High-Performance Automated Trading Network Architectures

Performance Without Loss—Performance When It Counts

Introduction

Firms in the automated trading business recognize that having a low-latency infrastructure can improve the profits of the business. In this paper, we explore the key considerations to implementing an ultra- low-latency network and highlight some key misconceptions when evaluating the suitability of network platforms for automated trading businesses.

High-Frequency Trading Market

Trading in financial markets is increasingly being performed by software programs. These programs run a variety of businesses including proprietary trading for a firm and services such as best execution for the firm's customers. In particular, the term "high-frequency trading," or HFT, has gained popularity to describe the most intensive of the automated trading businesses.

A competitive advantage in all these businesses is to be able to access the liquidity pool faster than the competition. To this end, firms are focused on lowering the latency of the end-to-end trade flow.

From an exchange perspective, there is desire to provide deep liquidity pools with the quickest access to the market. With securities now trading on multiple venues, being the liquidity choice is instrumental to the exchange's competitiveness.

> Mondo Visione asked 14 exchanges to chart how tick size reduction had affected average order size by value and found it had fallen by 52 percent since 2005.
>
> Falling trade sizes means more trades, which offsets the impact of lower trade fees, but it also means increased network traffic to contend with.
>
> **—Average trading order size falls by half, Jeremy Grant, Financial Times, Feb 22, 2010**

While the absolute latency is often a focus, the capability of the infrastructure to deliver predictable latency and handle ever-increasing volumes is also a critical requirement. Market data volumes continue to increase at exponential rates. Business factors such as smaller order sizes and reduced tick sizes are driving the increases in data volumes.

Firms will generally disaggregate the trade flow into functional components—application, middleware, server, network—and each functional owner will attempt to remove latency from their area, while improving the predictability and increasing the message capacity of their portion of the trade flow.

Finally, financial firms see the ongoing enhancements to the infrastructure, especially ongoing latency reduction, as a requirement to sustain a competitive advantage.

Latency Contribution Factors

Given the importance of latency reduction then, it makes sense to analyze in more detail the various contributors to the end-to-end trade flow latency. From a network perspective, the latency can be divided into two broad categories—the latency attributable to the network platform and "all-other" latency.

The "all-other" latency category is necessarily broad and not within the scope of this paper. However, the key components include the latency contributions from the application, middleware, OS, NIC, and most importantly the application architecture. Specifically, solutions such as the Cisco® Unified Computing System (UCS) server platform, with extended memory technology, can enable customers to consolidate the trade flow into one server. The benefit is that applications are now working at memory speeds and not network speeds. If customers have the flexibility to do so, re-architecting the application may provide the biggest performance benefit. Please refer to http://www.cisco.com/go/hft, for benchmarks that include the full infrastructure stack.

From a network perspective, there are five latency contributors, listed in increasing order of importance.

- **Serialization delay:** This is the delay to place a packet on the wire and is tied to the interface speed. At 10 Gigabit Ethernet, the serialization delay to send a 128-byte frame is 0.1 µsec. At 1 Gigabit Ethernet, the serialization delay for the same frame is 1 µsec. While this is not a large delta, it should be considered that the delay would be accumulated at every port. Further, 1 Gigabit Ethernet switches generally have larger nominal latencies and work in store-and-forward mode, where latency increases with packet size. In general, firms that desire to deploy low-latency infrastructures would implement 10 Gigabit Ethernet wherever possible.
- **Propagation delay:** Light takes about 3 µsec to traverse 1 km in fiber. To reduce this contribution, firms implement their environments to be as close to the liquidity pool as possible. This may be at the exchange data center itself using a co-location service or at a service provider's data center. However, it is useful to note that for some businesses like best-execution, the algorithms need to check multiple venues, so co-locating centrally to multiple liquidity venues will be more important than to a specific exchange.
- **Nominal switch latency:** This is the latency to traverse a switch hop. Many 10 Gigabit Ethernet switches can operate in cut-through mode (rather than store-and-forward), where the nominal latency is the same regardless of the packet size. This latency is measured in first-in-first-out (FIFO) mode, and can be less than 5 µsec per hop. This latency is what is measured by standard RFC tests (RFC2544, RFC2889, etc.), and is usually used by the industry to determine the suitability of a network platform for HFT environments. However, this turns out to be a faulty conclusion, because of the much larger contribution of the following two delays.
- **Queuing latency:** This is the latency when packets are queued within a network platform. Packets are generally queued due to egress port congestion. Larger buffers can result in more traffic being queued, *but note that simply having larger buffers does not increase latency*. The buffer is only used if the traffic needs to be queued or if the buffer is not available, the traffic is dropped. (See next point.) The efficiency of queuing algorithm is a key switch attribute, which is not often measured in benchmarks. With queuing latency ranging from tens to thousands of microseconds, depending on the traffic patterns, it can completely dwarf the nominal latency. (Please see the section on microbursts for a longer discussion of the traffic patterns that result in queuing delays.)
- **Retransmission delay:** Delay incurred when an application needs to resend a packet, typically due to packet loss in the network. If the packet buffers are not deep enough, traffic may be dropped instead of being queued. The application design needs to be considered to determine the latency impact of the dropped packet. If the drop was during a TCP session, TCP will take care of the retransmission. However, the delay to begin the retransmission is usually 200 milliseconds (RTO min timer in Linux). Also, a substantial rate of loss can result in congestion collapse on TCP sessions. In the case of UDP, the application may itself retransmit or just lose any dropped data. Application developers have a consistent preference to queue and deliver packets rather than dropping the traffic.

