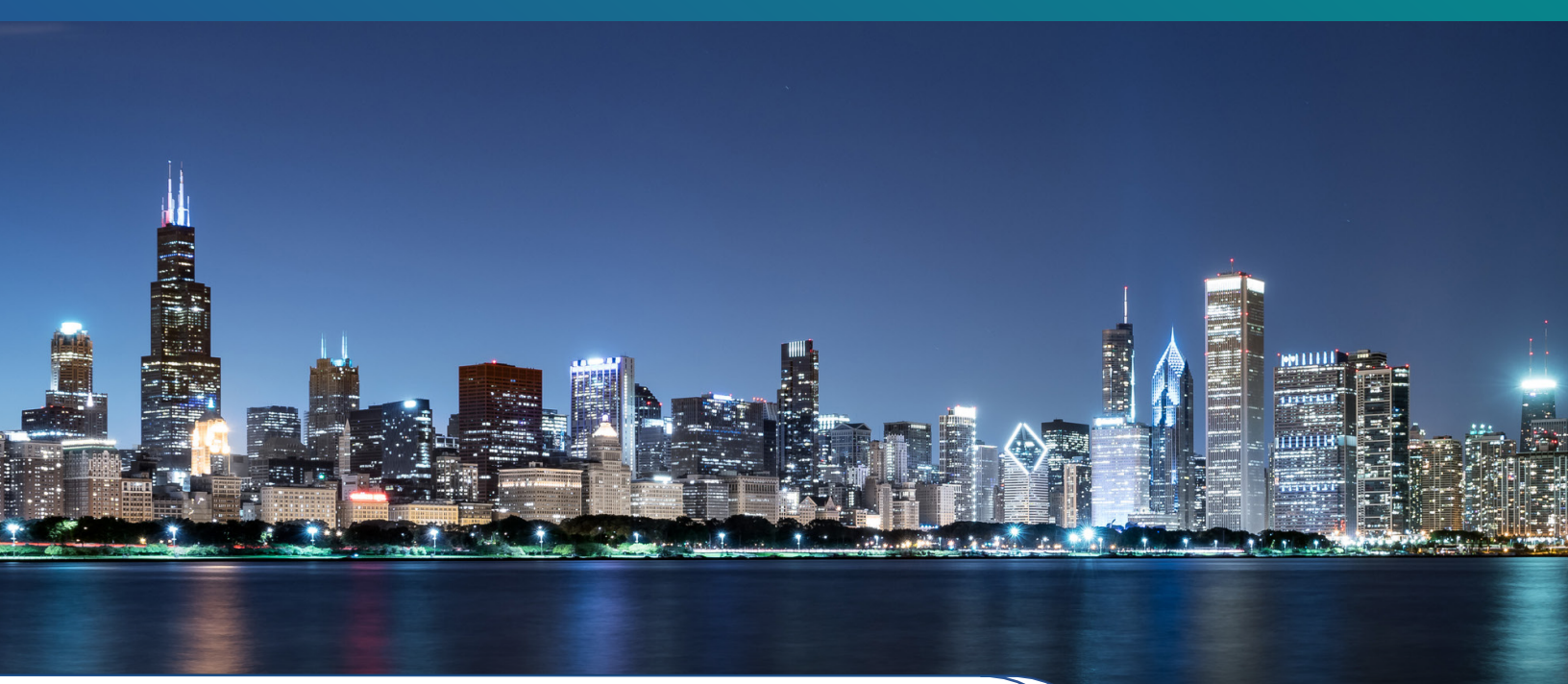

WHITE PAPER

Understanding Latency and Its Impact on Trading Profitability

How to Accurately Measure Latency in Your Global Trading Network

The New World of Electronic Trading

Benjamin Franklin's *Advice to a Young Tradesman*, written in 1748, is perhaps best known for coining the phrase "time is money." A quarter millennium later, this quote fits the modern world of algorithmic trading perfectly. The speed by which an algorithmic trading application, used by a sell-side institution or hedge fund, can access market information, place an order, and have that order filled is of paramount importance to achieving long-term profitability and maintaining competitive advantage. In highly fluid markets, however, raw speed is not enough. Accuracy is also critical. The speed and sequence with which market participants place and execute orders at the matching engine of a given exchange venue depend on many technical variables. This paper outlines considerations to make when deploying such trading platforms and monitoring their performance.

What does it mean to be fair in trading?

Buying and selling decisions happen at increasingly blazing speeds. Stock exchanges such as Nasdaq, Intercontinental Exchange (ICE), and the London Stock Exchange are under pressure to manage growing transaction volumes while ensuring fairness in order execution. They must accurately order the millions of

What does it mean to be fair? First venues should do what they are supposed to do. Two equal orders, placed in succession, should be executed in that sequence. Two market data feeds deemed equal should deliver the same information at the same time, within stated tolerances.

KEYSIGHT TECHNOLOGIES

stock trades on their networks every second. Recent volatility in stock markets worldwide, combined with increasing network complexity, exacerbates the issue of fairness. It allows market participants such as quant traders to gain an advantage over slower firms. In his book Flash Boys, Michael Lewis quotes financial services executive Brad Katsuyama as saying, “That’s when I realized the markets are rigged. And I knew it had to do with the technology.” That quote illustrates the need for some sort of impartiality.

