



By Donald Finucane,
Vice President of Product Management
and OTC Data Services,
Interactive Data Real-Time Services
www.interactivedata.com

Cause and FX: Taking a closer look at the issue of latency

Market data latency has become a white-hot issue in the past couple of years primarily because of the rapid adoption of electronic trading, and in particular algorithmic trading. Prior to electronic trading, when traders used screens to get prices and traded over the phone, it didn't matter so much if the data latency was 30 or 300 milliseconds – the latter representing the time it would take a trader to blink. Now, with computers doing much of the trading, every millisecond counts...

So what exactly is data latency?

In its simplest form data latency refers to the time it takes to get data from point A to point B.

Unfortunately, in the securities industry, there is nothing simple about data latency, particularly when it comes to managing and measuring it. It is difficult to manage because there are several sources of data latency that must be examined individually and as a system, and it is difficult to measure because the units of measurement of latency are becoming increasingly smaller. Indeed, latency is now typically measured in microseconds...and a microsecond is one millionth of a second.

Since data latency can stem from many sources, one must be able to measure each contributing source in order to manage it. In the case of securities markets and foreign exchange data, the primary sources of latency are application, network and geographic. Application latency refers to both the latency caused by the application that is generating the data and the latency in the application that is consuming the data. This could be due to any number of factors including the efficiency of the application and/or the processing power and configuration of the computer server(s) that the application is running on.

Network latency occurs when data leaves the application and enters the network over which the data is transmitted. This could be due to a poorly designed distribution network or insufficient bandwidth. In addition, geographic latency is attributable to the geographic distance that the

data has to traverse. This can range from microseconds for side-by-side applications, to hundreds of milliseconds for data travelling across continents.

Why has latency become such a burning issue?

The quest for lower latency has been led by the equities market, the first market to really embrace electronic trading, and more recently, automated algorithmic trading. But as other markets – including the foreign exchange marketplace – increasingly adopt electronic and algorithmic trading, data latency is becoming a central issue for all concerned. The Aite Group* estimated that the electronic FX market (eFX) represented 56% of the total FX trade volume at the end of 2006 and is estimated to reach 75% by 2010. The number and types of electronic venues for eFX have grown steadily in recent years, including single bank portals, multi-bank portals, inter-dealer portals, FX Electronic Communication Networks (ECNs), and retail trading applications.

Whenever trading is electronic, there are obviously advantages to be gained from being able to react more quickly to the market than your competitors. The early bird gets the worm, and a firm's ability to trade on an executable quote before someone else is key to profitability. All electronic trading depends on market data, i.e. quotes on which to trade, so undue latency in the data or application of one user or group will put them at a disadvantage over others using lower latency data.

And while data latency is crucial for all eFX trading systems, it is most important for the fully automated and/or algorithmic trading systems. This is because computerised mathematical models are being employed to generate trades automatically on the basis of incoming market data, and all trades are instantaneously executed via direct connections to electronic trading venues without human intervention. The decision to trade and ultimate execution can take place in milliseconds. On the other hand, on electronic systems where there is human intervention – for example, when a trader is entering a trade on a portal (inter-dealer, single bank or multi-bank) or ECN – the trader's involvement will add precious seconds at best to the transaction. This is because of the time it takes the trader to interpret the incoming data and execute the order. And even though the human interaction with the eFX system is much faster than the traditional conversational dealing alternative, it is considerably slower than a fully automated electronic trade generation and execution system.

Latency arbitrage

At the most basic level, market participants can generate profits exclusively by exploiting their competitive advantage in latency by engaging in latency arbitrage. Latency arbitrage involves using your speed advantage to profit from market inefficiencies and price discrepancies, while trading with counterparties that have latent data. While some market participants frown on this type of trading, others argue that it serves

* Aite Group, LLC: "Institutional FX Trading Platforms: Old Habits are Hard to Break", Report 200705071, May 2007



a purpose in forcing markets to be more efficient and transparent. Even for the majority of applications that run more complex models than pure latency plays, low latency is a core requirement of those systems.

Latency and liquidity

The key to survival for any trading venue is liquidity. Even in the very large and liquid foreign exchange market, with an estimated turnover of \$2.7 trillion a day, new e-FX venues must compete aggressively for liquidity to ensure survival in this very competitive arena. With more electronic trading venues, algorithms are being designed to seek and find the best execution, typically found on the most liquid markets. Whereas in the past the trading venue selection may have been more influenced by personal relationships, these algos trade purely on the basis of price. Hence trading venues that have latent data and/or latent execution systems will have a difficult time competing for order flow with their low-latency competitors. Multi-participant trading venues, including multi-bank portals, ECNs and Execution Management Systems (EMS) that aggregate order flow and quotes from across the market, provide a whole new level of market efficiency and transparency. A participating firm needs to be on par from a latency perspective in order to compete successfully with the others on the venue since all the data is going to be viewed side-by-side. Any discrepancies that a single participant may have in quality or latency – even on a single currency pair – will be very visible to all users of the system.

