
Holistic Latency Monitoring: Finding the Chain's Weakest Link



Vision

In motor racing, everyone has a fast car, so rarely does raw speed win the race. The decisive differentiator is instead the driver's ability to make decisions more effectively and faster than the rest of the field. Low-latency trading is exactly the same. Compared to the time it takes to acquire market data, send an order to an execution venue, or even for the matching engine to execute an order, the internal decision-making process has become the fat-end of the latency wedge and is where the largest opportunity gains are to be made in the reduction process. Not getting there faster, but rather more quickly deciding where to go, is key to winning this race.

If trade volumes remain relatively consistent and predictable, skilled operations teams at banks are largely able to massage a labyrinth of database management and operating systems, highly distributed trading applications, as well as various interacting communications links, into running smoothly. The trouble, of course, is that once a major market event blindsides the system, lacklustre capacity planning can become crystal clear as things quickly go horribly wrong. Such events are a dime-a-dozen and category-one system failures at some institutions have become a daily, if not hourly, concern.

Identifying where the trouble spots exist is no mean feat but it is a task that nevertheless must be tackled, since failure to identify them can result in business Armageddon. It is of little importance if a bank has invested millions in a bleeding-edge pricing engine if it's still relying on old messaging middleware to transmit data to the connection manager. In such a case, that ultra-competitive quote won't make it out in time to clients and they will take their business elsewhere.

Banks have to think carefully about capacity constraints and potential bottlenecks that can occur as a result of an increase in future business volumes. In other words, to be successful, they have to factor in future success. Otherwise, they will fail.

A first port of call for many institutions has been foreign exchange. In contrast to the equities markets, where so much emphasis has been on the acquisition of market data or the execution of orders in localised markets, FX is a global, more disparate marketplace. To compete, therefore, leading houses have sought to tackle in-house problems first, rather than strive to improve third-party communication links and improve the time it takes to transport data across a global fibre optic network.

The analysis power of application log files has made them unlikely heroes. Using these seemingly antiquated resource tools, firms have been able to identify and track a single piece of data that constitute a trade and so determine what happens over the course of its life within their four walls: where it was left to queue, where it got lost in translation, where it finally met its end. Once the data has been aligned with server loads, it has also been possible to see how certain applications weigh on infrastructure and, from there,

statistically model the impact that future business volume growth will have on businesses.

In other words, using application log files, banks are able to identify capacity constraints before they occur. This type of insight boils down to a very fine understanding of why latency is occurring today and why that might bear little relation to how it will manifest itself tomorrow. It is a rare clairvoyance reserved for the few intellectually equipped to understand it. And for those banks allocating an ROI on latency reduction, hiring such individuals constitutes a large part of their investment.

But, armed with this knowledge, financial firms are able to build out their businesses with confidence, both in terms of re-engineering systems architecture, deploying new infrastructural resources, implementing standardised messaging middleware or re-writing code to bring outdated, legacy applications up to speed for redeployment within the modern electronic trading world.

And the lessons firms have gleaned from reducing FX processing latency are being spread across other asset classes, particularly equities, which are today turning their attention in earnest to the internal compute challenge. The benefits of this cross-fertilisation process are being increasingly recognised and new, more holistic approaches to latency reduction are emerging.

Table of Contents

VISION	1
TABLE OF CONTENTS	3
INTRODUCTION.....	4
PROCESSING LATENCY	6
THE FOREIGN EXCHANGE	8
WHY HOLISTIC LATENCY?.....	10
A LATENCY MECHANIC.....	10
LATENCY ISN'T A SPEED, IT'S A DISTRIBUTION	11
MEASUREMENT; MORE THAN A MEANS TO AN END.....	13
LOG FILES – REALLY?	13
CHASING RED HERRINGS	14
THE (W)HOLISTIC ANSWER AND FUTURE NIRVANA	16
CONCLUSION	18
TABB GROUP	19
THE AUTHOR	19

Introduction

It feels like latency reduction is yesterday's news. Milliseconds have been slashed to microseconds, bandwidth has turned the 10 GigE (Gigabit Ethernet) corner and is heading for 40 GigE, new fibre-optic pathways have been built slashing hundreds of kilometres off the multi-asset trading round trip, and co-location has gone global. The perception is that we are at the limits of speed where only the laws of physics, the geography of the globe and the speed of light stand in the way of the zero-latency trade.

The reality is that we have only just begun the journey. Moreover, our successes in reducing network latency have given us a misleading, rather two-dimensional perspective of the long road ahead. It's as if latency is exclusively a matter of distance and speed. Nothing could be further from the truth. Not only is latency caused by a number of elements, each governed by their own sets of laws and parameters, but reducing latency is a dynamic enterprise in which any success is relative. As soon as one source of latency is eliminated, another is created, and another piece of the puzzle needs to be solved in order to realise any gains. To truly address the latency challenge requires the understanding that holistic latency is more than the sum of the parts.

To comprehend why this is, we must first look at what latency is, where it comes from and how it manifests itself. Broadly speaking there are three types:

- ▲ **Propagation latency** is the time it takes for a signal to move from one end of a communication link to another, and where latency is a function of the speed of light.
- ▲ **Transmission latency** is how quickly a data packet can be converted into a series of signals to be transmitted on a communication link.
- ▲ **Processing latency** is the time it takes to act once market data has been received. This can involve transforming the data into alternate representations, updating an internal order book and making the trade decision, as well as the creation of new orders or cancels.