The question to then ask is this: what does it mean to be fair? First, venues should do what they are supposed to do. Two equal orders, placed in succession, should be executed in that sequence. Two market data feeds that the venue deems equal should deliver the same information at the same time, within stated tolerances. Note that switches and gateways are not deterministic, and most gateway network cards and stacks do not deliver packet to software in order.

Consider the network architecture of a typical exchange. Participants are colocated with a top-of-rack, cross-connect, Level 3 network with Border Gateway Protocol (BGP) delivering unicast traffic through multiple gateways. Sparse multicast User Datagram Protocol (UDP) delivers market data. Multiple gateways and switches in this architecture receive and process an uneven load at any given time. If a single switch receives two packets of varying sizes, each packet processes through the fabric at a variable time. Figure 1 illustrates how packets traverse networks.

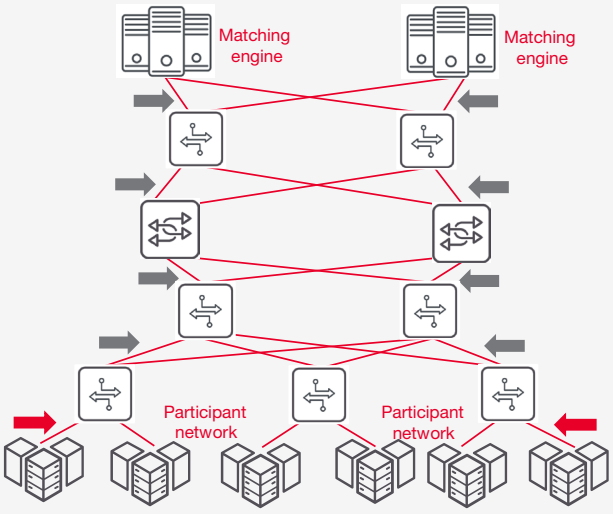

Figure 1. Typical exchange architecture

Typical traffic conditions necessitate contention queuing between traders across related or separate participant networks. Venues implement telemetry to measure between the aggregation switch and participant network cross-connects, denoted by red arrows in Figure 1. (However, empirical tests and resulting data have proven other places in the network valuable to know about, denoted by gray arrows.)

Understanding the latency footprint of your global network infrastructure

Figure 2 examines a latency footprint recently published by ICE. It shows latencies between ICE Global Network consolidated feed ticker plants across major sites.

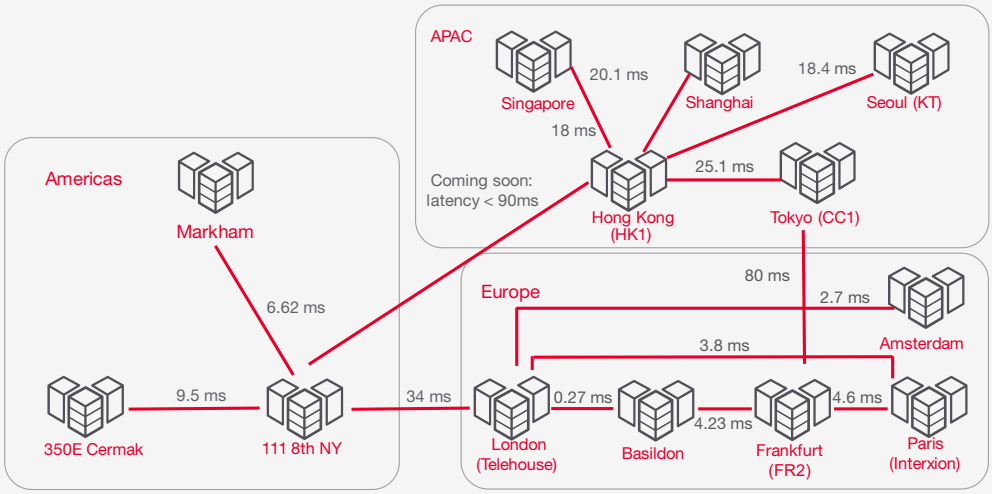

Figure 2. ICE global network latencies (source: Intercontinental Exchange, see https://www.theice.com/market-data/connectivity-and-feeds/network-topology-map)

Note that latency and contributing factors to quality vary considerably depending on the transport mechanisms employed between network sites, where most hosts fall into the wireless, low-latency, or standard category.

Is it enough to measure just latency?

Achieving low-latency distribution of market data depends on many variables, including the exchange, data provider, market participant network infrastructures, and implementation decisions. No matter the execution rate, if market data is delayed relative to competing, traders will not achieve their expected fill ratios. Irrespective of the network in question, latency is always directly dependent on available bandwidth and traffic load. That is, the bandwidth needed to deliver a specific latency objective depends directly on traffic load, irrespective of microbursts.

Another factor in qualifying the bandwidth requirement for a given latency expectation is the latency compliance objective: Is the provisioned bandwidth to deliver the latency objective for 99.99% of packets or higher? What about packet loss? It is not unusual for network service providers to use switches and routers with shallow buffers to minimize latencies.

