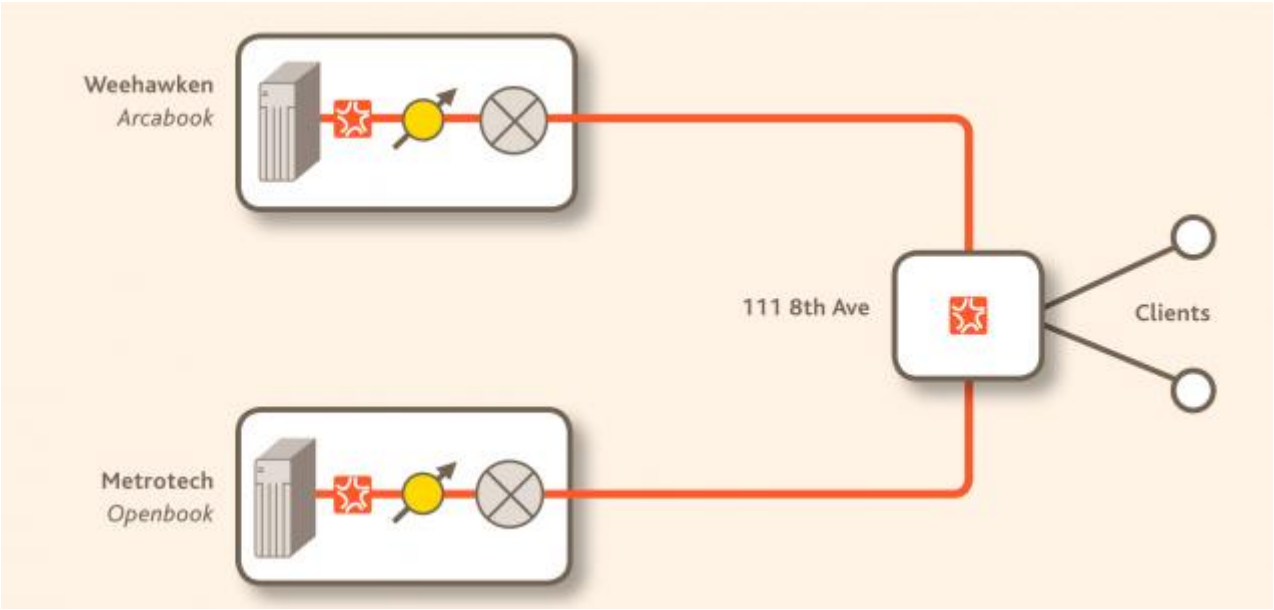
Published on *Latency Stats* (http://www.latencystats.com)
NYSE Technologies
Corvil

Latency Transparency for Market Data

Achieving low-latency distribution of market data has become one of the most important factors that determine success in high frequency trading. No matter how fast you can execute, if your market data is delayed relative to competing traders, you will not achieve the expected fill rates.

In spite of its importance, key questions about market data latency have traditionally been hard to answer. How long does it take for market updates to reach your systems? How much capacity will you need to process every update without delay, even when the feed is busiest? What impact do the various connectivity options available have on latency?

LatencyStats.com is a new initiative from NYSE Technologies and Corvil aimed at answering these questions and more, by providing full visibility into the performance and characteristics of NYSE market data feeds. Our goal is to ensure that traders feel confident they understand the nature and quality of the data they are receiving. We believe that greater transparency into market data performance will help traders refine their deployment strategies, and their investments in equipment, bandwidth and co-location services.



Instrumented sites providing measurements for LatencyStats.com.

