

Latency Measurement: Why and How

Henry Young, Co-chair of FIX IPL Working Group and CEO of TS-Associates, inspects the methods of latency measurement and relates new developments from the FIX Inter Party Latency Working Group.



We have all heard about the “cost of a microsecond” and the “latency arms race”. The relentless increase in price and transaction data rates in electronic trading and the drive to remain competitive are forcing financial markets’ participants to invest in the latest technology to support spiralling bandwidth requirements and drive down data flow latency. This has, in turn, given birth to a whole new industry in latency monitoring solutions, based on the premise that in order to manage something, you have to be able to measure it. This evolution has now played out to the point where traditional measurement techniques are no longer sufficiently accurate. The latency of hardware and software components used in trading systems is typically quoted in microseconds. Measurement solutions must therefore have resolution and accuracy in the sub-microsecond realm. This typically requires the use of specialist hardware, and out-of-band instrumentation techniques. The act of latency measurement should not itself add latency to business critical data flows.

But why is latency measurement important, and who uses it? The key driver behind latency measurement is competition at all levels in the capital markets. Execution venues compete for liquidity by being able to process orders, cancels and executions quickly. Deterministic behaviour is essential if market makers are to have the confidence to place large amounts of liquidity with execution venues, safe in the knowledge that they can pull orders from the book if the market moves rapidly against them. Prime brokers compete for order flow by being able to route those orders to market in the least time possible. Latency, both in absolute and deterministic terms, is also a vitally important factor in the success of algorithmic trading services hosted by many prime brokers. Some buy-side customers with routes to market through multiple prime brokers routinely monitor the latency performance



Henry Young,
Co-chair, FIX IPL Working Group
and CEO, TS-Associates

“Latency, both in absolute and deterministic terms, is a vitally important factor in the success of algorithmic trading services hosted by many prime brokers.”

of each broker and factor this information into their own systematic routing decisions. Finally, latency is perhaps the most important factor in arbitrage trading, where profitable trading opportunities may only exist for fleeting moments.

Latency is a multi-dimensional quantity. It can be viewed at different layers in the communications stack and at different levels of scale. The latency of network packets may be of interest to networking engineers ducking the blame for trading losses, or communications service providers reporting against service level agreements (SLAs) couched in terms of latency. However, the latency of application layer messages is often greater than that of the underlying network packets, due to the overhead of messaging protocols and correction of network errors, such as packet loss. So those with an interest in the latency experienced by applications will be less interested in network oriented monitoring solutions.

In terms of scale, as with many other human endeavours, latency has its macro and micro aspects. Inter-party latency is concerned with the latency of data flows between

different organisations, often at different physical locations. Here the concern is the time distance between trading applications and the various execution venues they have access to. This contrasts with intra-party latency, which is focussed on the latency of data flows within individual organisations. The purpose here is more often operational monitoring of trading systems, the ongoing performance optimisation of trading systems, or the wish scientifically to determine where best to invest effort in performance improvement. Latency has been referred to as the “canary in the coal mine”. Any operational problem will show up as an increase in data flow latency.

One of the latest developments in latency monitoring has been propelled by a recent trend in high frequency trading (HFT) systems architectures, with the move from the traditional distributed multi-server model, to the consolidated multi-core model. This evolution is driven by the desire to eliminate unnecessary sources of latency, such as network hops. The traditional technique used by latency monitoring solutions has been to tap into these network hops using passive network monitoring devices called network

taps. It has become necessary, however, to adapt to a world where there are no wires to tap between

An application tap enables visibility inside an "HFT in a box" system with high precision and virtually zero

and various execution gateways. Indeed, application taps can enable significant performance

"It has become necessary, however, to adapt to a world where there are no wires to tap between components..."

components within the latest "HFT in a box" systems. Just as Kennedy's moon programme gave us Teflon and the transistor, this development in trading systems architecture has given rise to the application tap, an important breakthrough in high precision software instrumentation.

performance impact. It is essential, for example, for a prime broker to be able to monitor the hop-by-hop latency across a set of processes all running on the same multi-core server: e.g. client connectivity gateway, order management system (OMS), smart order router (SOR)

enhancements by enabling routine but essential processing to be moved out-of-band.

Any thoughts about this or other articles? Please send any comments direct to: editorial@fixglobal.com

FIX Inter Party Latency (FIX IPL) Working Group

The FIX Inter Party Latency Working Group, is part of the FPL organisation and it has over 160 members encompassing exchanges, brokers, funds, service providers and latency monitoring solution providers. The working group has brought together leading industry experts to standardise two specific aspects of latency measurement.

Firstly, the working group has developed a taxonomy of standardised measurement points. This is vitally important, so that when different organisations

publish latency statistics that comply with the standard, financial markets participants will know the data can be compared on a like-for-like basis. The FIXIPL taxonomy is shown in the schematic.

Secondly, the working group is developing a latency data interoperability protocol. Currently, a situation exists in the industry whereby none of the available latency monitoring solutions interoperate, placing unacceptable bilateral constraints on purchasing decisions. A widely adopted latency measurement interoperability protocol

will, as with FIX, break down the barriers to the uptake of standards-compliant latency monitoring solutions.

The working group has recently reached the point where an early version of the FIXIPL Protocol is in testing between three of the firms represented on the working group. It is important to note that the work of the FIXIPL Working Group supports FIX, but will also be applicable to any data flow transaction, standards-based or proprietary.