Additionally, when network elements process a 1,500-byte packet versus a 200-byte packet, the dynamic buffer memory allocated to store and forward that packet has an increased impact on latency. This affects not only the larger packet but also subsequent packets. If sufficient buffer memory is unavailable, the network will experience increased packet loss during microbursts.

Therefore, regarding the bandwidth question, it is equally important to consider both the loss objective and the latency objective.

Observations from network architects and operations teams

While working with exchanges and market participants, Keysight Technologies met with IT teams from network infrastructure and operations. When discussing market data performance analytics, these teams articulated the clear need for near-instant identification of the root cause of data errors, as well as the contributing factors. Is the reported sequence gap on a particular data feed linked to the exchange, the data distribution network, or the sell-side network where the feed handlers are? What went wrong and why? Where should these teams take measurements in the global network infrastructure, and how?

Network architects, or anyone responsible for network operations, should consider the following to have continuous insight:

- Measure one-way latency from the exchange to the network edge.
- Measure latency between consolidated feed ticker plants and shared colocation sites.
- Measure latency and jitter throughout the network at all critical interface points (that is, aggregation switches, feed handler ingress and egress, participant network handoff).
- Understand the impact on latency as a function of network bandwidth and traffic load for a given network segment, transport path, and data feed in question.
- Determine the impact of time synchronization and timing accuracy on latency measurements.
- Complete all of the above with consideration of operational costs and impact on profitability.

Measuring One-Way Feed Latency Using Application Time Stamps

You might need to measure the one-way latency of a market data feed packet published by the exchange venue to a remote colocation site that consumes the feed. To do so, you must calculate the time difference in nanoseconds between the embedded application time stamp (send time) and the time of arrival at the market data analytics device performing this measurement. For this measurement to meet timing accuracy requirements, you need to synchronize the analytics device to an accurate GPS-connected time source. Time synchronization and accuracy are essential; see below for best practices and requirements.

Note that data feeds from different exchanges have different formats for date and time conventions. For example, the Exchange Data Publisher (XDP) feed from ICE date and

time encoding uses Coordinated Universal Time (UTC) Epoch format. XDP uses the concept of a time reference that identifies the whole number of seconds in UTC, and each data message contains the nanosecond offset from that time reference value. The Cboe US Equities / Options Multicast Depth of Book (PITCH) specification, on the other hand, uses the time offset field embedded in various market data messages to indicate the nanosecond offset from the last unit time stamp. Figure 3 illustrates how certain market data can vary depending on the exchange where the information originated.

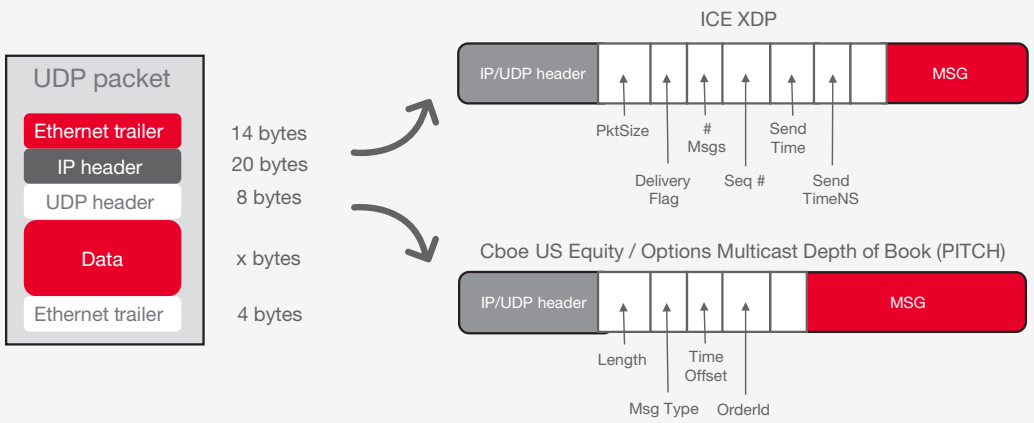

Figure 3. Exchanges use different formats for market data time stamps

Trade performance analytics solutions deployed for latency measurement purposes need to support market data decoders for all exchange feeds that are analyzed onsite. It is also essential for all such analytics to happen in real time, irrespective of the number of feeds and channels in use simultaneously. Regardless of how each respective team wishes to consume latency and latency variation (i.e. jitter) measurements, the trade analytics solution should support computation of at least average and maximum latencies for a given channel over a defined time interval. The solution should allow the user to visualize this data in time-series charts, natively or remotely, via a third- party security information and event management (SIEM) tool streaming that streams meta data generation in real time. Operations and network architecture teams often require tools that alert when the latency or jitter observed for a given feed observed surpassed surpasses the acceptable threshold. These alerts should be easily consumable by existing analytics and SIEM tools native to the network environment where they are deployed.