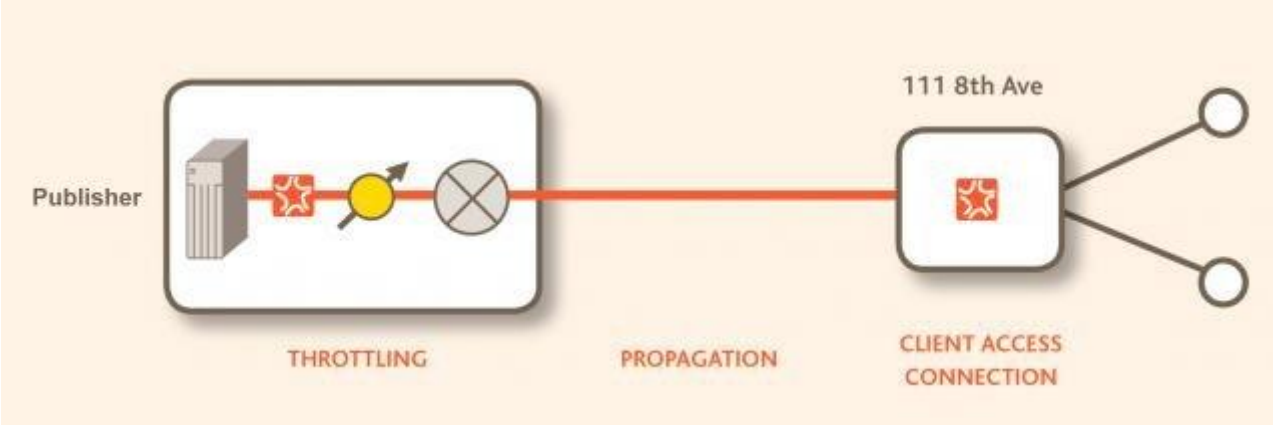
At launch LatencyStats.com will focus on the ArcaBook™ and OpenBook™ data feeds. These are two of the most popular feeds for NYSE equity market data. The presented performance stats are based on Corvil instrumentation installed at the SIAC data centers where the feeds are published, and at a Secure Financial Transactions Infrastructure © (SFTI©) access point where many trading clients connect to receive the feeds. The website provides a live view of current market data activity that is updated in real time, as well as historical views that will show how activity evolves over time -- and what impact this evolution has on latency.

To get the most from LatencyStats.com it will be helpful to understand the various possible causes of market data latency, and how these relate to the information presented in the site. With this in mind let's review what happens to a feed between the point where it is published by the exchange, and the point where it is ultimately consumed by a trading client.

Sources of Latency

Many trading algos still make the basic assumption that the market data they use is an instantaneous real-time view of the state of the market. In reality, all market data experiences some unavoidable latency before it is delivered to the end user. At a minimum there is propagation delay based on distance from the matching engines. Information in today's networks travels very fast (at up to 75% of light-speed depending on the type of cabling used) -- nevertheless even at these speeds, propagation over metropolitan area distances can still add hundreds of microseconds of latency (and more if the path taken is not direct).

A second potential source of latency is processing delay in the various active components in the market data path, such as switches, routers, firewalls and other elements. These elements inevitably take some time to process each message they forward, and queues can build up during periods when the data rate temporarily exceeds processing capacity. Backlog periods of tens or hundreds of microseconds are too short to notice using conventional instrumentation, but these timescales are significant in today's trading environment. If backlogs become excessive then buffers can fill up, leading to data loss and kicking off lengthy retransmission cycles.



A feed may experience latency on its way to trading clients for several different reasons.

Queuing latency is most likely to be experienced at speed mismatch points – for example at the point where a client connects to the exchange network. The extent and frequency of delays at this point will be determined by the amount of bandwidth available to carry the data onwards to the client's systems. Sizing this connection adequately is therefore an important consideration for latency sensitive traders.