Propagation and transmission latency are within the terrain of the external network and, as discussed above, the financial services industry has already seen great success in reducing delays from these sources. Processing latency, however, is a more internally focused source of latency stemming from application processes that sit within the four walls of a trading firm.

Typically, general-purpose computer systems are used to host application processes in the form of trading systems software and, just as their name implies, are designed to be as general as possible in order to achieve a wide variety of computational chores. This versatility is a wonder of modern technology but means that, by design, such computer systems have to process instructions sequentially in a logical, serial way. Each instruction takes a fixed

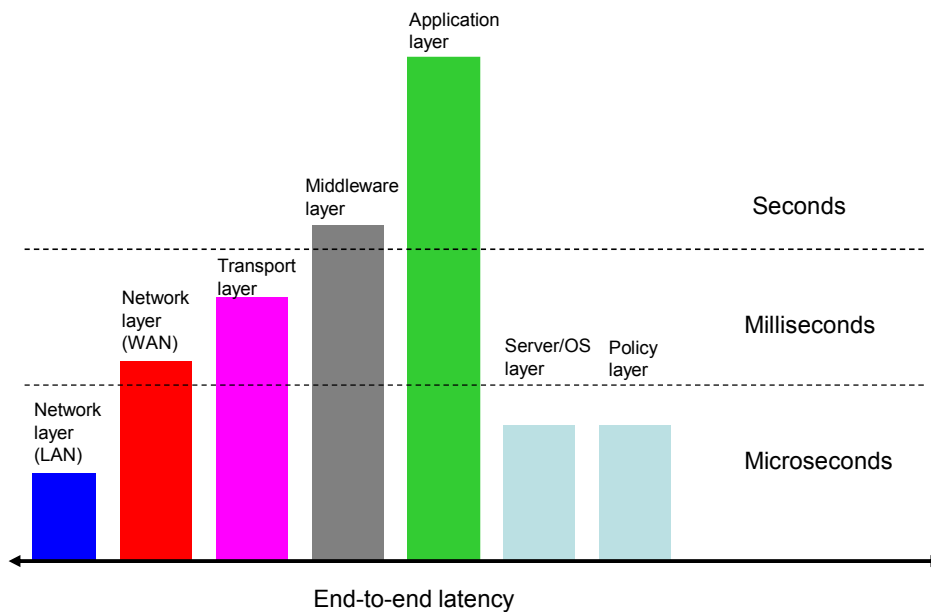
amount of time to complete, something that is related to the CPU's clock speed. The more instructions needed to implement a given piece of functionality, the longer the sequence takes and the greater the degree of processing latency. Multi-core processors and parallel programming techniques can make these processes more efficient, but the fundamentals of generic computer systems remain the same.

When one considers the hundreds of lines of code that are executed by an operating system simply to receive or transmit a data packet over a network, coupled with the many thousands of lines of code needed to process a market data feed through the pricing engine, the connectivity manager with a client, the session manager, the FIX gateway, the hedging system, the dark pool, the Smart Order Router (SOR), the execution venue – often in varying formats and with a number of messaging middleware applications – it soon becomes easy to see how application processing can dwarf other sources of latency.

Compared to propagation latency over the network, which can be measured in microseconds, application processes can easily take several thousands of microseconds (milliseconds) or even seconds to complete (Exhibit 1).

Exhibit 1

Comparison of Latency Effects on a Relative Basis over the Life Cycle of a Trade



Source: Cisco, TABB Group

A bank's ability to compete in the financial markets is only as strong as the weakest link in its application processes. It doesn't matter if the pricing engine is a state-of-the-art example of software engineering; if it's relying on old messaging middleware to transmit data to the connection manager, then that

ultra-competitive quote won't make it out in time to clients and they will have taken their business elsewhere. Likewise, the fastest fibre optic connectivity can quickly lose its usefulness if the servers on either end are incapable of keeping up with the data being fed into them.

These system inefficiencies can occur for a whole range of reasons and some can be solved with incredible ease. For buy-side firms, it may be as simple as de-fragmenting the hard drive or ensuring that the systems virus check doesn't occur during the market close. Others are considerably more complex, many of which we will explore in more detail later in this Vision Note.