Using TradeVision for Advanced Transaction Analytics

Keysight's TradeVision is an advanced network visibility solution that serves as a market data analytics tool and a network packet broker. TradeVision, shown in Figure 4, allows IT teams to bring their market feed monitoring infrastructure together with network visibility management while allowing access to preprogrammed support for hundreds of trading venues. Easy to deploy, TradeVision detects sequence gaps and microbursts from more than 4,096 multicast channels in real time. It simultaneously monitors venue feed connectivity health continuously and provides visual dashboards and statistics with time stamping, accurate to the sub-microsecond. Most importantly, TradeVision supports advanced latency analytics, allowing the user to measure one-way application latency and jitter between the exchange venue and remote colocation sites throughout the network at various points of interest, as well as directly between colocation sites.

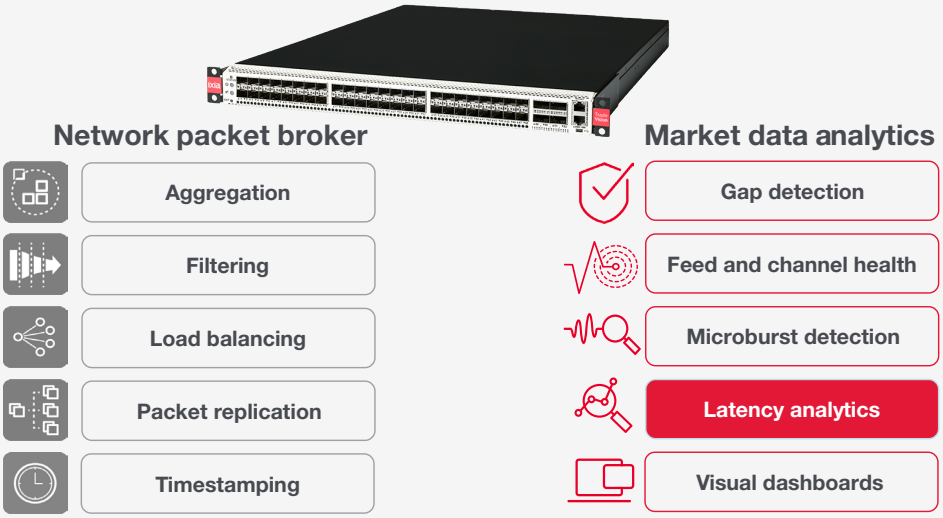

Figure 4. TradeVision for transaction analytics

Colocation deployment options for OWL and internetwork latency measurements

Where are the best sites to deploy TradeVision in the colocation environment for measuring one-way application latency from the exchange and between two points in your network infrastructure? Many TradeVision customers that are market participants consuming market data feeds deploy TradeVision at their core data center locations in major financial hubs. Those hubs include exchange cross-connects in Mahwah, NY4, Carteret, Aurora, Cermak, and Basildon.

The network diagram in Figure 5 illustrates the TradeVision deployment architecture. Each TradeVision appliance interconnects to aggregation switches in the local network

infrastructure via 40 Gb/s links, with two ports connected to national exchange market data feeds and two ports connected to local exchanges. Many of these colocations also use several 10G tool ports for feeding downstream capture devices. The remaining 10G ports on TradeVision patch data traffic through to other switches, such as aggregation, top-of-rack, and explicit congestion notification (ECN) solutions. One-way latency (OWL) measurement takes place at the individual channel level, with 4,096unique channels supported simultaneously per appliance.

TradeVision takes the average, minimum, and maximum latency and jitter measurements in real time. It allows the user to create visual dashboards and threshold-based alerts. If IT teams need remediation to determine root cause analysis, they must to be able to correlate excessive latency with other possible quality metrics such as sequence gaps and microbursts for the feed(s) in question.