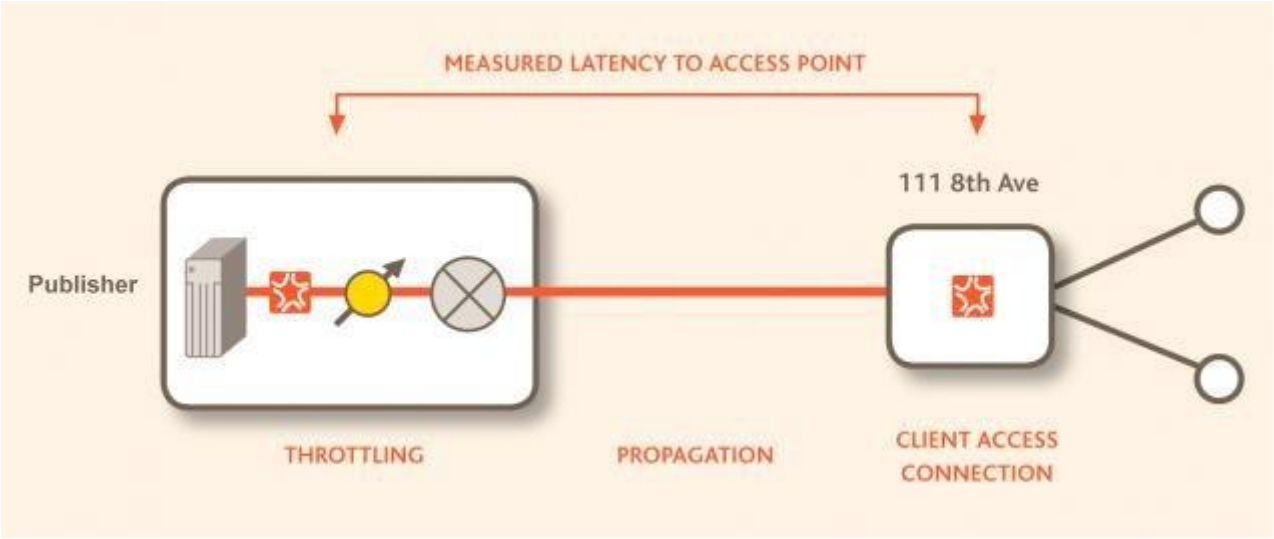
To reduce the likelihood of congestion-related effects, market data providers normally apply some 'throttling' to the feed close to its source.

This limits the peak feed rate to a level that is less likely to overwhelm client connections or the network elements along the way. Throttling can add latency if the throttling system is forced to occasionally reduce the volume of data sent, requiring some data to be re-transmitted.

Finally, a trader's own systems can add latency to market data if they cannot keep up with the pace of the feed. Market data activity varies over many timescales from microseconds upwards, due to both machine- and human-related factors. To avoid delaying the feed, elements along its path must be able to keep pace with short spikes of activity – microbursts – that may last fractions of a second.

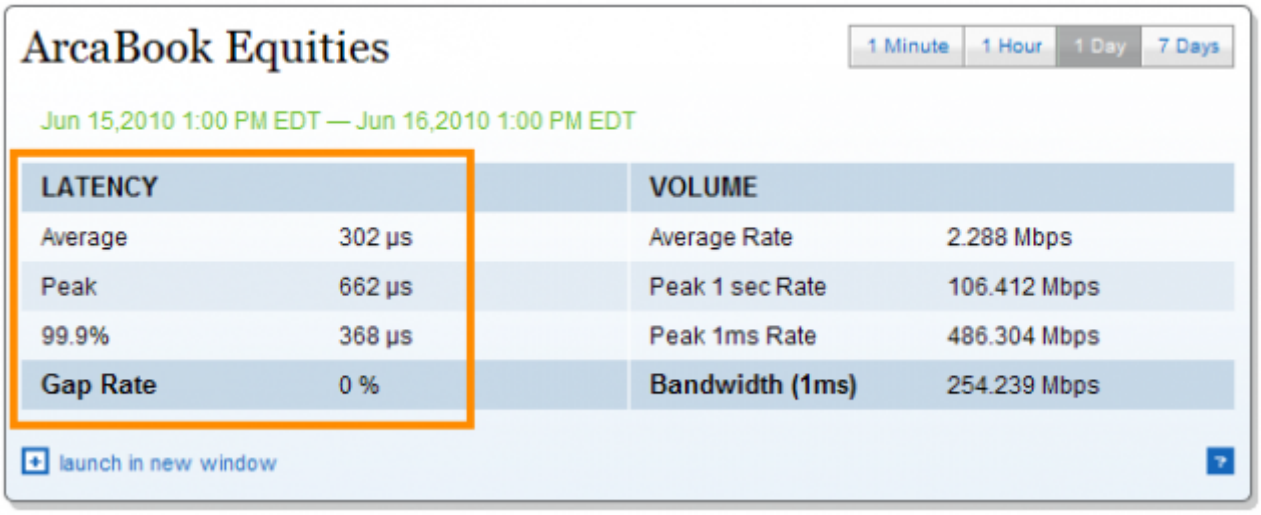
From Feed Publication to the Client Access Point

Our goal in establishing LatencyStats.com is to give users visibility into all of the possible sources of latency along the path from the feed publishers to the client access connection. Firstly, the site uses Corvil's measurement technology to accurately determine latency from the feed publication point to the access point where clients receive the feed. Latency values are measured with microsecond precision for every individual market data packet and stored for later presentation via the website. The measurements include the latency across the client CPE switch (assuming a 10G connection).