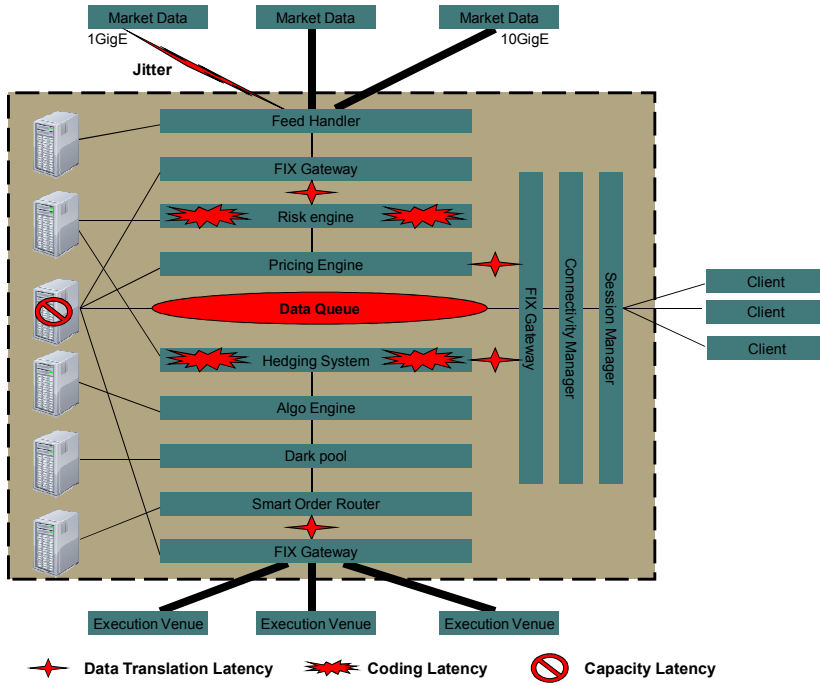
Processing Latency

Processing latency can be broadly categorised into four main types:

- ▲ **Data queues and capacity latency** relate to infrastructural bottlenecks that occur because of high data traffic caused either, in the short term, by a market event or, over the long term, by an overall increase in business volumes.
- ▲ **Quality of data and absolute latency** is the result of jitters caused by congestion in the network. This type of processing latency is largely caused by bursts in data traffic. Utilising 10GigE is proving to be an effective solution to the microburst challenge.
- ▲ **Application performance and coding latency** occurs for the simple reason that many firms rely on outdated applications in a fast-developing electronic trading world. Dealing with legacy applications in a world of ever decreasing latencies is the bane of most banking operations teams. So-called "Category One" crisis situations are a near-daily occurrence, and many financial services firms have had to set up triage units manned by former military personnel who are used to crisis situations in order to cope.
- ▲ **Inter-application messaging and data translation latency** is caused by the simple fact that not all applications speak the same language and need a translation protocol in order to communicate. FIX Gateways are the most common bottleneck since all firms, regardless of the internal application code, ultimately have to translate orders into FIX.

This Vision Note focuses on processing latency in order to demonstrate the importance of adopting a holistic approach. Any two-dimensional examination of latency ignores the fact that eliminating a bottleneck in one area often only leads to a problem occurring elsewhere. By viewing latency as a landscape made up of multiple, inter-related moving parts, meaningful success can be achieved (Exhibit 2).

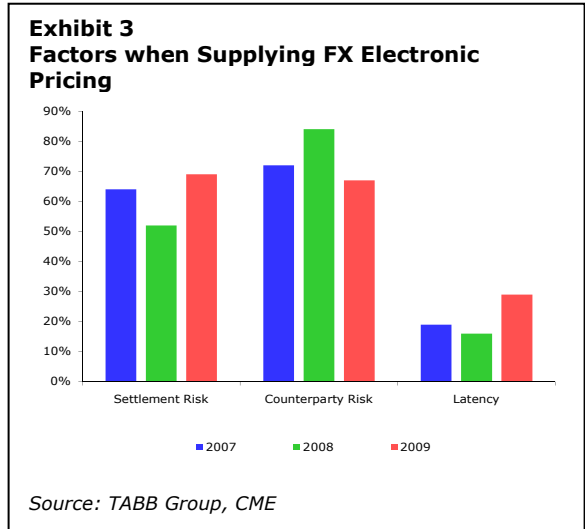
Exhibit 2
A Holistic View of Internal Compute Latency and its Various Flavours



Source: TABB Group

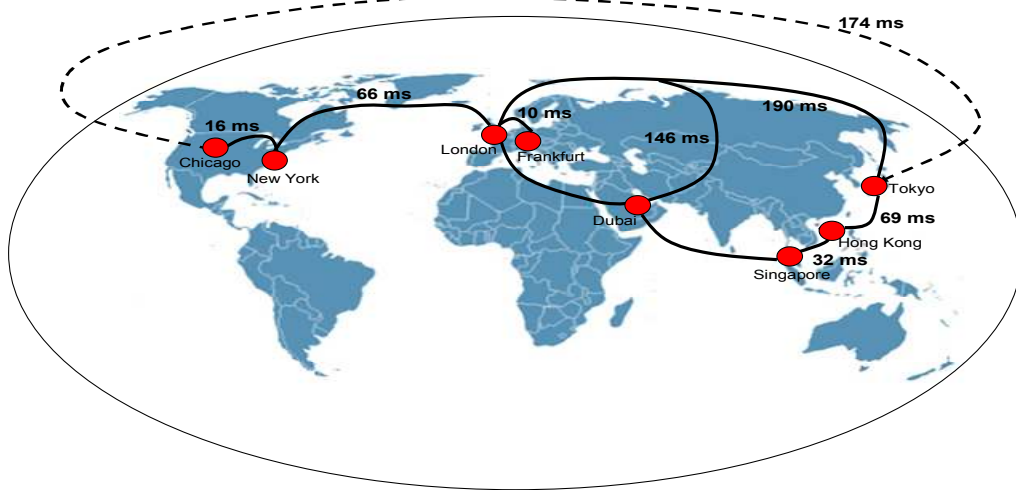
The Foreign Exchange

Latency has long been the bugbear of the equities markets but it is becoming increasingly important in other asset classes as well. Many trading strategies take into account movements in the futures, interest rates, fixed income and commodities markets to predict price moves in the equities markets. The futures markets in Chicago, in particular, are directly linked to the equities markets in New York, both for arbitrage trading opportunities, as well as for reduced hedging costs. In the options markets, some trading shops have embedded the Black-Scholes model directly onto computer chips to reduce processing latency. And according to the Chicago Mercantile Exchange's (CME's) *Annual Global FX Market Study*, concerns about latency in the FX markets have almost doubled to 29% in 2009 from 16% the previous year (Exhibit 3).