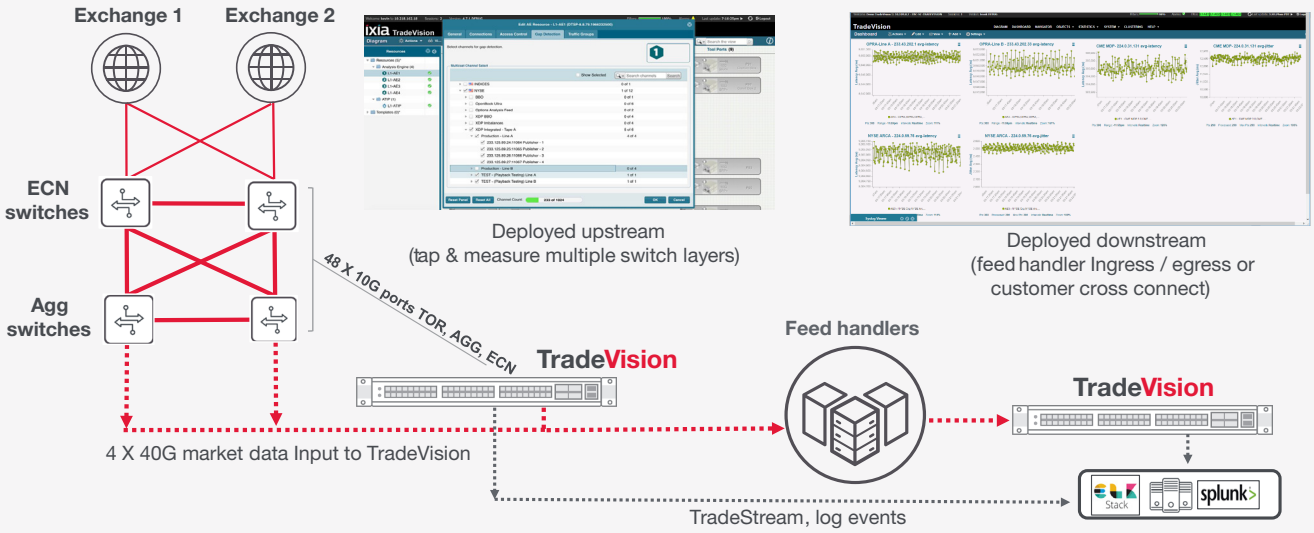

Figure 5. TradeVision deployment

These capabilities give network architecture and operations teams significant flexibility and advanced telemetry in comparing A and B feed latencies or jitter measurements from the same exchange across network segments and transport paths.

The impact of a network's infrastructure and its elements on fixed and variable delay is another dimension of latency. Suppose latency measurements from the exchange to the network edge are within expected thresholds, but the trading application team is complaining about CTA's CQS / CTS price freeze. Is the issue then linked to packet drops caused by microbursts that result from excessive queueing because of oversubscription on a specific network segment on a store-and-forward switch? What type of jitter do you see on a given market data feed between ECN switches at the exchange cross-connect and the customer network handoff? Who is the managed services provider?

Determine two-point latency (TPL) by computing the time delta between nanosecond time stamps at two points in the network (usually between the upstream time stamp inserted at a production switch or upstream TradeVision appliance) and a downstream TradeVision Precision Time Protocol (PTP)-synchronized hardware analysis engine. Both systems are deployed on the same data center in different locations and can answer the questions posed above. As with OWL, TPL allows the user to create flexible time-series graphs via real-time dashboards, showing minimum, maximum, or average latency or jitter measurements over a defined time interval. Threshold-based alert generation via syslog, TradeStream, and SNMP is supported as well. Figure 6 illustrates a software interface in the TradeVision system.

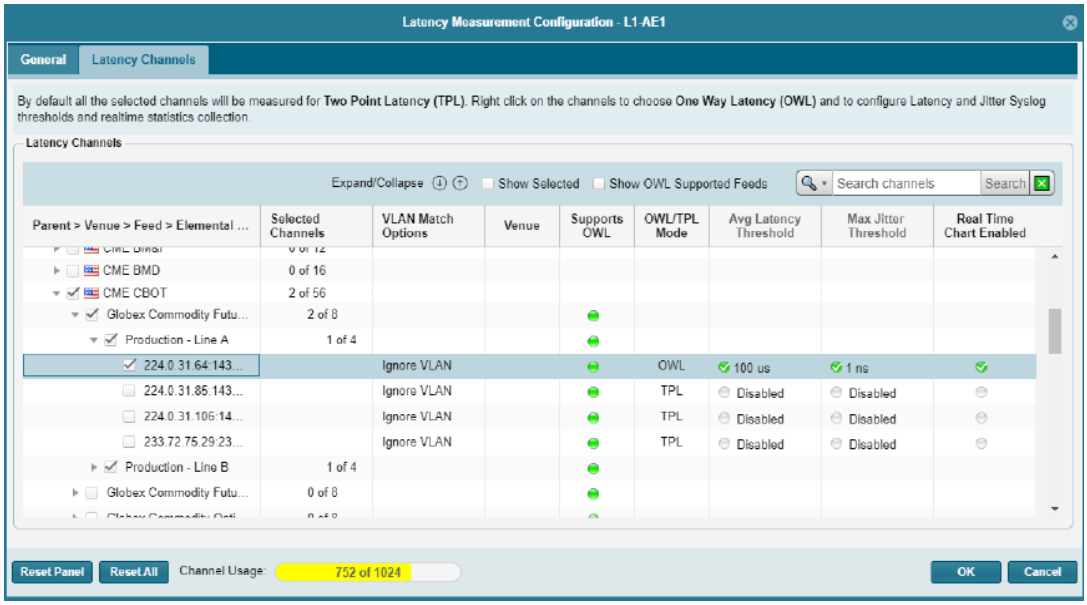

Figure 6. Configuring OWL and TPL measurements

Equally important to the depth of transaction analytics and market data latency monitoring capabilities is the system's ease of use. User interaction workflow on TradeVision enables a new user with no prior solution knowledge to configure data feeds to measure on-way latencies, set up threshold-based alert notifications, and create visual charts displaying maximum, average, or minimum values over time. In addition to creating these real-time charts of various latency measurements, users can drill down from any triggered event of interest, shown as a point on a chart, to the associated event log metadata for further analysis and problem identification. Figure 7 displays some charting options.

