

Measuring Data Latency in The Ultra-Low Latency Space



An Interview with David Brukman, Vice President, Technology, Interactive Data Real-Time Services.

When it comes to the ultra-low latency space, latency monitoring and measurement become even more important than in the low latency arena. Attaining the best possible latency is one of the highest priorities of those firms looking to achieve ultra-low latency... and they devote considerable resources to obtain it. That makes measurement much more crucial.

Measuring and monitoring latency in the ultra-low latency space presents some particular challenges. For instance, it's critical to compare apples and apples, and determine at what point the measurement clock starts and stops. Some vendors do not measure the entire data processing chain – the period from when the data is received from the exchange until it is accessed by the customer's application. Others test for latency one time in a lab, but not on an ongoing basis under real-time production conditions.

Another challenge is that because the timescale being measured is so small, the measurement must be done with very sensitive tools. David Brukman, Vice President, Technology, Interactive Data Real-Time Services, recently described the company's latency monitoring in detail, and discussed the challenges involved in developing the measurement system.

Interactive Data's DirectPlusSM fully managed, ultra-low latency direct exchange data service has been measuring data latency of no more than 130 microseconds (0.13 milliseconds) and generally with a range of 80 to 130 microseconds

since being rolled out in June.

Q: At a high level, what does your latency monitoring system look like?

David Brukman: We use two different methods in order to be confident in the accuracy of our results. The first is based on diagnostic equipment that is independent of any production software. This system is designed to time-stamp the incoming exchange data at the switches where it's delivered to our network, and again when it's processed and written to storage by a simulated customer application.

The diagnostic tool is independent of the production software in that it measures the latency performance of the production system, without being part of the system. We believe that makes it a more impartial and comprehensive method of measurement, as it takes into account all aspects of latency, including impact of network equipment, interface cards, operating system and device drivers.

The second method is built into our ticker plant. This method is designed to time-stamp the data when the ticker plant receives it, and again when the software is finished processing the data and is sending it out to the customer's application.

Q: And then you compare the results?

DB: Yes, we compare the results of the two to ensure that they're producing

similar numbers. That gives us confidence that our latency measurements are accurate and representative of what the customer experience is. We use the first system once every week; the second is designed to be active continuously – as long as our ticker plant is operational and processing exchange data.

Q: When did you implement these two methods?

DB: In mid-June, when we announced that DirectPlus was available.

Q: Have you seen much variation in data latency over that period?

DB: There is some variation as traffic rates go up and down, but even at times when volumes have surged – as they did in July – our measurements indicate the number hasn't gone beyond the 130 microsecond ceiling.

Q: Some people have talked about data latency measurement as an art as well as a science. Do you see it that way?

DB: To me this is more of a science. But it's important to emphasize that you need to find ways of measuring it that reflect real user experience, and sometimes you need to develop specialized tools because when you're dealing with time segments as small as microseconds, a lot of conventional measurement tools aren't really effective.

Q: What types of specialized tools have you needed to develop?

DB: We have two that address time skew. Let's say you're timing a sprinter. One stop watch is pushed when he starts. Another is pushed when he crosses the finish line. If those two watches are off by even a fraction of a second, the resulting time would be inaccurate. That's time skew. One way to address it, which we use in the first measurement method, is to use one computer that has access to both segments of the network during the measurement period.

The second way we address it is to use ticker plant facilities that are engineered to be very efficient in using high resolution time stamps. When you're measuring tens of thousands of messages per second to the microsecond resolution, the actual action of measuring and re-

coding time stamps may load the ticker plant and increase the latency. So we had to work very carefully to design the high resolution time stamp to be very, very efficient.

Q: What are some of the ways you have designed the DirectPlus ticker plant for ultra-low latency?

DB: We have optimized both hardware and software, and have tried to avoid operations that cause high latency, such as multiple copying of the buffers, which are pieces of computer memory where a message is stored.

We've also designed the ticker plant to minimize system calls. That's a computer operation that requires support from the operating system. Minimizing the number of system calls in data processing is another way to reduce latency.

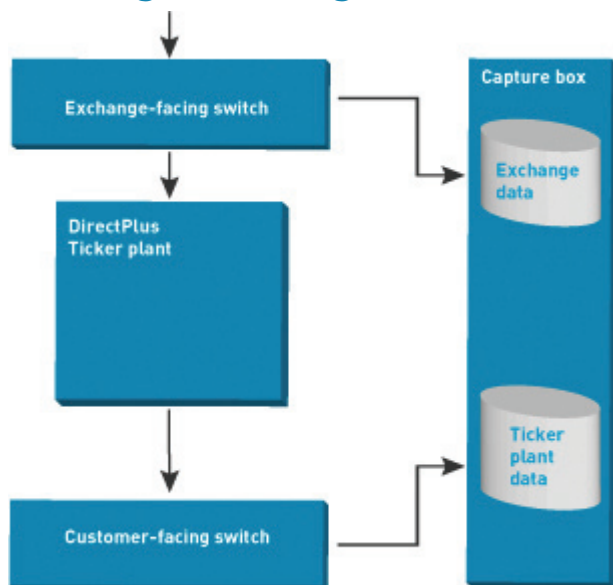
Q: Anything else you'd like to add about the kinds of monitoring tools you've discussed?

DB: Any direct exchange feed solution should have sophisticated monitoring tools designed to ensure that data latency is optimal. As a fully managed service, DirectPlus enables customers to benefit from the work we've done to develop these tools. Of course there are many other benefits of a hosted service like DirectPlus, including the opportunity to share communications costs and off-load ticker plant maintenance.

This article is provided for information purposes only. Nothing herein should be construed as legal or other professional advice or be relied upon as such.

External Latency Calculation (method 1)

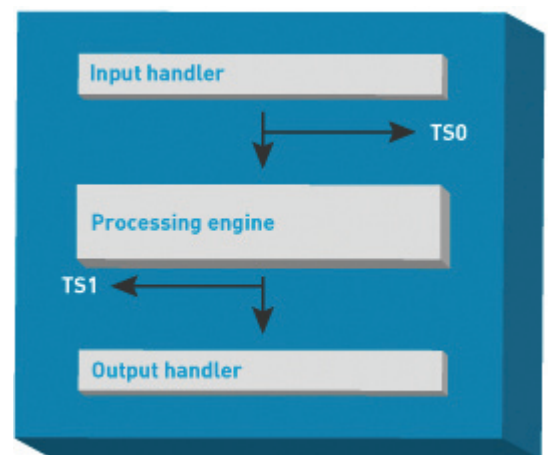
Exchange incoming data



Total latency is measured as the time difference between an exchange-received message and its output from the DirectPlus ticker plant. Both outputs are captured on the same box.

Internal Latency Calculation (method 2)

DirectPlus Ticker Plant



The internal latency calculation is the difference between TS1 and TS0 (TS1-TS0).

This article has been reproduced by permission of A-Team Group, publishers of *Low Latency- Are You Performing*. For further information about Low Latency please visit www.low-latency.com.

A-TEAMGROUP
www.a-teamgroup.com