However, while the equities markets are used to exchanges advertising matching engine speeds in order to attract business, this is not the norm elsewhere. Foreign exchange proves a prime example as trading FX is a less localised, more global, affair. As a result, propagation latency is a far-wider and more challenging proposition for those trading FX compared to the equities markets (Exhibit 4).

**Exhibit 4
The Global Nature of the FX Markets Makes Propagation Latency a Larger Challenge**



Source: FXecosystem, TABB Group

This means that the FX markets are a good starting point for understanding application processing latency. Since latency reduction is about solving the easiest parts first, and the geographically disparate nature of FX markets is not going to change, participants in the FX market have been more drawn towards solving the internal compute challenge rather than the external network.

Banks have dedicated substantial resources to ensuring their matching engines are the best in order to garner market share, building business dominance on their ability to process market data feeds through their application processes more quickly than their competitors, just as equity execution venues do in the equities markets. Generally speaking, the FX market is a more application-intensive asset class compared to equities. While equities pose a greater challenge in terms of the variables of data, pricing FX in real-time based on multiple data feeds (as opposed to a handful of exchanges) requires more complex application processes. In particular, because FX transactions are carried out bilaterally, pre-trade credit checks and margining calculations continually have to be made before each and every transaction depending on the risk profile of each client. While banks only have to absorb a few price streams, they also have to offer prices based on their own book as well as adjust the spread that goes out to multiple clients throughout the day. Thus, once a tradable price has been determined from a market tick, it will typically result in hundreds of quote variations, as various customers require different spreads and precisions. Typically, OTC-streamed quotes have a limited lifetime of about 10 seconds, so this validity tracking also has to be built into the flow.

As a result of these two drivers, banks in the FX space have become experts in solving application processing latency. However, it should be borne in mind that a holistic approach also offers the attendant benefit of a cross-fertilisation of skill sets across asset classes. As firms re-organise themselves to address latency on a firm-wide rather than divisional basis, lessons learnt in one asset class can more easily be shared with another, shaving time and effort from future latency-reduction rollouts.

Why Holistic Latency?

The notion of a holistic approach to latency won't come as news to some. For those at the forefront of high-speed trading, for example, a holistic approach has been their mantra from the very beginning – you're only as fast as your slowest link. Indeed, most high-frequency firms started life by looking inwards, to what were, for them, the most obvious and more easily solvable sources of latency. In the days when co-location wasn't an option, being the best was about how fast you could make a trading decision as opposed to how fast you could execute that decision once it was made. Code was therefore modified to run on multi-core processors, secondary processes were offloaded to hardware-accelerated appliances, and switches and routers within their private network were removed to reduce hops. And when high-frequency firms did turn their attention in recent years to the external network, they did so with an awareness of what it might mean to their internal computing processes.

When considering an upgrade to 10 GigE from 1GigE for example, they asked themselves: what will be the impact on the rest of their systems? The upgrade would mean there was enough headroom in the pipe to handle microbursts, avoid queuing, packet jitter and the attendant knock-on effect on the performance of their trading algorithms; but what kind of bottlenecks would such a steady flood of data cause at the application-processing level? Would it make any sense to upgrade to 10GigE without first considering the use of a graphic processing unit (GPU) or field-programmable gate array (FPGA) to handle the fire hose of data? And what implications would that change have in terms of messaging middleware and FIX encoders?

A Latency Mechanic

Such engineers have been compared to the scientists working on Formula One racing cars. They know the impact that introducing a wider fuel pipe has on the performance of the overall engine. They are the so-called latency intolerant. They have been at the forefront of the latency reduction race and hold advanced perspectives on the problem.

But different firms have different needs and, as a result, are coming into the latency reduction challenge from differing perspectives. Global investment banks, for example, are not driving Formula One racing cars. As institutions they more closely resemble an entire fleet of vehicles made up of 20-ton trucks, some Porsches, a few high speed Hummers, and a range of Mercedes-Benz. The latency challenge for the engineers at these firms, then, is an altogether more complex and multi-faceted proposition compared to that of the high-frequency trading engineer responsible only for a single, super-light race car zipping around the ultra-low latency network track.

The race to reduce latency is about relative rather than absolute speed. This understanding has prompted many top-tier banks to identify latency as a competitive differentiator and they are in the business of allocating an ROI to

latency-reduction efforts, down to the finest microsecond. The business proposition is simple: if we can improve trading by x microseconds, how much will we need to spend on a new route, a new piece of hardware, a new piece of software, or an entire team of engineers (and it should be noted here that much of the 'I' in ROI relates to the high salaries of the talented few) in order to make it happen?