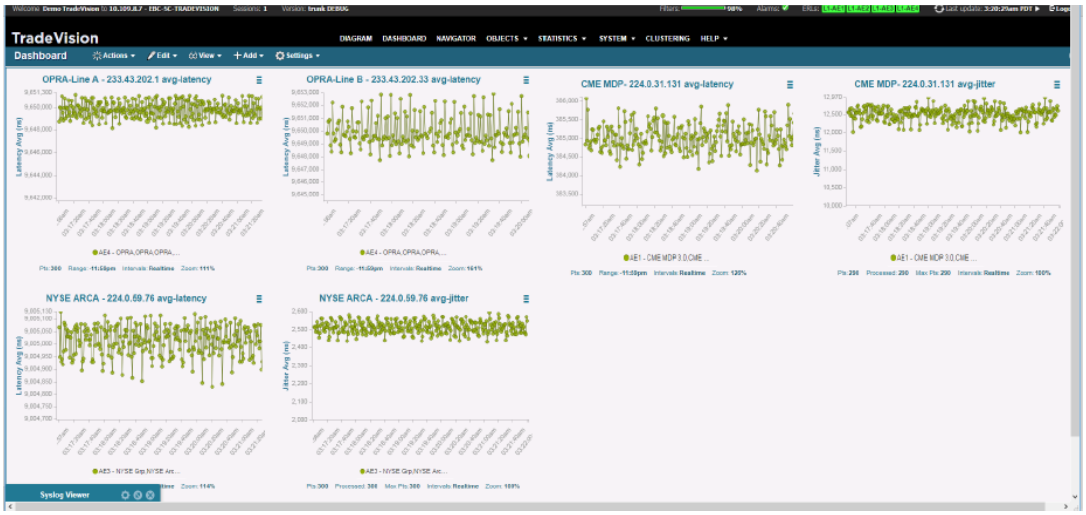

Figure 7. Visualizing latency and jitter measurements

Measuring (WAN) latencies between your colocation sites

Up to this point, we have discussed only OWL measurement from the exchange to the network edge and throughout the network environment. An equally important element to understanding the global latency footprint is by measuring it between consolidated feed ticker plants and shared colocation sites. Suppose the trading applications and analytics tools are deployed in the NY4 or Carteret sites, and the user is consuming MDP 3.0 market data from the Chicago Mercantile Exchange. The user would want the ability to continuously measure latency and jitter for all feeds and channels between these sites, across all transport paths: standard, low latency, and wireless.

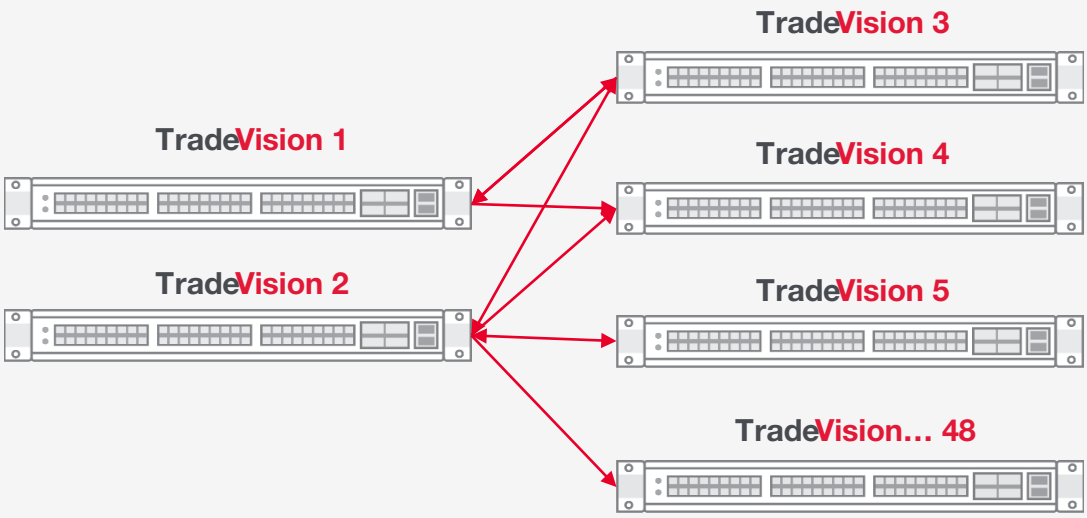

Figure 8. Synthetic mesh network

TradeVision extends advanced latency analytics to accomplish continuous measurements by generating synthetic UDP packets sent between source and destination “mesh pairs” deployed throughout global colocation sites, as shown in Figure 8. Each TradeVision appliance in such a deployment configuration generates synthetic packets with a unique nanosecond resolution time stamp inserted as a 15-byte trailer before the packet’s frame check sequence. A single source appliance may be configured to send to as many as 48 destinations, leveraging the existing Layer 3 transport network. Since both devices in a given synthetic latency mesh pair sync via PTP to a GPS-connected grandmaster, you can make latency and jitter measurements on the destination TradeVision with the same visual dashboards employed for continuous network operations.

Figure 9 illustrates how the TradeVision source appliance connects to respective sites.

