, ATSes, and Dark Pools. The last two are private venues that do not publish the names of the parties involved in transactions.

  The order matching engine is the heart of an execution venue's trading system, maintaining a very large list of buy and sell orders that are matched in microseconds. When a match occurs, both parties in the sale are notified and the notice of the trade is sent to the market data distributor. This entire process usually occurs in less than 10 milliseconds.

  The market data distributor sends trades to a wide distribution and can amount to millions of trades per second for large exchanges. The aggregated volume of market data from all feeds is monitored real-time by www.marketdatapeaks.com. The statistics for May 3, 2012 are shown in Figure 3.

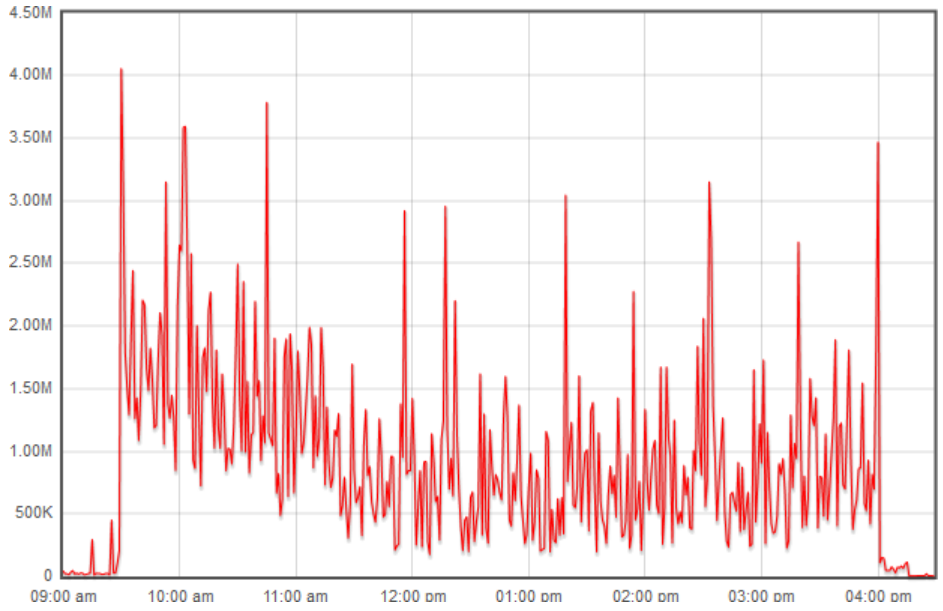


Figure 3. Market Data Volume for May 3, 2012

- **Sell-Side Broker/Dealer**. These traders receive market data feeds from a number of sources. In addition to trades requested by their clients, they use algorithmic trading to decide when to execute a trade. The actual execution venue used is determined by order router software that considers the current prices at multiple exchanges, as well as the latency to reach those exchanges.

- **Buy-Side Hedge Funds**. These funds operate in much the same manner as the sell-side brokers, but operate through them.

Larger brokers and traders often use direct market access (DMA) technology to place buy and sell offers directly into the exchange's match book. They, in turn, can provide this service to their larger clients. This minimizes the latency associated with trades by the brokers and their customers.

Larger brokers and traders often use direct market access (DMA) technology to place buy and sell offers directly into the exchange's match book. They, in turn, can provide this service to their larger clients. This minimizes the latency associated with trades by the brokers and their customers.

Algorithmic trading accounts for a growing percentage of trades in equity markets. In many cases, larger trades are broken up into smaller ones and distributed into multiple

exchanges so as to not affect the price at any one exchange. Low latency here means less price slippage from when the buy/sell decision is made until the trade actually happens.

An aggressive form of algorithmic trading is high-frequency trading (HFT), which involves very rapid buying and selling of securities so as aggregate small profits over a day. There are approximately 400 firms in the U.S. that engage in HFT, accounting for 60-70% of all trades. Together with algorithmic trading, this accounts for about 80% of all trading.

An aggressive form of algorithmic trading is high-frequency trading (HFT), which involves very rapid buying and selling of securities so as aggregate small profits over a day.

Unexpected trading can happen when market movers are not careful about their trades. See the article at the bottom of the next page for a description of the May 6, 2010 "Flash Crash." This type of surge in trading activity has an adverse effect on a large number of exchanges – increasing update traffic as well as computational and network delays.

Key Trading Latencies

Typical latencies for different types of electronic trading applications are shown in Table 3.

Application	Description	Typical Latencies
Low-Frequency Alpha Trading	Long-term trading designed to generate consistent returns	100s of milliseconds
FX and FI Market Making	Foreign Exchanges (FX) and Financial Institutions (FI) that offer both buy and sell price quotes	10s of milliseconds
Prime Brokerage Services	A bundled package of services offered by investment banks and securities firms to hedge funds and other professional investors	10s of milliseconds
Derivative Pricing	Pricing of a security whose price is dependent-upon or derived-from other assets.	Milliseconds
Equities DMA Services	Direct market access (DMA) to exchange market books	Milliseconds
High-Frequency Trading	A mechanism by which large numbers of small transactions are placed to capitalize on small stock price variations	Milliseconds
Latency Arbitrage	Trading designed to take advantage of different markets' price differences	Microseconds

Table 2. Typical Latencies in Electronic Trading

Key Trading Transactions

Although there are a number of electronic trading applications, the key latency that the financial community is aggressively addressing is that associated with co-located brokers executing trades with their associated exchange. The primary protocol[3] messages associated with such trades are shown in Figure 4.