Latency isn't a Speed, it's a Distribution

For banks whose business models depend primarily on trade volumes, success is less about how fast a trade was transacted and more about what volume of trades were transacted within optimum time bands. Capacity bottlenecks – whether in terms of CPU, memory, network bandwidth or disk I/O – slow down processing speeds and throttle a bank's ability to compete. And bottlenecks are more likely to occur at times of high traffic when a bank most urgently needs to be able to demonstrate its ability to process market data quickly.

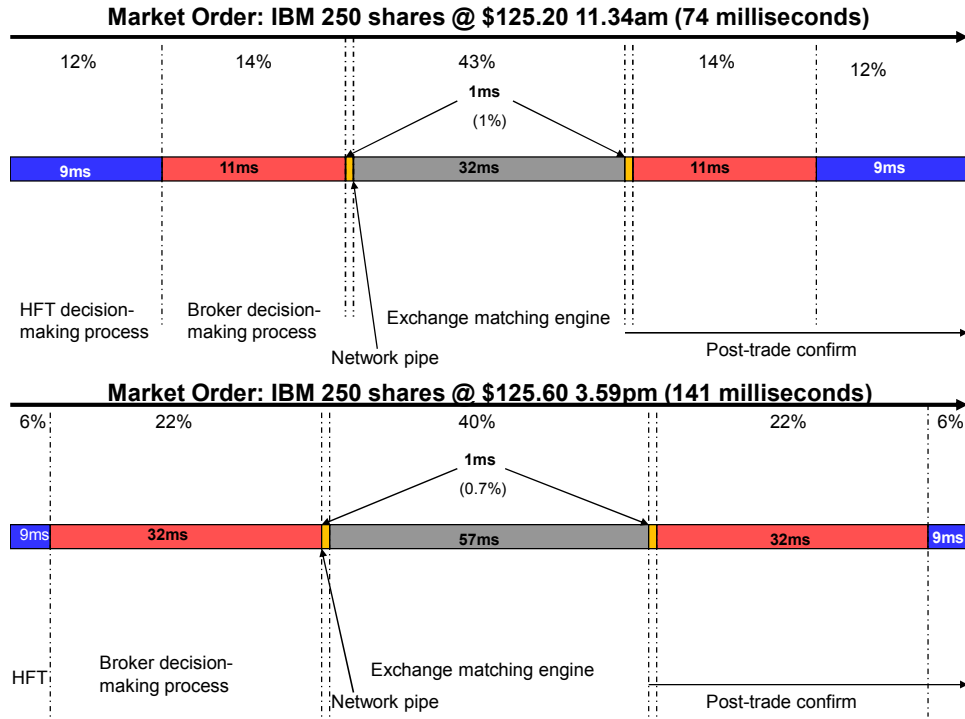
Consider a taxi journey to the airport. A taxi driver may be able to boast his ability to get to the airport quickly but the 'when' is as important as the time and distance. It makes little difference that he can get to the airport in half an hour at 3am if all the important flights leave in the evening when the same journey takes two hours. And what would it mean if he was, say, able to complete the journey in 30 minutes during the rush hour period?

Similarly, the 'when' is an important factor when it comes to considering capacity latency. There is little need for understanding latency over a 24-hour period, if latency during the busiest 1 millisecond of the day is a thousand times more than the average. This is a much more relevant consideration for application processing than for latencies that occur at the external network layer, which generally remain the same regardless of business volumes and traffic.

Exhibit 5 illustrates relative latencies over the various stages within the life cycle in a trade. The first is a market order for 250 IBM shares executed at \$125.20. At 11.34am the trade takes 74 milliseconds (ms) to complete, with the matching engine taking up the bulk of the time. The pre-trade decision-making process by a high-frequency trading firm and the broker take just 9ms and 11ms respectively, while the propagation delay is a mere 1ms, or 1% of the trade life cycle.

At the market close, however, the same trade takes 141ms to complete. The most significant sources of latency occur because of the pre- and post-trade broker decision-making process and at the exchange matching engine. The largest increase in percentage terms is at the broker level, taking up 44% of the trade life cycle. Propagation delay remains at a notably low level of just 1ms, now less than 1%, quite literally a slither in the entire trade life cycle.

Exhibit 5
The Life Cycle of a Trade at Two Different Times of Day



Source: TABB Group

Thus, with a holistic approach come new metrics for measuring latency. Capacity and distribution are just as important as speed and distance. Moreover, it is only by identifying and uncorking bottlenecks in application processing that banks can manage their infrastructural resources to better handle future growth in trading volumes.

Measurement; More than a Means to an End

It has famously been observed that if you can't measure something, then you can't manage it. How do we know we're truly saving the microseconds promised if no benchmark exists? Much time and money has gone into solving this metrics problem. However, despite cutting-edge hardware tracking tools and GPS timestamp mechanisms, another source of data as old as the computer itself provides huge potential, but has been largely overlooked.

Log Files – Really?

In these days of non-invasive hardware clocks, network taps and other sophisticated latency monitoring tools, application log files may seem like an antiquated resource tool. However, they hold principal advantages over the alternatives – they are readily available and free.