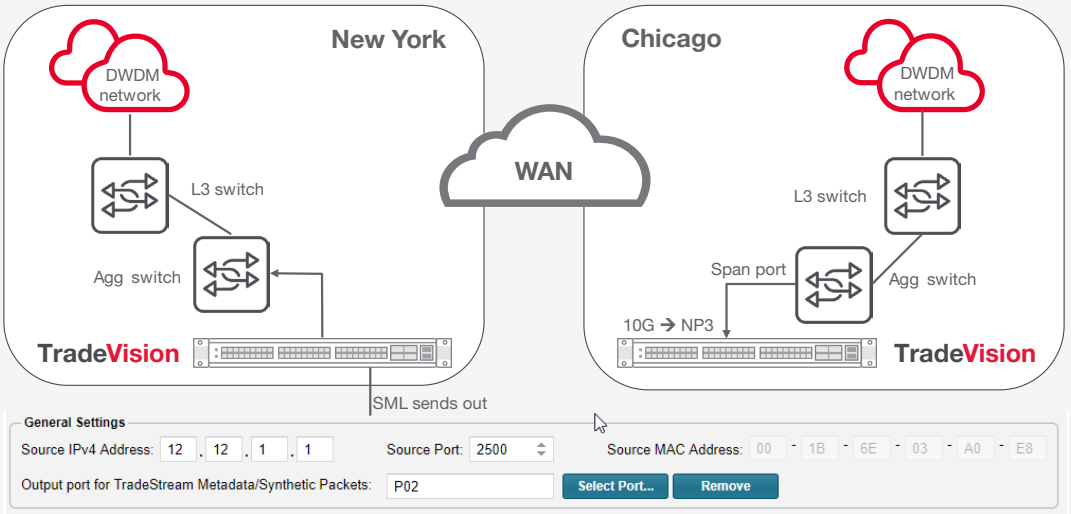

Figure 9. Synthetic mesh pair deployment

Exchange and data provider recommendations — latency vs. bandwidth

An earlier discussion asked about bandwidth for low latency. It established an association between latency objective and the bandwidth required for a given percentage of data traffic when provisioning a circuit for market data connectivity to an exchange. The challenge of appropriately sizing bandwidth for low-latency market data feeds is irrespective of the type of connectivity to the exchange. It could be via IPSec VPN, by colocation cross-connect, through a financial extranet provider, or directly connected via Ethernet Private Line.

That said, based on our experience working with high-frequency market participants, exchanges and market data aggregators provide their own best practices, which include connectivity and bandwidth recommendations referenced above. Cboe, for example, publishes and regularly updates its recommendations in its “US Equities / Options

Connectivity Manual." Section 4.1 of this document emphasizes some of the challenges involved in circuit provisioning. It furnishes useful empirical data for market data consumers to advise on minimum bandwidth requirements for the Multicast PITCH, Gig-Shaped, and WAN-Shaped feeds. Figure 10 provides an excerpt from the manual showing bandwidth statistics and messages-per-second peaks for historical highs for 1-, 5-, 10-, 30-, and 60-second intervals. The figure also includes 1- and 10-millisecond interval peaks.

Cboe market	Interval seconds	Multicast PITCH		TOP		TCP PITCH	
		MPS	Mb/s	MPS	Mb/s	MPS	Mb/s
BZX Exchange	.001	17,421,000	1,951	1,098,000	322	1,953,000	907
	.010	13,699,800	1,546	757,400	208	1,182,300	842
	1	500,539	173	168,714	45	363,976	163
	5	251,988	87	85,237	22	249,388	116
	10	178,283	62	79,303	21	229,095	107
	30	159,833	55	70,997	19	207,030	96
	60	140,200	48	64,124	17	180,113	84

Figure 10. Example illustration of bandwidth recommendations for Cboe market data feeds. Excerpt from the "Cboe US Equity / Options Connectivity Manual" (version 10.1.0).

Note that, as expected, the rates increase as the measurement interval decreases. The 1-millisecond rate for Multicast PITCH feed is 1.95 Gb/s versus 48 Mb/s for the 60-second peak rate. Also, as the manual specifies, buffers in the end-to-end path strongly determine the extent to which the network connection to the consumer of the market data feed will handle microbursts exceeding the available bandwidth without packet loss.

During spikes in quote updates, market participants using less-than-sufficient bandwidth will experience queuing of their market data. Those consumers of data feeds using the same bandwidth to receive quotes and transmit orders may expect slightly delayed orders if the bandwidth is insufficient. Many companies will find delays unacceptable and should provision bandwidth to reduce delays.

TradeVision's market data analytics include real-time bandwidth utilization measurements during microbursts for any top-level feed, channel, or IP, in addition to advanced latency. Data density measurements take place at 15.2 µsec sample interval to calculate the average burst over 1 second and generate events when the average burst exceeds user-specified thresholds. Such a low-level sample rate / granularity in microburst

measurements ensures that specific service-level agreement (SLA) objectives (that is, 99.99% of packets) meet latency SLA objectives, even during the smallest microburst time intervals, thus accounting for more realistic circuit bandwidth provisioning requirements. Figures 11 and 12 illustrate latency measurements at single-digit nanosecond resolution for one of the production A channels on the CTA feed, and aggregate bandwidth over the same period for all of the CTA channels, with 15.2 us measurement sample rate, averaged over a second.

