Measuring Up With Latency

By Sergio Bogazzi and Rachel Dale

For such a small unit of measure (effectively microseconds), market data platform latency has received a large amount of attention in recent years. Couple this with front-office concerns about escalating market data volumes, especially those driven by exploding options data rates, and you have a sharp need to properly measure the performance of your market data platform.

Latency in particular has become the key differentiator when evaluating performance of competing direct market access platforms. These platforms target low-latency demand of time-sensitive trading strategies, such as statistical arbitrage applications, by allowing them to even bypass the securities information processors (i.e., SIAC, NASDAQ) and receive information directly from exchanges or pools of liquidity.

Latency has also become a central component of the SEC's Regulation NMS Order Protection Rule. Rule 611(a)(2) states the following:

"Trade prices should be compared with protected quotations at the time of execution, as reliably documented, to affirm that such quotations were not traded through." - 3.02 SEC FAQ 10/31/06

Clearly the need to accurately measure latency has become a central performance and regulatory requirement for many front office market data systems.

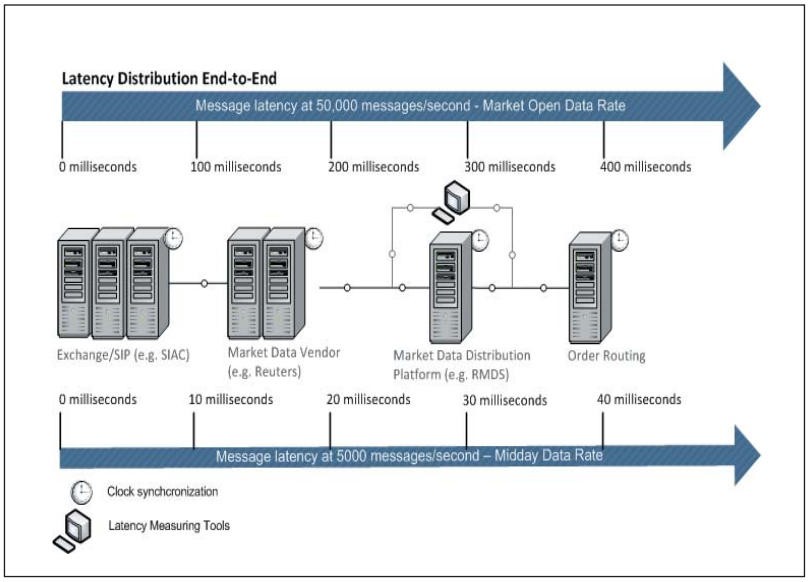


Figure 1 – Market Data Latency Distribution

Latency Defined

Put simply, latency is the amount of time it takes for messages, containing time-sensitive price and size information, to move from the originating source to the ultimate destination as shown in Figure 1. Simple as it may sound, latency measurements may be unfavorably inflated by the measuring and clock synchronization tools adopted. These tools may deplete precious network and CPU resources dedicated to message processing, as well as factoring their own processing time into the latency result; a concept with familiar roots in Heisenberg's uncertainty principle. As stated above, regulatory drivers are now demanding accurate latency measurements. Regulation NMS, in particular, asks the following of broker dealers:

"Adopt reasonable policies and procedures for synchronizing its internal clocks, to the extent that it uses different clocks to assign time-stamps to its order, trade, and quotation data." - 6.02 SEC FAQ 10/31/06