While no one would argue with the unrivalled quality that modern monitoring tools bring to any firm engaged in an analysis of application processing latency, it is important to remember that we are not as far down the latency reduction track as we'd like to imagine. The level of granularity provided by these modern measurement solutions isn't always necessary – why measure in microseconds when an application takes milliseconds to complete? Log file timestamps containing a margin of error of plus/minus one millisecond are not as redundant as they may seem. Finally, the analytical tools that use application log files for understanding latency have come a long way. No longer must throngs of IT support staff 'grep' through log files looking for an answer. Specialty software now allows us to paint quite a detailed, sophisticated picture of latency using this most basic of resources.

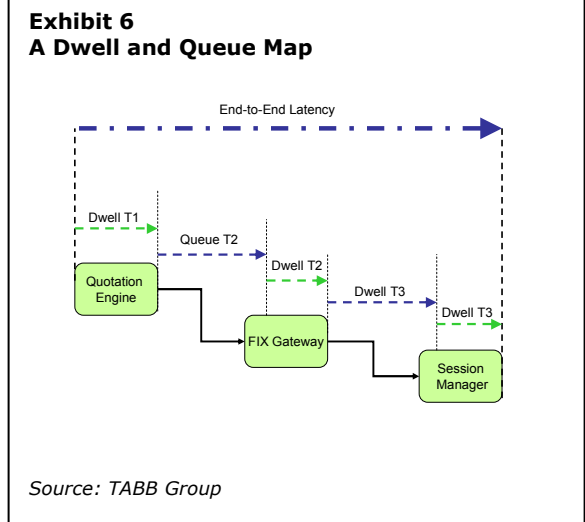
The application log records information about activities within application instances, such as date, time and server process ID. By joining up all the application log files across the stack, it is possible to single out individual packets of data as they move through the layers and determine how long each spent at each particular application. This level of visibility is only of limited use. Chances are a trader frustrated with a slow system will be able to more quickly identify a bottleneck in the system than a latency reduction analyst pouring over reams of application log files.

The real benefit comes when the log files are correlated with server loads to determine the available processing capacity left in the system and for identifying potential bottlenecks in the event of future business growth. By identifying the variability of latency according to business volumes, banks are able to introduce a level of consistency that they can then use as a benchmark for building out their businesses confidently.

The analysis is a multi-stage process. First, application log messages must be grouped into steps to identify the start and end of the processing of a quote or trade through the stack. As a transaction moves through the various layers it

changes status. Some market ticks will become quotes and some of these will finally become trades. The processing of transactions moving through the system consumes infrastructure resources (CPU, memory, I/O etc) and the consumption profile depends on the state of the transaction. There are two useful metrics here:

- ▲ **Dwell time** is a measure of how long a quote or trade takes to pass through an application component;
- ▲ **Queue time** is how long it takes for a quote/trade to leave an upstream application, such as a quotation engine, to being recorded on the next application down the stack, such as a FIX gateway.

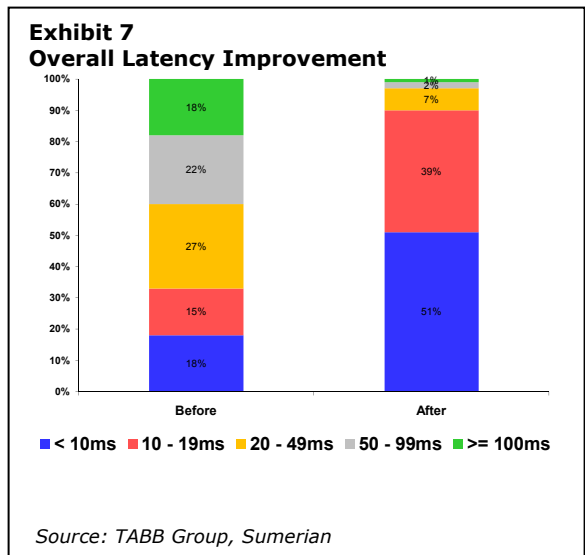


End-to-end latency for a quote or trade can then be determined by summing the dwell and queue times for each application/hop in a flow (Exhibit 6). The next stage of the process is to look at server loads and see how they correspond with the dwell and queue analysis. By aligning application activity with their corresponding servers it is possible to identify the maximum throughput capacity available before load limits are breached. Doing this provides a series of consumption data, which in turn allows for the statistical modelling of resource demand according to varying intensities of transaction volume in order to identify where bottlenecks in CPU may occur as business volumes increase. Having identified future potential bottlenecks ahead of their actual arising, action can be taken to remedy the situation. Applications can be moved onto different servers to increase CPU headroom, new infrastructure resources can be deployed, architecture can be re-engineered or, if necessary, application code can be re-written if coding latency is the culprit.

