

BY
HENRY YOUNG
DIRECTOR OF PRODUCT DEVELOPMENT
TS-ASSOCIATES PLC

TS-Associates
Trading Systems Associates plc

Precision Instrumentation

In recent years, players in the automated trading industry have been investing heavily in accelerated trading infrastructure. This trend looks set to continue into 2010 and beyond. If you haven't been jumping on this trend yourself, get jumping! Whether you're an execution venue, connectivity service provider, market maker, liquidity provider or high frequency trading house, to ignore the fact that your competitors are all playing in the fourth dimension is tantamount to death by a thousand slow trades.

What is "accelerated trading"?

Simply put, it is the business of reducing the elapsed time across any or all parts of the trading cycle — price dissemination/discovery, decision making, decision validation (credit, risk, etc), order/trade submission and execution. But we're not finished, because the cycle comes full circle with the impact of trades and orders on market price. Repeat ad infinitum, or at least until market close. This is the pernicious feedback loop at the root of the latency arms race.

It has famously been observed that if you can't measure something, then you can't manage it. Applied to accelerated trading, this means that if you're investing in technology and services, with the objective of latency reduction, then it is equally important to ensure you're accurately able to measure the improvement achieved, and to do so on an ongoing basis as an integral part of your trading systems infrastructure. To know you're faster is not good enough if your aim is to be fastest. You need quantitative measurement, not qualitative estimation. You need precision instrumentation.

Measurement of the latency of data flows through distributed systems provides a key performance indicator (KPI) that is the "canary in the coal mine". Any type of operational problem will show up as an increase in latency. Whilst it's common practice to monitor other technical KPIs such as bandwidth or CPU utilisation, it is latency that impacts trading systems. By choosing to monitor KPIs that correlate more closely with profitability, you're empowered to fine tune your systems to optimize profitability and to identify rapidly the technical causes of any loss in profitability.

Existing users of precision instrumentation solutions are starting to realize that latency data can also be leveraged back into trading systems with the potential for profit enhancement, culminating in actual latency arbitrage. For example, accurate knowledge of your time distance from the various order books you could execute a trade on is useful tertiary input into smart order routing systems. Knowledge of how long an execution venue is taking to make your impact on their order book public to your competitors is useful input for fine tuning algorithmic trading systems, and those designed to trade against other firms' "algos."

Let's drill into the two key aspects that need to be factored into a holistic approach to precision instrumentation — network latency and compute latency. The measurement of each of these requires the use of specialist hardware instrumentation techniques in order to avoid the act of measurement impacting performance. This is a significant architectural challenge that one might consider to be the contextual equivalent of the Heisenberg Uncertainty Principle. In abstract terms, this states that you can't measure something without modifying its behavior. The trick is to use "out of band" techniques where instrumentation is implemented in parallel to, and is therefore not disruptive of, main line business processing.

Network latency is best instrumented using entirely passive and noninvasive network taps connected to monitoring appliances incorporating specialist packet capture and precision time stamping hardware. Although still a relatively young discipline, this form of network latency monitoring has been around since 2005, and is currently practiced by a number of vendors, including Correlix, Corvil, NetScout, SeaNet Technologies, NetQoS Trade Monitor, and TS-Associates with our TipOff product.

These products are all similar in terms of hardware architecture. However there are specific features that distinguish between them that need to be carefully considered in relation to your requirements. Important features include time synchronisation integration (NTP, PTP, PPS, IRIG, GPS, etc), message layer vs. packet layer latency (the former is required for multi-protocol data flows, the latter is only useful for network rather than application centric views), layer convergence (ability to function at all layers of the OSI 7 layer stack rather than requiring deployment of a patchwork of different monitoring appliances), support for emerging interoperability standards (enabling the integration of solutions from different vendors across inter-party demarcation lines), support for all your required wire protocols (some vendors focus heavily on standards based protocols such as FIX and omit support for proprietary middleware vendor protocols such as LBM, TIB/RV, RMDS, Activ, etc) and finally the critically important issue of performance (some vendor solutions designed for transaction monitoring don't perform well when applied to high frequency market data).

One of the more publicized aspects of network latency instrumentation relates specifically to an area known as inter-party latency. This is the latency of price and order data flows between execution venues and trading organizations. The industry currently lacks an open standard to enable interoperability between different latency instrumentation

solutions. The consequence is that vendors who focus on inter-party latency, rather than other aspects of precision instrumentation, are pursuing a policy of land grab and proprietary lock-in. Solutions are marketed on the basis that in order to have compatibility with exchange X, you must buy vendor Y's solution. This trend is damaging to the industry, and indeed the interests of the very vendors implementing such policies. The correct solution is for the industry as a whole to cooperate in implementing an open standard for the interoperability of network latency instrumentation solutions. The good news is that this process has started under the auspices of FIX Protocol Ltd, the not-for-profit organization that acts as independent guardian of FIX and related standards. The FIX Global Steering Committee recently approved the setting up of a working group tasked specifically with establishing and promoting such a standard, currently with the working title of FIPL — FIX Inter Party Latency. The group had its first meeting January 13th, and in only three weeks since its announcement has become FPL's third largest working group. Expect to see vendors (including TS-Associates) lining up to support this emerging standard, and customers who have already chosen one of the proprietary solutions looking rather red faced and short sighted.