In implementing architectures for ultra low-latency infrastructures, it is important to consider all the five latency contribution factors above. Most importantly, it is critical to benchmark to match the application traffic characteristics. Solely relying on RFC results, for instance, would lead to faulty conclusions.

Measuring Latency Accurately

To paraphrase a popular business adage: “You can only improve what you can measure.” This is true of HFT infrastructures. In particular, since firms care about microsecond latency performance, they need the ability to measure latency accurately to the microsecond level. Ideally, the latency is measured on an end-to-end trade-flow basis. However, since the data is being transformed en-route, from tick data to an actual order, it may only be possible to measure latency performance in segregated functional units, such as latency of the market data infrastructure and order management system. Cisco’s Application Visibility and Management (AVM) suite as well as platforms such as those from Corvil can be utilized to get accurate measurements of latency across the trade flow.

It is also important for the servers to be synchronized to accurate, microsecond granular clocks. Cisco offers an implementation of the Precision Time Protocol (IEEE 1588) for this purpose.

> The extent to which the network connection to the member will cope with the microbursts exceeding the available bandwidth, without packet loss, will depend heavily on the buffers in the end-to-end path.
> **—BATS Connectivity Manual**

Switch Architecture Considerations

Automated trading environments typically are implemented in co-location facilities to reduce propagation delay and typically have tens to low-hundreds of server ports. For this reason, “top-of-rack” form factor access-layer switches are used most often to interconnect the servers.

Traditionally, these access-layer switches have been implemented using the “shared memory” switch architecture with a “switch-on-chip” implementation. In more recent times, access-layer switches are also being implemented using the port-ASIC/crossbar switch architecture, which was previously only available in high-end modular switches and routers. The rest of this section considers the differences between the Cisco Nexus® 5000 Series and Arista 7100 Series switch platforms.

There are two main benefits to the port-ASIC/crossbar architecture from a performance perspective.

- **Adequate buffering:** The architecture allows switch designers to build adequate buffering to handle short-lived network congestion. Since designers have multiple chips, they can accept the high cost of packet buffers to ensure traffic does not get dropped. Switch-on-chip implementations are constrained in that there is a lot of functionality that must be squeezed into one chip, which generally means a significant compromise in the amount of buffering for the switch. This compromise leads to packet drops during even mild periods of congestion.
- **Predictable latency at scale:** Unlike shared memory architectures, the performance does not degrade as the number of ports connected increases. Innovations like the fabric extender architecture enable scaled-out implementations to hundreds of server ports at predictably low nominal latency. With shared memory architectures, the amount of buffering per port is reduced as the number of ports used is increased. This can lead to a situation where an environment works fine at low port counts, but as the customer adds more ports, they experience packet drops. Further, once the building-block size of the switch is exceeded, the switches need to be arranged in a fat-tree implementation to build larger environments. These fat-tree designs may have less performance than the basic building block as detailed in the section below.

There is a tradeoff with the port-ASCI/crossbar architecture in that it will have a slightly higher nominal latency than the basic building-block-size, switch-on-chip architecture. This is due to the traffic having to traverse multiple chips, unlike in the single chip design. However, as detailed in a previous section, the nominal latency is not material with real-world traffic. The following section provides more details.