Chasing Red Herrings

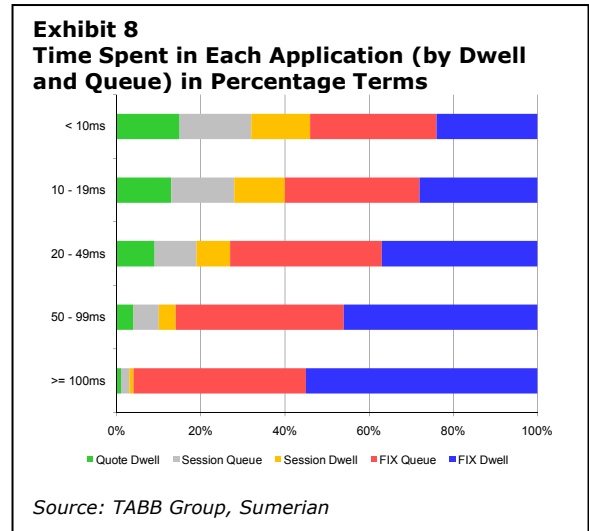
Take bank A as an example. Following an analysis of log files, it was able to identify bottlenecks that led to hardware upgrades and a re-balancing of software instances across servers. As a result, overall latency improved significantly, with 51% of quotes issued in under 10ms, compared to 18% in the period prior to the analysis (Exhibit 7). The volume of quotes in the high latency bands (greater than 50ms) was only 3%, compared to 40% prior to the analysis.

What emerges most clearly is the ease with which sources of latency can be



misidentified and the way in which banks can end up chasing red herrings if they fail to analyse application log file data correctly.

Exhibit 8 shows the percentage of time a single quote spent in dwell and queue terms through a section of the application stack – the quotation engine, the FIX gateway and the session manager. Each horizontal bar represents a different end-to-end latency band for each layer. Immediately it becomes clear that for those trades taking the longest time to complete (over 100 milliseconds), latency is occurring at the FIX gateway – the long queue and dwell times for which are highlighted in red and blue. The natural conclusion, therefore, is that a capacity constraint occurring at the FIX gateway is causing more trades to complete in more than 100 milliseconds and, in order to solve the problem, more servers and CPUs should be dedicated to that application in order to improve performance.



But once the dwell and queue profile becomes aligned with the relevant servers, it emerges that the problem is not occurring because of a capacity constraint. In fact, the FIX Gateway (along with the Session Manager) has plenty of CPU headroom. Its servers have the capacity to handle an additional 1.2 million trades – or a five times increase in business volume.

In this instance, it turned out that the real culprit was coding latency. A coding error meant that the application was unable to process trades efficiently as business volumes increased. A correction uncorked the bottleneck and the problem was solved. Just as importantly, the analysis also identified a future CPU bottleneck in the making, since the quotation engine server only had enough headroom to handle 66,000 more trades. After that, CPU limits would be breached and the bank would have another major latency concern. Thus, through the analysis, the bank was not only able to correctly identify the type of processing latency that was occurring in the present, it was also able to identify a future bottleneck that hitherto had not been on the radar. Had it followed its initial impression and dedicated more CPU capacity to the FIX Gateway, it would likely have only made matters worse by drawing rare infrastructural resources unnecessarily away from where they were really needed to solve a problem caused by an altogether different type of processing latency.

The (W)holistic Answer and Future Nirvana

Clearly, log files coupled with statistical analysis provide good insight into application processing latency and, when aligned with server load metrics, they can help firms understand the variable relationship between business volumes and latency. They also adequately demonstrate the point that latency reduction is a holistic challenge since, no sooner has one problem been solved, than another hotspot source of latency is identified.

But there are drawbacks.

First, there is complexity because log data files can vary and tend to be spread over a wide area. A bank may be using multiple operating systems and database management systems as well as a number of other bespoke systems – all with their own data formats and standards for recording processing events. Applications are spread across thousands of server nodes, with each server collecting hundreds of metrics every minute of the day, encompassing all the variable details that go into a single quote – price, quantity, source, execution venue and so on. Any attempt to mine such a vast, complex and inconsistent resource pool therefore needs a hugely scalable analytical system that can efficiently bleed key performance indicators (KPIs) from the raft of log files.

Second, using application log files for latency monitoring, especially when a high level of granularity is required, comes at the expense of the CPU. This violates Heisenberg's Uncertainty Principle, which states that you can't measure the property of some things without modifying other properties. In other words, the effort to measure latency becomes a source of latency or other problems.

Network measurement techniques provide a good example of how best to remove Heisenberg from the picture. Many of these latency-monitoring solutions offer 'out of band' techniques where passive, non-invasive hardware clocks are used in parallel to main-line business processing (Exhibit 9). This includes third-party network appliances and tools built directly into traditional networking hardware. Tapping into a small number of carefully chosen network points can provide good visibility into the activities of multiple applications spread across many servers.

For this reason, some firms deploy network packet capture tools that can record these activities in large volume. However, capturing the data while not introducing additional forms of latency is only one part of the solution. Adding application-layer awareness on top of high-capacity packet capture can turn the recorded raw data into a goldmine of information about how well applications are doing and where the bottlenecks are using a unified approach to both application and network latency management – all without adding any additional latency to the process.

Third-party latency reduction solution providers can use them for statistical modelling purposes remotely. Many banks will smart at the suggestion to allow third-party providers access to such proprietary information but it should be noted that third-party providers tend to have a perspective that internal specialists, with their narrow focus, lack. A broader level of experience means a third-party will, for example, instantly understand that pushing one nail down may only succeed in making another pop up elsewhere. Internal teams, despite the calibre of the professionals that work within them, can sometimes chase their tales in this regard. A level of co-operation is needed between the solution providers and also the internal operations teams at banks if a holistic approach to latency reduction is to be successfully implemented.