3 FIX is a standardized protocol used by most exchanges. Proprietary APIs, however, are in widespread use. In either case, the types of messages shown here are used.

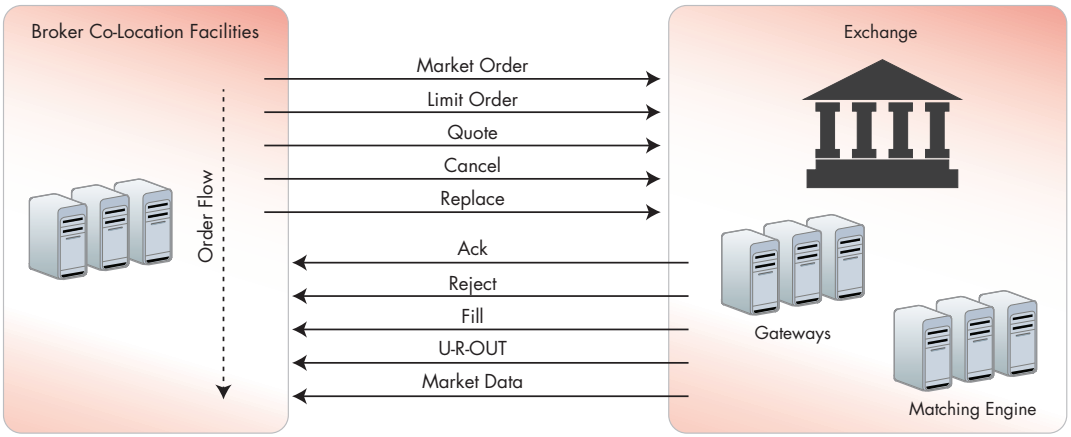


Figure 4. Primary Trading Messages

The messages shown cover those used for most types of orders and states. The key transactions that account for the vast majority of trades are:

- Order to acknowledge
- Quote to acknowledge
- Cancel to confirm
- Order to market data update

The latencies associated with these transactions are critical in responding to trading signals for HFT.

From Wikipedia: "Flash Crash"*

On May 6, US stock markets opened down and trended down most of the day on worries about the debt crisis in Greece. At 2:42 pm, with the Dow Jones down more than 300 points for the day, the equity market began to fall rapidly, dropping more than 600 points in 5 minutes for an almost 1000 point loss on the day by 2:47 pm. Twenty minutes later, by 3:07 pm, the market had regained most of the 600 point drop.

The SEC and CFTC joint report itself says that "May 6 started as an unusually turbulent day for the markets" and that by the early afternoon "broadly negative market sentiment was already affecting an increase in the price volatility of some individual securities." At 2:32 pm (EDT), against a "backdrop of unusually high volatility and thinning liquidity" that day, "a large fundamental trader (a mutual fund complex) initiated a sell program to sell a total of 75,000 E-Mini S&P 500 contracts (valued at approximately $4.1 billion) as a hedge to an existing equity position." The report says that this was an unusually large position and that the computer algorithm the trader used to trade the position was set to "target an execution rate set to 9% of the trading volume calculated over the previous minute, but without regard to price or time."

As the large seller's trades were executed in the futures market, buyers included high-frequency trading firms … and within minutes these high-frequency trading firms also started aggressively selling the long futures positions they first accumulated mainly from the mutual fund. The Wall Street Journal quoted the joint report, "'HFTs [then] began to quickly buy and then resell contracts to each other—generating a 'hot-potato' volume effect as the same positions were passed rapidly back and forth.'" The combined sales by the large seller and high-frequency firms quickly drove "the E-mini price down 3% in just four minutes."

*Source: http://en.wikipedia.org/wiki/Flash_crash

The Low-Latency Arms Race

The Low-Latency Arms Race is an effort by market exchanges to reduce latency, not only for key transactions, but for all exchange processing. Efforts cover the gamut of computing and networking technologies.

Most of the world's exchanges are working to reduce the exchange order response times (rather effectively). Figure 5 shows a steady, exponential decrease in response times. In October of 2011 the Singapore Stock Exchange (SGX) claimed a world-leading door-to-door latency of 79 microseconds as part of their Reach initiative.

The Low Latency Arms Race is an effort by market exchanges to reduce latency, not only for key transactions, but for all exchange processing.