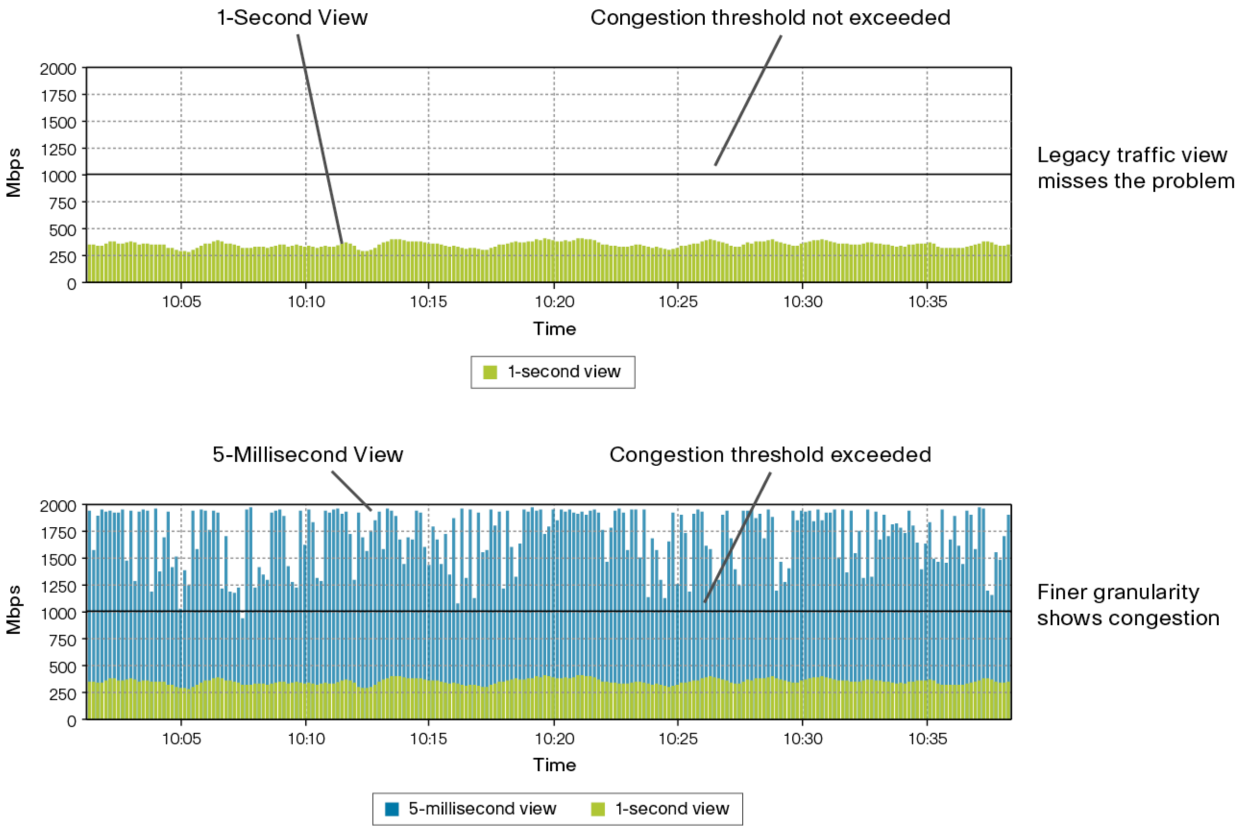
> It is convenient to measure the DUT performance under steady state load but this is an unrealistic way to gauge the functioning of a DUT since actual network traffic normally consists of bursts of frames.
> ***—Section 21, RFC-2544***

Microbursts and Packet Loss

A microburst is a traffic pattern that causes short-lived congestion in the network. This is often due to a period of high activity that causes network endpoints to send bursts of traffic into the network, for example during high-volatility periods in the market. Benchmarking infrastructure performance during these high-volatility periods is especially important because this is when the automated trading businesses make the most profits.

Ironically, these microbursts are often not well understood in customer environments and platform evaluations do not include them during the benchmarking. The primary reason is customers do not have the tools to characterize microbursts in their environments. As Figure 1 shows, when monitoring traffic at relatively large time-scale intervals, the microbursts tend to be smoothed out. Customers may believe that there is small probability of congestion when every link is significantly underutilized. However, since buffer sizes in some implementations tend to be small, it is critical to monitor traffic at the microsecond time interval to detect packet drops due to microbursts.

Figure 1. Two Views of the Same Traffic



Benchmark Results

Miercom recently completed a study of the performance of two network platforms that are positioned for HFT infrastructures, the Cisco Nexus 5010 and Arista 7124S, to understand the impact of microbursts. Please refer to www.miercom.com/cisco for full details.

The benchmark tests four different traffic patterns with microbursts:

- Full mesh multicast
- Full mesh unicast
- Two-to-many multicast (a specific case of few-to-many multicast)
- 23 to one multicast (a specific case of many-to-one multicast)

The results show that the Arista platform drops packets, up to 90 percent, under mild short-lived congestion. It further shows that queuing delay overwhelms the nominal switch latency. Finally, the results show that performance degrades on the Arista switch as the number of ports increase, while the performance stays consistent on the Cisco Nexus 5010 independent of the number of ports used.

Conclusion

Low-latency infrastructures increase the profits of automated trading businesses. However, when evaluating network platforms for suitability for these environments, there is unwarranted focus placed on the nominal latency of a switch. Recent tests by Miercom reveal that queuing delay and packet drops due to short-lived congestion overwhelm the nominal latency of the switch. Moreover, data from end customers and latency monitoring vendors reveal that microbursts that cause short-lived congestion are prevalent on all the links of an automated trading environment. It is critical for customers who are evaluating network platforms for suitability for HFT environments to carefully consider the impact of microbursts during the evaluation.

For More Information

To learn more about high-performance automated trading network architectures please visit http://www.cisco.com/go/hft

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-600126-00 04/10