LatencyStats.com measures the latency of the feed from the point where it is published to the client access point.

To let users see how latency changes over time, we present results for a set of time-frames ranging from the most recent minute to the last 7 days. You can select a time-frame to view using the tabs at the top of the panel for each feed. When the one-minute time-frame is selected, for example, the statistics shown are computed by analyzing the measurements for all the packets within the most recent minute. Because latency can vary appreciably even at very short timescales, the panel presents three different statistics giving the average, peak or maximum, and 99th percentile of the observed latency values.

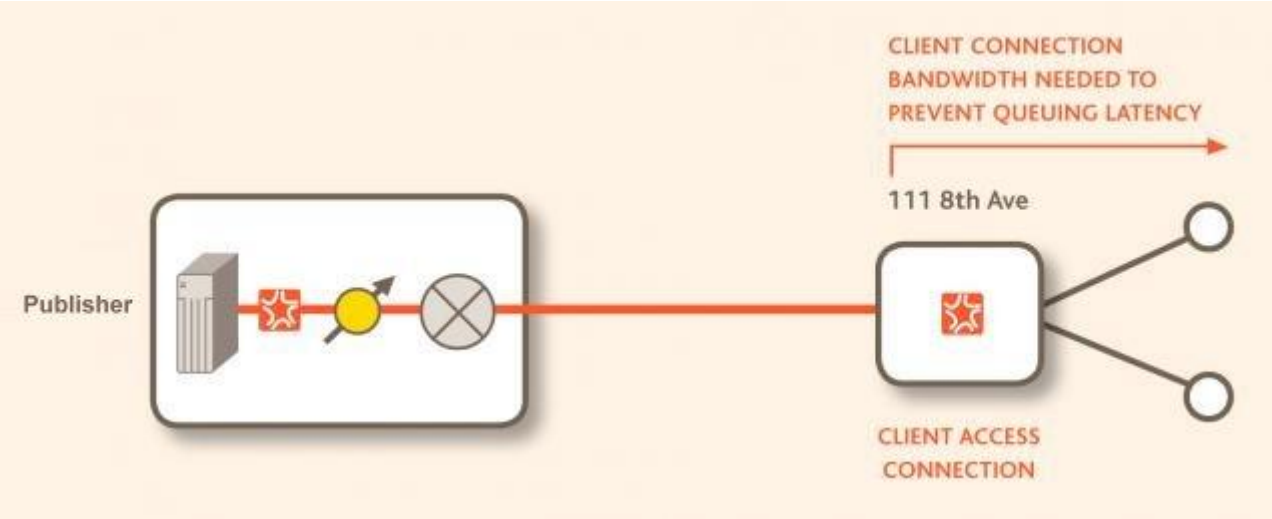


The panel for each feed shows the average, peak, and 99th percentile latencies measured for every market data packet over the time-frame selected at the top of the panel. Also shown is the percentage of market data messages detected as missing (if any) at the point where the feed is delivered to the client connection.

Also shown in each feed panel is the rate of sequence gaps seen in the feed at the access point where it is delivered to clients. Sequence gaps may occur if market data messages are inadvertently dropped somewhere along the path from the feed publication point and are consequently not available for delivery. To actually receive these messages, clients would have to identify and re-request the missing sequence numbers, a process which adds extra latency. The gap rate figure presented in the panel gives the percentage of messages that we have identified as missing. Together with the latency measurements from the publication point, these figures provide a complete picture of the quality of the feed as it is handed over to the client connection.

Latency Across the Client Connection

The client connection to the access point is a potential source of variable queuing latency due to speed mismatch. The faster the connection is, the lower the maximum latency will be; but of course faster connections are also more expensive. The best way to determine latency across the client connection is to install instrumentation at the client side to timestamp the arriving market data messages, and correlate the timestamps against corresponding measurements on the exchange side of the link. In the absence of client-side measurement, LatencyStats.com takes the next best approach: we analyze the market data traffic heading onto the client connection to determine how fast the connection needs to be to prevent queuing latency from exceeding a specified maximum value.