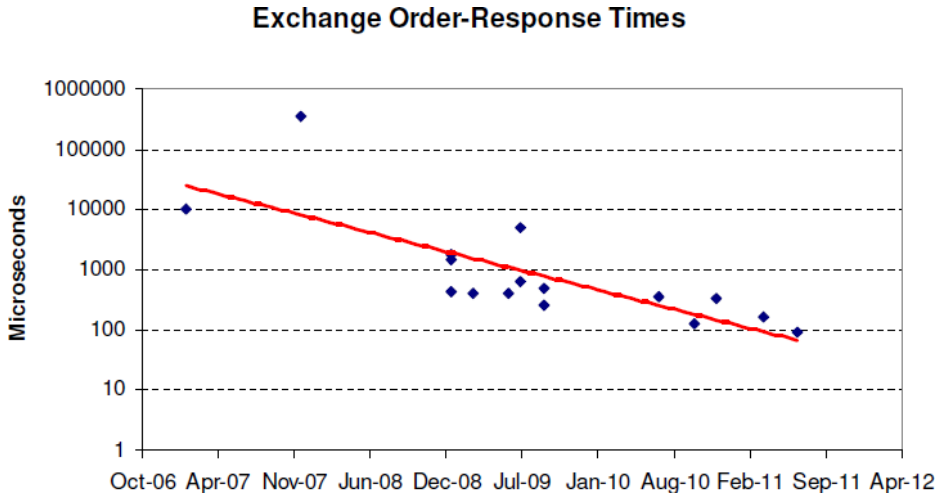


Figure 5. Decreasing Exchange Order-Response Times[4]

The door-to-door latency is the time that it takes for an exchange to receive, process, and acknowledge an order as measured from the customer side of the exchange firewall/gateway within the exchange's data center.

The variation in latency – jitter – is another critical component. If there are significant deviations in latency, then exchange users must base their trading calculations on a larger-than-the-average value. Figure 6 shows an analysis of 5 minutes from the SGX, as measured by their Corvil network monitors. The circled item shows a single event that exhibited significant latency above the average.

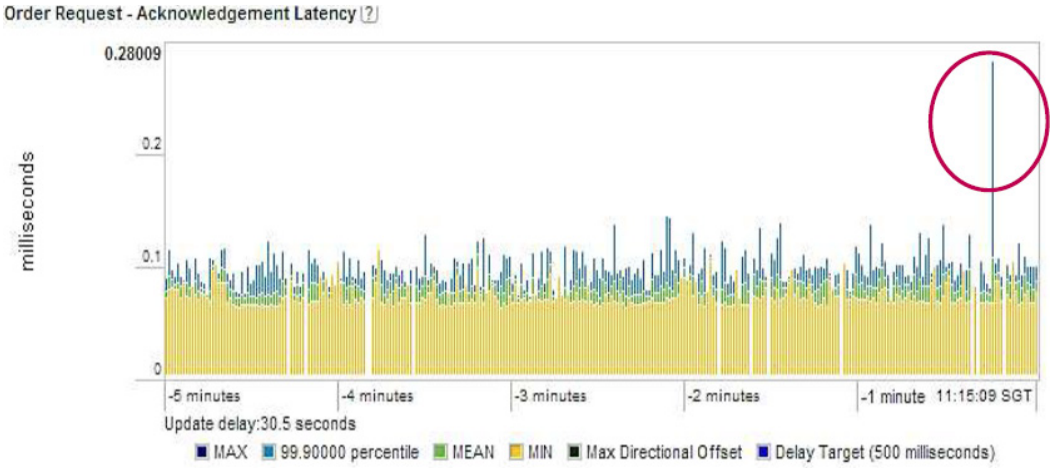


Figure 6. Jitter

4 Source: Low-latency.com white paper, authored by Corvil.

The Sources of Latency

Trading latency can be broadly broken into three general areas:

- Transmission delays – associated with transmitting light over fibers
- Networking delays– associated with forwarding Ethernet packets
- Computational delays – associated with processing trades

Transmission Delays

Long-Distance Delays

Larger financial institutions are co-located with the exchange(s) that they operate through. That is, their computers are in the same building as the exchange and other financial institutions' computers. Exchanges go to extraordinary efforts to ensure that none of the many institutions that they support has any latency advantage over another.

Smaller institutions go to great efforts to minimize the latency of their connection to their exchange(s), often buying high-speed connections: 1Gbps or even 10Gbps. They do not necessarily need the bandwidth, but the connection rate minimizes the delay inherent in transmission rates and network processing. The New York Stock Exchange (NYSE) has painstakingly worked to build a sub-millisecond network that connects approximately 50 markets in the U.S. Their attitude is: if we can't make the trade, we'll make money moving the trade.

Smaller institutions go to great efforts to minimize the latency of their connection to their exchange(s), often buying high-speed connections: 1Gbps or even 10Gbps.

Where co-location is not available, traders must depend on the Internet or dedicated connections. The latency of any connection is determined by physics: fiber optic network transmission occurs at approximately 2/3 of the speed of light: 5 microseconds per kilometer. Table 3 is a recent report from two major carriers of their measured long-line latencies.