Continued on page 12

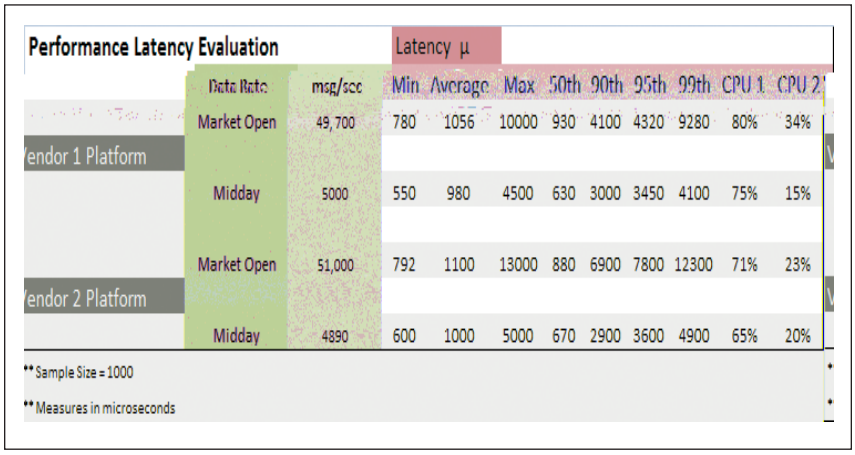
Measuring Up With Latency

Continued from page 10

In order to reach meaningful conclusions regarding the latency measurements your business requires, it is important to understand how message rates, clock synchronization, and the metrics you compute influence the results.

Message Rate

A meaningful latency measure depends on high network traffic. In the market data world, various segments of the trading day exhibit distinctly different message rates (e.g., market open, market close). Establishing an appropriate rate requires a balanced view that factors in the needs of the business as well as the capabilities of the underlying technology solutions. If you are using replay tools to simulate a real-time market data feed, the message rate should reflect the feed's present day maximum peak, future maximum peak (such as those forecasted by the Market Data Capacity working group of the Financial Information Forum), or a rate that reflects your platform's saturation rate.



Performance Latency Evaluation			Latency µ							
	Data Rate	msg/sec	Min	Average	Max	50th	90th	95th	99th	CPU 1	CPU 2
Vendor 1 Platform	Market Open	49,700	780	1056	10000	930	4100	4320	9280	80%	34%
	Midday	5000	550	980	4500	630	3000	3450	4100	75%	15%
Vendor 2 Platform	Market Open	51,000	792	1100	13000	880	6900	7800	12300	71%	23%
	Midday	4890	600	1000	5000	670	2900	3600	4900	65%	20%

** Sample Size = 1000

** Measures in microseconds

Figure 2 – Evaluating latency performance in competing platforms.

Clock Synchronization

Once the proper message rate has been established, you need to ensure the integrity of your timings. Latency measurements require the time a particular message first left its source, followed by the time the message arrived at its destination. If both of these times are collected from the same machine, using the same system clock, you will have favorably eliminated the need for synchronization altogether. If instead your latency measurements depend on the injection of timestamps by more than one machine in the platform, then it is necessary to synchronize the clocks between these machines. The reality of clock synchronization is, at best, accuracy on the order of milliseconds, not microseconds. If your measurements require microsecond accuracy, you should bypass clock synchronization approaches and measure from a single machine.

Computing the Result

After establishing an appropriate message rate and clock synchronization technique, you are now ready to compute your latency measurements. Like any statistical process, it is important to collect an appropriate number of timing samples to ensure a quality result. Minimums, maximums, and averages are useful when estimating the latency result as shown in Figure 2. Additionally, the variances between the 50th, 90th, and 95th percentiles will justify any concerns with spikes, allowing you to take corrective action.

These seemingly small details are extremely important towards ensuring the quality of such a key measurement. Today's performance and regulatory requirements should urge financial firms to uncover the real meaning of their latency measurements and ensure the techniques that produce them will meet the needs of their businesses.

Sergio Bogazzi (bogazzi@jandj.com) is Director – Technology Services and Rachel Dale (dale@jandj.com) is a Senior Software Developer at Jordan & Jordan, 212-422-8567. Jordan & Jordan (www.jandj.com) offers practical, insightful services and technology solutions to address the challenges and opportunities facing the securities industry. Combining domain expertise with a disciplined approach to problem-solving, we serve a global client base of broker-dealers, investment managers, hedge funds, exchanges and financial information and technology vendors.