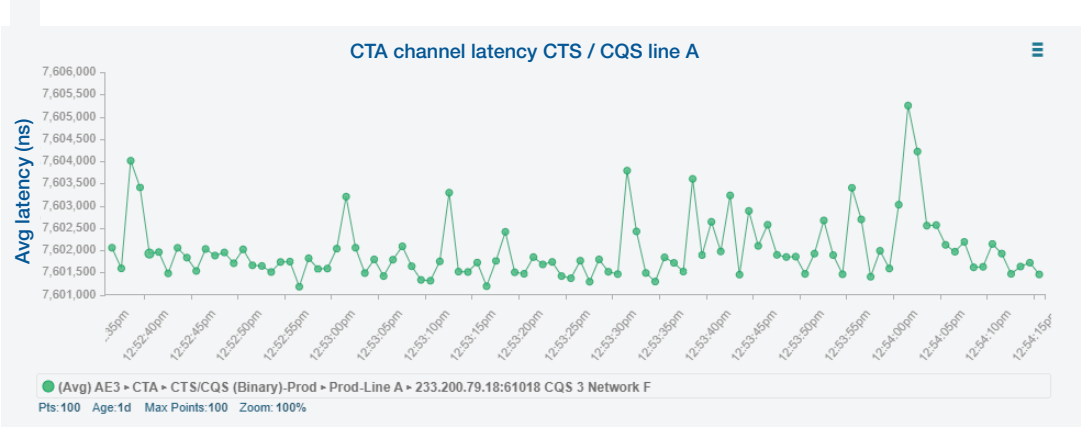

Figure 11. Consolidated Quote System (CQS) 3 CTA production line A channel latency

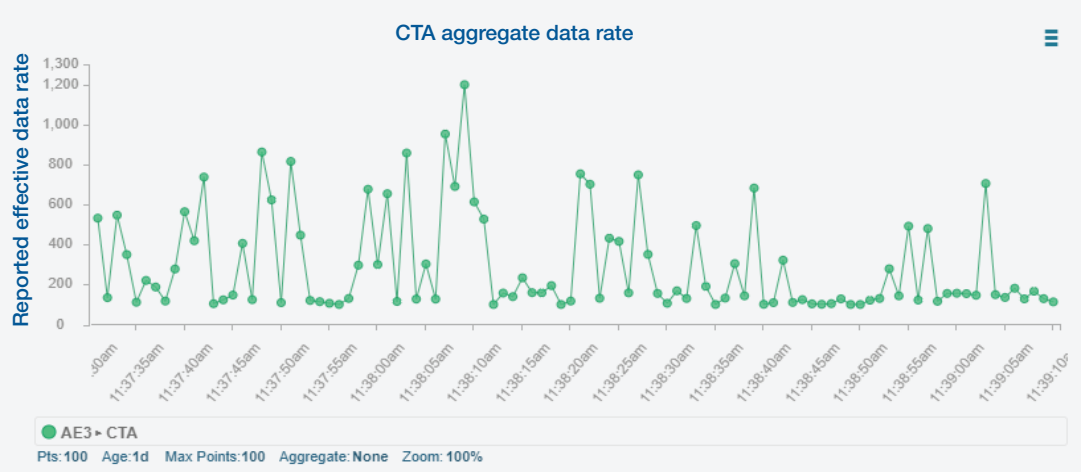

Figure 12. Consolidated Tape Association (CTA aggregate feed bandwidth with 15.2 us measurement rate

Summary and Conclusion

Before acting, consider the following recommendations for bandwidth connections for low-latency market data, deploying market data analytics tools across the colocation infrastructure, and aims for benchmark latencies throughout the global network infrastructure.

1. When provisioning market data circuits for a given connectivity type to an exchange, specify latency and loss objectives in conjunction with bandwidth requirements during peak utilization periods.
2. Monitor continuously, as market volatility conditions change rapidly. Volatility impacts average and peak message rates, causing higher-than-specified or -expected levels of latency and packet loss because of microbursts. Having monitoring infrastructure in place with trigger-based alert notifications is essential for proactive issue identification and problem resolution.
3. Understand and continuously verify the global latency footprint from the exchange to remote colocation sites throughout network infrastructure, and between data centers or colocation sites that consume real-time market data feeds.
4. Ensure that your transaction analytics tools provide detailed visibility into crucial quality events. Ensure that enough empirical data is available to conduct root cause analysis when needed — that is, associating excessive latency with possible other quality metrics such as sequence gaps and microbursts for feeds in question

Appendix: Which Latency Measurement to Use Where

This white paper has covered various challenges involved in measuring latency throughout the global network infrastructure. The following two tables summarize latency objectives, associate them with one of the three latency measurement options offered, and provide additional configuration insight for each.

Table 1. TradeVision latency analytics based on objective

Objective	One-way latency	Two-point latency	Synthetic mesh latency
One-way latency between exchange and a remote site	✓		
Latency between two points within the network		✓	
Inter-data center / colocation latency over WAN			✓

Table 2. Configuration details for OWL, TPL, and SML latency methodologies

Configuration	One-way latency	Two-point latency	Synthetic mesh latency
Support for external switch time stamps		✓	
Min # of TradeVisions needed	1	1 (when using external time stamps)	2
Maximum supported	N	2	48
Latency and jitter	✓	✓	✓

Learn more at: www.keysight.com

For more information on Keysight Technologies' products, applications or services, please contact your local Keysight office. The complete list is available at: www.keysight.com/find/contactus

KEYSIGHT TECHNOLOGIES