Carrier: Verizon	Measured Latency	Carrier: AT&T	Measured Latency
Trans-Atlantic	< 80ms	Los Angeles – New York	68ms
Intra-Europe	< 19ms	New York – London	74ms
Intra-North America	< 45ms	San Francisco – Hong Kong	129ms
Intra-Japan	< 16ms	Los Angeles – Chicago	47ms
Trans-Pacific	< 120ms		
Intra-APAC	< 110ms		
Intra-Latin America	< 138ms		

Table 3. Sample Long-Line Latency

The numbers and variation indicate that there's more going on than the speed of transmission. The distance between Los Angeles and New York, for example, is about 4000km – which equated to a light transmission delay of 20 milliseconds. This is less than half of the 68 millisecond latency observed by AT&T. As we'll see, assorted signal processing, routing, and switching account for additional delay.

Networking Delays

Networking delays are caused by optical and copper cables, and networking devices.

Optical and Copper Cables

Most long connections between data centers takes place over fiber optic cables: single-mode for longer distances and multi-mode for shorter distances.

Most long connections between data centers takes place over fiber optic cables: single-mode for longer distances and multi-mode for shorter distances. Within the data center and between racks (Figure 7), copper cabling is frequently used for cost considerations. As in optical networks, the speed of transmission is 2/3 of the speed of light.

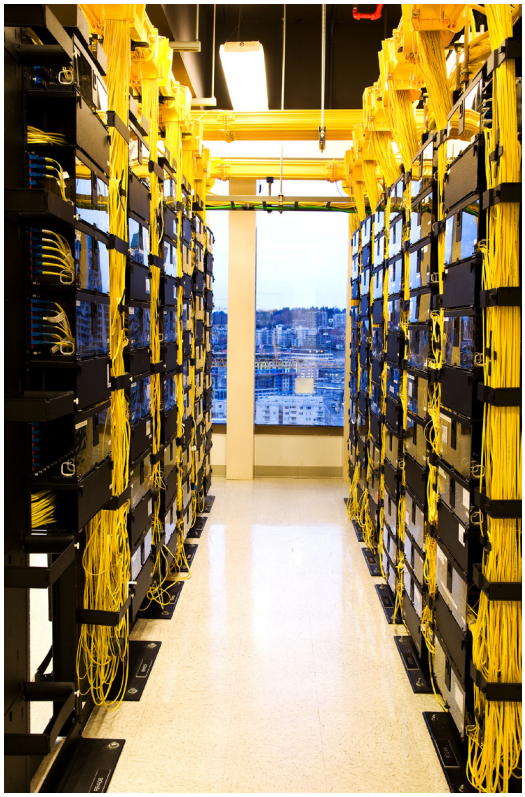


Figure 7. Data Center Rack

Many financial data centers are quite large. As discussed above, the transmission of Ethernet signals requires 5 microseconds per kilometer. Some data center connections are up to a kilometer in length. Although 5 microseconds is a small amount, we will see that this delay is of the same order of magnitude as that seen in some Ethernet switches.

Networking Devices

Information must go through a number of networking devices between parties. The devices vary in function, sophistication, and amount of latency that they contribute.

For multi-kilometer connections the addition delays above and beyond transmission through the media is due to:

- **Optical transmission/retransmission devices**. Color conversion, amplification, dispersion compensation, and regeneration are all common optical networking functions that are carried out by different devices. Transponders and muxponders on each end of a connection convert "gray" signals to/from specific light colors and

aggregate them into single high-speed channels. The best of these devices operate with about 1 microsecond of latency.

Optical signals weaken as they travel. Depending on the length of the connection and the type of cable used, additional amplification and/or regeneration steps may be necessary. This may add hundreds of nanoseconds to small numbers of microseconds of latency.

- **Routers**. Routers are the principal devices that connect the backbone of the Internet and major nodes in private networks. They handle large volumes and types of traffic, numerous protocols, and operate over a large number of network connections of various speeds. Large core routers, such as Cisco's CRS-1, have latencies in the range of 50-500 microseconds[5] in uncongested networks. The variation can be due to the number of routing protocols in operation and number of routes that need to be held in the router's memory. Smaller routers used in data centers exhibit average latencies of 20-200 microseconds.

  When a router is in the process of transmitting a packet and another packet becomes ready for transmission, the latter packet must be queued. Depending on the instantaneous contention for the connection, queuing delay can result in addition latency. This variation shows up as jitter on the line.

Routers are the principal devices that connect the backbone of the Internet and major nodes in private networks.

The large-scale devices that are the "front door" to exchange, broker, and other data centers include:

- **Firewalls and other network security devices**. These protect communications from outside the data center, offering a wide range of security services. The simplest function is basic firewalling – only allowing communications from specific sources, protocols, and UDP/TCP port numbers. Additional services include data encryption/decryption, anti-virus protection, and intrusion protection. The number and type of services determines the amount of latency that may occur, from microseconds for basic firewalling services to seconds for intrusion protection.