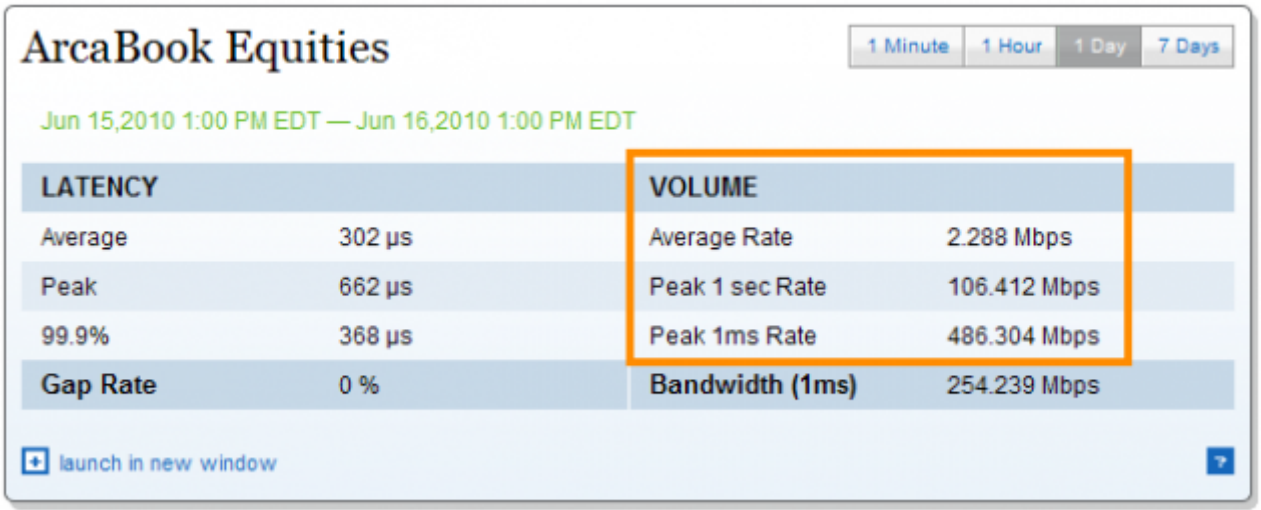


The feed Bandwidth figure presented on LatencyStats.com is the minimum client connection speed needed to ensure that queuing latency never exceeds a specified target value.

The results of this analysis are presented in the Bandwidth section of the panel for each feed, which shows the connection speed needed to ensure that queuing latency will never exceed 1 millisecond. This value is computed under the assumption that all of the feed (every multicast group) is received by the client. It represents the bandwidth that this individual feed needs to stay within the latency target; clients receiving multiple feeds or other traffic on the same connection may require a higher overall connection speed to achieve the same latency.

Consuming the Feed

For clients trying to size their own internal market data systems, it's helpful to know what the feed activity pattern looks like at short timescales. To help with this task, the Volume section of the LatencyStats feed panel shows the measured feed activity over multiple timescales ranging down to 1 millisecond. In this section of the panel, the peak 1-second and peak 1-millisecond rates show the highest bit-rates measured at these timescales during the period selected by the user at the top of the panel. The average rate figure shows the average bit-rate over the whole of the selected period. By varying the selected period you can get a picture of how both the average rate and burstiness of the feed are trending over time.



The volume figures presented in the feed panel provide information on feed activity over multiple timescales, ranging down to 1 millisecond.

A downstream system such as a feed handler or a firewall that is able to process feed data at a sustained speed exceeding the 1-millisecond peak rate, will add no more than 1 millisecond of latency to the data. This can be a useful, conservative rule of thumb for capacity sizing; the added latency will often be much less than 1 millisecond, provided that the downstream system can indeed process data continuously at this speed without interruption.

New Levels of Market Data Transparency

LatencyStats.com represents a new and more comprehensive level of visibility into market data than users have been able to access in the past. Market data is continuously evolving in time, and usually growing in volume. Managing this changing flow of data within the tight performance requirements of today's traders is a challenge to all parties involved in the production and distribution of the feeds. We hope that LatencyStats.com will help the industry to meet this challenge.

1.jpg [1]
2.jpg [2]
3.jpg [3]
p31.png [4]
4.jpg [5]
p51.png [6]