“Hence trading venues that have latent data and/or latent execution systems will have a difficult time competing for order flow with their low-latency competitors.”

For firms that aggregate the markets internally and do not use a centralised multi-contributor platform, data latency is also a key concern. Take, for example, a hedge fund or proprietary trading desk that has integrated the APIs of a number of external trading venues and is using those data streams internally to power a trading algo. If there is latency in the incoming data stream due to the API or other factors, then the system will not work as designed ...with potentially significant revenue implications.

What can be done to reduce data latency?

It is important to understand that data latency and capacity, in terms of the ability to process and distribute data, go hand in hand. If there are any capacity bottlenecks in the system, these will cause data latency in the same way that traffic jams will delay you in getting to your destination. The best designed and lowest latency distribution systems can keep capacity utilisation below 50% to ensure speedy data transfer.

To ensure that latency is kept to a minimum, firms need to look for application latency and examine the application that is being used to process or generate the data output for bottlenecks. Firstly, from a software perspective, application code should be reviewed to ensure that it is highly

efficient. As data volumes grow, certain processes may become overloaded and need to be rewritten and/or moved to different systems. Volumes may have grown to a point that the code base made need to be written in a more speed-friendly language. For example, Interactive Data's Real-Time Services business rewrote certain ticker plant components from Java to C++ because it considered that the latter was more suited to processing larger data volumes. Certain processes may need to be moved to other servers. It is also equally important to review the hardware infrastructure of the data processing/generating system; performance may be suffering due to growing data volumes or obsolete hardware, or a combination of both. Generally, better performance can be obtained from newer hardware utilising such features as faster processors and/or multi-processor configurations.

When it comes to application latency, it is critical to examine the application that is consuming the data in order to ensure that all the efforts to get the data into the application with the lowest possible latency are not negated by inefficiencies in the consuming application itself from a software or hardware perspective. For example, a complex algorithm running on the application that

“The global nature of foreign exchange, where trading counterparties can be a continent away, is more susceptible to geographic latency.”

needs to analyse thousands of data points in real-time may take seconds, so eliminating the benefits of reduced latency on the inbound data.

With network latency, network topology needs to be reviewed once again to ensure that there are no bottlenecks anywhere in the system, from routers to bandwidth constraints. As data volumes grow, higher bandwidth connectivity is required to carry the market data between counterparties. Higher performance routers may also be required to keep latency to a minimum.

And with geographic latency, there is no getting around the laws of nature; hence the greater the distance between the origin and destination of the data, the longer the transmission time will be. There is only one way to reduce geographic latency and that is to reduce the distance over which the data has to travel. To this end, firms have been moving data infrastructures closer to their clients, and there has been a big trend toward co-locating applications at central hosting data centers, so that the data generation and data consuming applications sit side-by-side to achieve the lowest geographic latency. The global nature of foreign exchange, where trading counterparties can be a continent away, is more susceptible to geographic latency.

One way firms can manage this is to have trading operations in multiple key FX liquidity centers closer to their trading partners.

Is FAST really a latency fix?

The FAST (Fix Adapted for Streaming) protocol has often been mentioned as a latency panacea. FAST is a data compression technology that reduces message size and bandwidth utilisation. This is an important development in an era of burgeoning data volumes; however its impact on latency is not yet entirely clear. It is true that smaller message sizes should allow messages to travel more quickly across networks; however the number of messages is not reduced and there are some inherent processing costs for compressing and decompressing data that are likely to claw back some, if not all, of the latency gains from smaller messages.

And in the end...

The equities market – even with the fragmentation of the past decade – has a much more concentrated number of trading venues than the FX market. This has been a key contributing factor to the growth on electronic and algorithmic trading in stocks. The race to reduce data latency in equity data has been underway for some time. Firms are taking extraordinary measures at

considerable cost to eradicate latency, including getting data directly from exchanges as opposed to market data vendors that aggregate exchange data. That said, for those firms that do need ultra-low latency data, Interactive Data offers DirectPlusSM, a fully managed direct exchange data service designed to provide access to sub-one millisecond data.

A direct exchange connection is established in order to eliminate the application latency for data processing at the vendor's ticker plant, and to reduce network and geographic latency by bypassing the vendor ticker plant. Much headway has been made here, and one has to wonder if much more advantage can be gained from squeezing out the last remaining microseconds of latency. Those players relying on their latency advantage would be well advised to explore new ways to generate alpha beyond pure latency plays.

On the other hand, market data latency in the foreign exchange markets will present the opportunity for competitive advantage for some time to come. This is because of the fragmented nature of the over-the-counter market, with multiple liquidity pools across multiple time zones. Expect to see FX market participants explore whatever means they can to reduce latency at every possible juncture to beat out their competitors.

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.