- **Load balancers**. Data centers must service large numbers of request, requiring a large number of computational servers. Load balancing devices, often referred to as application delivery controllers (ADCs), are used to distribute requests among the servers. Intelligent ADCs used a technique known as deep packet inspection (DPI) to sort out one type of requests from another, using that information to direct requests to the best server for the job. Load balancers in general, and DPI functions specifically, add additional latency to requests.

5 Deploying QoS for Cisco IP and Next-Generation Networks: The Definitive Guide, by Vinod Joseph, Brett Chapman.

Switching

By and large, data centers are switched networks. That is, traffic is routed based on low-level, layer 2 MAC addresses and other bits of information included in the layer 2 packet headers. Historically, data center networks have been organized in three layers[6], as shown in Figure 8.

By and large, data centers are switched networks.

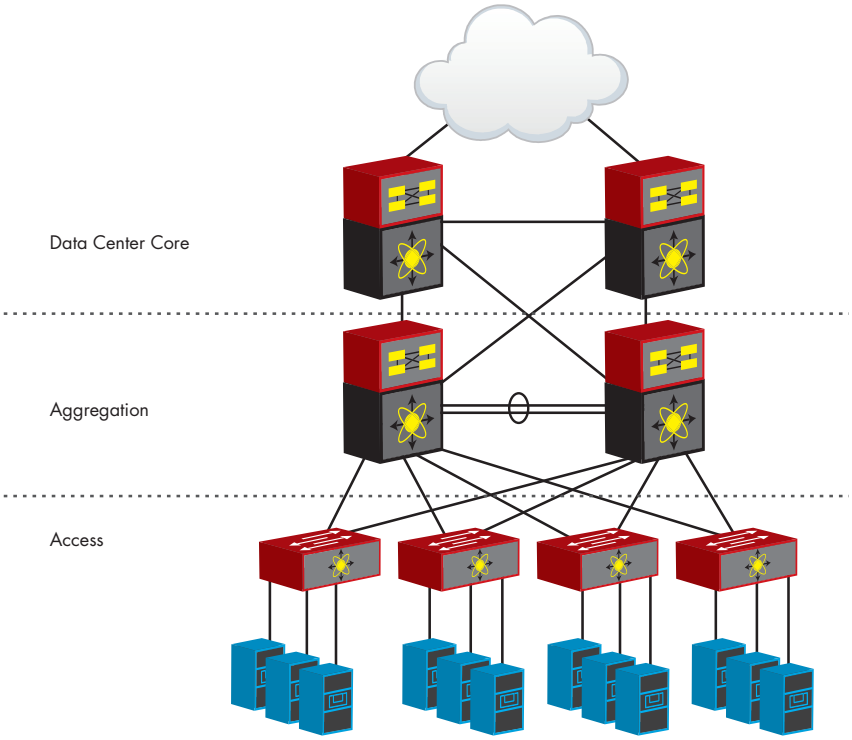


Figure 8. Three-Level Data Center Network

Layers of switches are used at:

- **The access level**. These switches are used to connect servers within a rack and to other racks. These are often referred to as top-of-rack (ToR) switches.
- **The aggregation level**. These switches are used to connect a set of racks to each other. They're often placed at the end of a row of racks, and are referred to end-of-row (EoR) switches.
- **The data center core**. These connect all of the aggregation-level switches to create an environment where any data center device is accessible from any other.

Each switching layer adds latency, as we'll detail in a minute. There has been a recent effort to reduce the number of layers to two by merging the access and aggregation levels. For example, Juniper's QFabric architecture (Figure 9) connects ToR-type nodes (outer dots) through a high-speed interconnect of 40Gbps cables (white lines).

6 There is actually another layer of switching that has no physical manifestation. It's called a virtual switch and is a software component within a virtualized server that connects the virtual machines (VMs) inside the server to each other.

Figure 9. QFabric Architecture

When a switch receives a packet on one of its ports, it forwards it to another in one of two manners:

Cut-through. As soon as it reads the layer 2 header of the packet and can determine which port to forward the packet to, it immediately transmits the header and follows with the remainder of the packet. This can only happen if the speed of the output port is less than or equal to that of the input port. Cut-through processing is found largely in ToR and EoR switches.

Store and forward. The entire incoming packet is read and stored on the switch before forwarding. This technique is frequently used in larger core switches, due to the requirement to implement advanced features such as quality of service (QoS) and network address translation (NAT).

The minimum latency of a packet sent through a switch using store and forward processing is determined by the length of the packet and the speed of transmission. This is often referred to as the serialization delay. Table 4 describes the store and forward latency for various packet lengths and network speeds.