Ultimately, the instrumentation of inter-party latency is a sticking plaster hiding more fundamental problems with execution venue timing services. Exchanges insert time stamps into their messages. The state of the art is currently millisecond precision, with some exchanges still providing second precision time stamps. Technology has moved on, and even millisecond precision is no longer adequate. Consider a co-located trading engine that is typically a few hundred microseconds from the execution venue's matching engine. What relevance does a millisecond precision time stamp have in this scenario? For execution venue time stamps to be useful for latency measurement, the precision must be greater than the baseline latency being measured, otherwise no reading is possible.

Exchanges need to improve their timing services. This means improving the precision of time stamps by three or more orders of magnitude, ensuring that the clocks generating these time stamps are accurately synchronised with UTC time, and being considerably more transparent about what actual event in their processing flow is triggering each time stamp.

Here at TS-Associates, we have a product framework called TradeSync, which aims to move execution venue timing services into the 21st century, with a four order of magnitude improvement in precision and accuracy. Adoption of TradeSync has already been announced by Chi-X Europe. A key part of TradeSync is its ability to generate order match time stamps to an accuracy of 100 nanoseconds. This is achieved using a hardware assisted software instrumentation technique called the Application Tap.

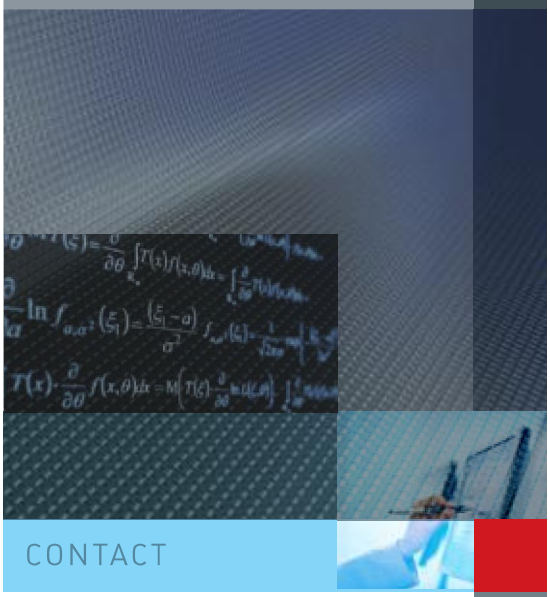
Measuring network latency only tells part of the story. Just as important is compute latency. The measurement of compute latency requires the generation of accurate time stamps triggered by software events. Traditional techniques for time stamping in software rely on operating system time services which are inaccurate and impair application performance. Enter the Application Tap. This is a new hardware product, delivered as a PCIe card which contains a precision hardware clock, FPGA co-processor and on board network interfaces. Using the Application Tap API, developers are able to implement precise software instrumentation with minimal performance overhead. The job of time stamping and processing instrumentation data is handled "out of band" by the FPGA firmware and optional external monitoring appliance linked to a network interface on the Application Tap.

Whilst the Application Tap enables precision compute latency instrumentation, it also has other uses including software logging, real time code profiling, device driver instrumentation (a technique we used to instrument the NYSE Data Fabric on Infiniband) and precise process/thread CPU utilisation statistics.

Taken together, a comprehensive network and compute latency instrumentation solution delivers real time monitoring of multi-hop multi-protocol data flow latency, with visibility not only at network hops, but with the ability to drill into software at the process, thread and even function call level. The Application Tap from TS-Associates is the first and currently the only product that makes this possible. ■

Latency Trends for 2010

- **Flattening of traditional layered network architecture**
- **Greater use of cross-over cables for direct interconnection of servers**
- **Greater use of kernel bypass techniques**
- **Precise time synchronisation fabrics deployed across enterprise data centres**
- **Standardisation of inter-party latency interoperability through the FIPL standard**
- **Improvement of exchange timing services — second to millisecond to microsecond to nanosecond**
- **Hardware assisted precision application and kernel software instrumentation**



CONTACT

TS ASSOCIATES
WEB www.ts-a.com
EMAIL sales@ts-a.com
PHONE +44 20.7415.7028