Packet Size	1 Gbps	10 Gbps	40 Gbps
64	0.51 microseconds	51 nanoseconds	5 nanoseconds
128	1.02 microseconds	102 nanoseconds	10 nanoseconds
256	2.04 microseconds	204 nanoseconds	20 nanoseconds
512	4.08 microseconds	408 nanoseconds	40 nanoseconds
1024	8.16 microseconds	816 nanoseconds	82 nanoseconds
1518	12.1 microseconds	1.2 microseconds	121 nanoseconds

Table 4. Serialization Delays

In 2011, the Lippis Report performed an exhaustive test of data center and core switches[7] using Ixia test equipment. Among the top performers in the ToR category employing cut-through operation was the Dell/Force 10 Networks S4810. The range of observed latencies was between 800 to 900 nanoseconds, typical for best-in-class ToR 10Gbps switches.

In 2011, the Lippis Report performed an exhaustive test of data center and core switches using Ixia test equipment.

7 Fall 2011 Open Industry Network Performance And Power Test Report

Within the data center, traffic is characterized as north-south or east-west, as shown in Figure 10.

Improving computational delays is the province of the trading application developers, along with their software suppliers, and is beyond the scope of this white paper.

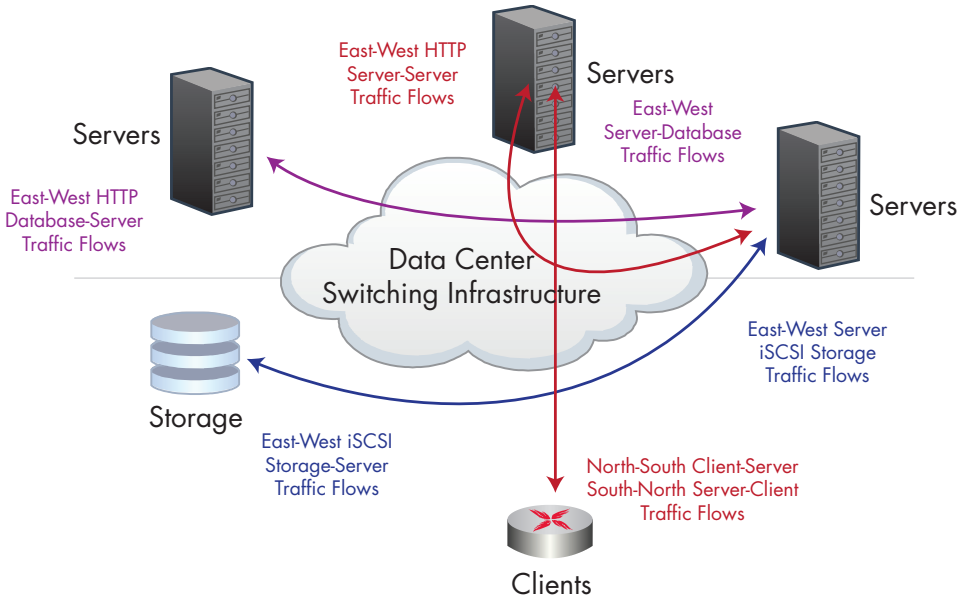


Figure 10. Data Center East-West and North-South Traffic

North-south traffic flows between client and servers in the data center and comprises 20% of observed data traffic. East-west traffic connects servers to other servers, to their storage arrays, and comprises 80% of the observed traffic.

Computational Delays

Improving computational delays is the province of the trading application developers, along with their software suppliers, and is beyond the scope of this white paper.

There are, however, situations that cause the amount of market data and trading to increase dramatically. This includes market opening and closing, breaking news, and algorithmic and high-frequency trading (see the article on page 6). This increases computational delay, but also increases network traffic that can result in additional delays.

Such events result in jitter (the variation in latency), which wreaks havoc on algorithmic trading dependent on specific latencies for specific markets.

Measuring Latency

Improving latency of all types is a desirable objective. To know where to focus the improvement, it's essential to understand the latencies of the network's component systems. Testing provides the means and measurements for identifying sources of latency. Vendor specifications are fine as far as they go, but do not provide a true picture of when their devices are subject to a particular data center's workload. Different usage models will result in different workloads for individual devices, sub-systems, and the entire system. Networking devices are very complex combinations of hardware, software, and firmware that are configured by human beings. A failure to prioritize types of traffic and workload can lead to unanticipated latencies.

Several testing scenarios must be addressed:

- **Component selection and characterization testing**. This type of testing occurs when a new system is being built or an expansion is planned.
- **Pre-deployment testing**. This generally takes place in a special development/test lab that recreates a subset of the live system.
- **Full system test**. This type of testing uses the complete system when it is offline.
- **Live system test**. This type of testing happens while the system is live, which restricts the type of operations that can be performed.

Component Selection and Characterization Testing

Component selection is a critical part of data center design. The specifications of networking, computing, and storage components need to be verified – those that come from the equipment manufacturers are the best possible values and may not be realistic. In order to establish the proper performance parameters that relate to a site's activity, all components must be tested with the type of network traffic and applications that will run on the component.

The specifications of networking, computing, and storage components need to be verified – those that come from the equipment manufacturers are the best possible values and may not be realistic.

For example, an ADC must be tested with a mix of real-world traffic that mimics the traffic the ADC will experience when deployed. This provides a true measure of capacity that can then be matched with other components. In the case of an ADC, the number of servers per ADC can be accurately determined so that the ADC is not a limiting the overall system performance.

Component selection testing is also essential in ensuring interoperability between components. Although networking components from different vendors have become relatively interchangeable, there's always a possibility that networking protocols are implemented differently. For example, the TRILL protocol[8] is a relatively new protocol for optimizing networking paths with a data center. As the protocol matures, there are opportunities for implementation differences by different vendors. Conformance testing, as it's called, confirms that network devices adhere to published standards.

Pre-Deployment Testing

This category applies the techniques used in component selection and characterization testing to complete systems and sub-systems. For example, the following data center sub-systems may be tested and characterized:

- Firewall and related security appliances
- Multi-level switching infrastructures
- ADCs and virtualized servers
- Servers and storage arrays (SANS)

8 Transparent Interconnection of Lots of Links, http://tools.ietf.org/html/rfc5556.

As in component testing, it is essential that real-world conditions be used. The principal technique used for this type of testing is called "emulation," wherein interactions are simulated with software. For example, a web server that is used to host customer-facing account management is tested by performing the same interaction that an end-user's browser would have with web server.

Modern test equipment, such as that offered by Ixia, can be used to emulate many thousands of end-users interacting with the service. During this emulation testing, key performance indicators (KPIs) are measured to determine acceptable behavior. KPIs vary by application; for web applications the KPIs include the number of concurrent connections, transactions per second, throughput, latency, and jitter. Latency is measured in time to first byte (TTFB) and time to last byte (TTLB), which indicate when the response page first starts to display and when it is finished, respectively.

As in component testing, it is essential that real-world conditions be used. The principal technique used for this type of testing is called "emulation," wherein interactions are simulated with software.

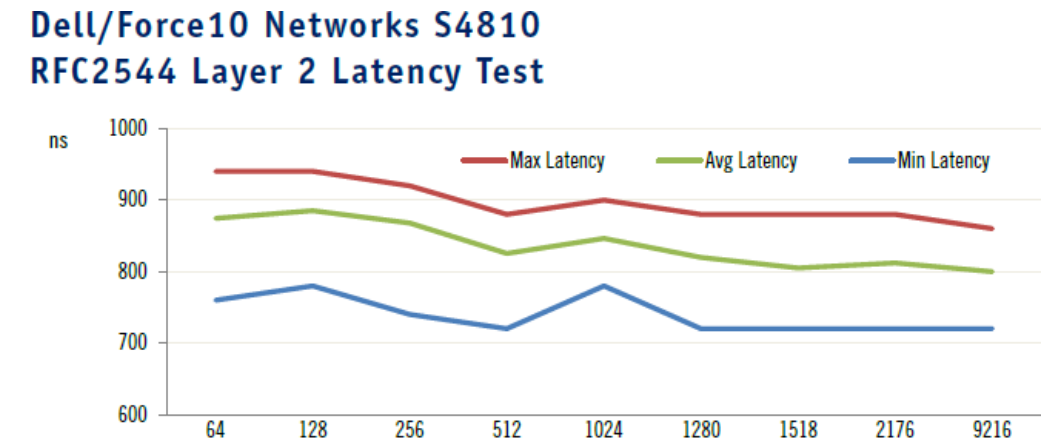
Emulation can likewise be effective in measuring trading latency. Trades embodied in FIX or proprietary APIs can be emulated, and key latencies measured. In order to determine where processing overheads differ, it is important to measure all KPIs for:

- Different trading sessions – to ensure consistency,
- Different securities and symbols – to ensure consistency across different parts of the market, and
- Different order types – which can affect trade strategies.

The transmission of network traffic within the data center is a key factor in transactional latency. Emulation is particularly effective in measuring infrastructure performance. Various levels of traffic, up to line rate at 1Gbps and 10Gbps, can be generated across network paths in anticipated quantities. Overall system performance becomes immediately evident and bottlenecks quickly identified. It is particularly import to generate high-volume micro-bursts to determine the infrastructure's elasticity.

Emulation is used to test larger and larger sub-systems. For example, after measuring web server performance it would be logical to test the performance of multiple servers with their associated ADCs. Continuing the technique, after testing the security and switching sub-systems an entire system could be performance-tested to determine the latency of interactions from the firewall at the front door through to the web server and their back-end components. Pre-deployment testing also proves interoperability on a larger scale.

As discussed previously, the Lippis report did an extensive measurement of data center switch latencies. Figure 11 shows the measured latencies across 48 ports of 10Gbps operating at full line rate.



	64	128	256	512	1024	1280	1518	2176	9216
Max Latency	940	940	920	880	900	880	880	880	860
Avg Latency	874.479	885	867.604	825.292	846.521	819.896	805.063	812.063	799.792
Min Latency	760	780	740	720	780	720	720	720	720
Avg. delay Variation	9.104	5.229	5.458	7.938	6.729	6	8.979	5	9.271

Figure 11. Dell/Force 10 Networks S4810 Latency Measurements

The range of observed latencies was between 800 to 900 nanoseconds, typical for best-in-class ToR 10Gbps switches. It is expected that this number will continue to fall toward 100ns by 2014.

Full System Test

Pre-deployment labs are an essential part of deploying and maintaining a large-scale system. Pre-deployment labs usually represent a small subset of the larger system that they support. There's no substitute, however, for the real thing. Where the opportunity exists, the full system can be used for end-to-end system test. Trading systems are often available on the weekends for such testing.

Pre-deployment labs are an essential part of deploying and maintaining a large-scale system.

The same tests that were used in pre-deployment testing may be used for full system test, but it may not be possible to create enough emulated sessions to match the capacity of the entire system. It may be necessary to use different techniques to fully stress the entire system. These may include:

- **Use of a customer's system**. For example, a larger broker's system might be used to generate mock trades.
- **Use of recorded sessions**. This is typically only useful where the protocols used by the application(s) are very simple; that is, not requiring an involved interaction.
- **Custom-built test hardware**. Systems built for the express purpose of generating appropriate traffic.

Full system test allows measurement of true end-to-end behavior. Queuing and other bottlenecks can be explored and tuned away.

The Lippis report extended its testing beyond latency measurements using Ixia emulations of storage and application traffic, with the results shown in Table 5. The use of real-world traffic, reflecting expected packet sizes and combinations, provided a better indication of latencies observed in real-world conditions. The values shown in Table 5 actually indicate better latency values than those of the individual packet size tests.

Traffic Direction	Traffic Type	Average Latency (ns)
East-West	Database_to_Server	6890
East-West	Server_to_Database	1010
East-West	HTTP	4027
East-West	iSCSI-Server_to_Storage	771
East-West	iSCSI-Storage_to_Server	6435
North-South	Client_to_Server	1801
North-South	Server_to_Client	334

Table 5. Dell/Force 10 Networks S4810 Cloud Simulation Test

The latency of metropolitan and wide area networks can be measured with "ping" packets, or more extensively tested with live system protocols using advanced tools such as Ixia's IxChariot.

Live System Test

Two techniques have proven useful in measuring live system performance:

- Injection of low-bandwidth transactions
- Monitoring

Low-Bandwidth Injection

While it is important to minimize impact a running system, short duration and bandwidth tests can be used in a number of ways. The latency of metropolitan and wide area networks can be measured with "ping" packets, or more extensively tested with live system protocols using advanced tools such as Ixia's IxChariot. IxChariot uses small software or hardware agents in multiple locations to send and receive network traffic, while reporting KPIs.

Within the data center the networking infrastructure's latency can be tested with tools similar to IxChariot, or more advanced network test tools such as Ixia's IxNetwork. KPIs related to latency can be obtained, although capacity measurements are not possible.

In some cases, special test accounts may enable full application testing. For example, in trading applications a fake broker could execute trades for a fake stock.

Such low-bandwidth tests are often run on a recurring basis, varying from once every second to several minutes. The values obtained can be plotted over time for a day or longer periods.

Monitoring

Monitoring involves the placement of high-speed network taps at key locations in the system. These taps are connected to intelligent concentrators/filters, such as those made by Anue Systems[9], and then forwarded to sophisticated analysis tools. High-speed, high-capacity devices record all or filtered network traffic. Depending on the applications being monitored, some instantaneous measurements may be available. For example, some monitoring equipment used in trading applications can match up FIX transaction messages to provide instantaneous latency measurements. More often data post-analysis is required to provide performance metrics.

9 http://www.anuesystems.com/products/5288-net-tool-optimizer

Conclusion

Latency is an important performance indicator in most networked applications and is the KEY performance indicator in high-frequency security exchange applications. There are many factors, however, that contribute to latency inside of and between data centers. There are a variety of tools available to discover the sources of latency so that those areas may be improved.

The low-latency arms race focuses on the need to squeeze every microsecond out of exchange-based transactions through improvements to hardware, software, and diagnostic tools. We hope that the active test techniques and passive monitoring techniques covered in this white paper help you with your efforts.

Ixia designs, develops, sells, and maintains a variety of test tools for the financial industry.

For more information on our testing solutions for finance, go to:
http://www.ixiacom.com/solutions/enterprise/finance/index.php

For more information on our Ultra Low Latency (ULL) Testing Blackbook, go to:
http://www.ixiacom.com/blackbook

Latency is an important performance indicator in most networked applications and is the KEY performance indicator in high-frequency security exchange applications.

Contents

Ixia Worldwide Headquarters
26601 Agoura Rd.
Calabasas, CA 91302

(Toll Free North America)
1.877.367.4942

(Outside North America)
+1.818.871.1800

(Fax) 818.871.1805
www.ixiacom.com

Other Ixia Contacts
Info: info@ixiacom.com
Investors: ir@ixiacom.com
Public Relations: pr@ixiacom.com
Renewals: renewals@ixiacom.com
Sales: sales@ixiacom.com
Support: support@ixiacom.com
Training: training@ixiacom.com