Exhibits 9
Sample Set of Latency Monitoring and Reduction Solution Providers by Type

Application layer	Software ▲ Ticker capture ▲ Pre-risk checks ▲ FIX Protocol Engines ▲ Algos ▲ SORs ▲ Dark pools	Application hardware ▲ CPU ▲ Memory ▲ Data Storage ▲ Disk I/O ▲ Log files	▲ Application Analysis (Cisco) ▲ Log file analysis ▲ Multi-data resource analysis (Sumerian) ▲ Hardware clocks (ITRS)	▲ CPU/Server resource re-allocation ▲ CEP (Progress Apama, Streambase, Sybase)	
Transaction layer	▲ Market data distribution ▲ FIX		▲ Transaction and network monitoring and measuring (Corvil, Correlix, NetScout, NetQos, Trade Monitor, Nimsoft, SeaNet Technologies, TS-Associates)	▲ Acceleration appliances (QuantHouse) ▲ Low latency message queuing (Tibco)	▲ Feed Handlers/Ticker Plants (Exegy, Coloxica, Fixnetix, Readline) ▲ Tick database (kx, Vhayu)
Network Layer	▲ Security: firewall, identity server encryption, etc.	TCP/IP overhead		▲ FIX adapted for streaming (FAST) ▲ Point-to-point middleware (IBM, Informatica 29West, RTI)	▲ TCP Optimisation ▲ Kernel bypass (Solarflare Communications)
Interface Layer	▲ Buffering ▲ Serialisation	▲ Ethernet ▲ WAN		▲ Serialisation ▲ Optimisation	▲ Ethernet 1Gb/10Gb (Arista, BLADE Networks, Cisco) ▲ InfiniBand ▲ Fibre Optics (ADVA, Ciena, Infinera)

Source: TABB Group

And with co-operation comes compromise. Banks are learning to tear down divisional silos and build out teams that are able to address latency holistically. For example, new application build outs are starting to occur in tandem with operations and infrastructure teams as well as the relevant business units. Without the necessary CPU, memory or data storage capacity, new trading applications are not worth the disk space they're written on. As a result, application developers are being asked to shed pre-conceptions that necessary memory and space will simply be made available to accommodate their inventions and are instead being asked to work side-by-side and in constant consultation with operations and infrastructure teams.

Conclusion

Holistic medicine encourages the co-operation of several specialists to improve the quality of life for the patient. The never-ending quest to reduce latency by financial markets firms is starting to look very similar.

The holistic latency approach isn't new. Those on the forefront of the low-latency quest have been espousing the philosophy since the beginning. They have long understood that latency is a challenge made up of inter-related parts which, when addressed together, deliver success in excess of their sum. The difference is that holistic latency isn't just a philosophy anymore – it's a reality. Advances in reducing latency across the external network, both in terms of propagation and transmission delay, have cleared the path so that we have the line of sight to ask new and more important questions.

Questions such as: what will increasing trading speeds do to overall system speeds? How should the business build out – in terms of infrastructure, new application development and connectivity – to avoid growth-throttling bottlenecks?

The answers lie within. Application log files have been at our disposal since the very beginning and are now proving to be an invaluable latency-analysis tool. However, when it comes to implementing solutions, addressing one component invariably affects others. Thus, piecemeal approaches are at best inefficient, and at worst unproductive. As a result, banks are fostering greater co-operation between competing third-party latency solution providers, and pressing for the introduction of standard terminology for latency measurement, particularly when it comes to agreeing on the points between which latency should be measured and expressed.

Sometimes, less is more. By rationalising the number of internal applications it takes to act, banks can react more quickly and better serve their clients. Many traders prefer to hit the lit price they can see than have their order routed to every execution venue available in hopes of a basis point improvement. As a result, more back-to-basics direct market access (DMA) products are being offered.

More revolutionary approaches that allow for greater efficiencies and the cross-fertilisation of latency reduction skill sets will emerge as the futility of piecemeal approaches to latency become further exposed. Second-tier banks seeking to compete in the FX markets are wooing technical expertise to help them build state-of-the-art matching engines. In return, the balance-sheet-rich banks are committing capital to funds that have been struck by redemptions and need investors. This is a symbiotic relationship that sees their mutual low-latency interests become aligned as one in the pursuit of low-latency Nirvana.



About

TABB Group

TABB Group is a financial markets research and strategic advisory firm focused exclusively on capital markets. Founded in 2003 and based on the methodology of *first-person knowledge*, TABB Group analyses and quantifies the investing value chain from the fiduciary, investment manager, broker, exchange and custodian. Our goal is to help senior business leaders gain a truer understanding of financial markets issues and trends so they can grow their business. TABB Group members are regularly cited in the press and speak at industry conferences. For more information about TABB Group, go to www.tabbgroup.com.

The Author

Will Rhode

Will Rhode joined TABB Group in March 2010 as an analyst based in London. He brings 14 years' experience as a financial journalist specialising in the risk management and derivatives industry, principally for *Risk* magazine and its associated publications. Previously, he was Editor Americas for *Risk* in New York and Editor of *AsiaRisk* magazine in Hong Kong. In addition to journalism, Will has also worked as a novelist and has had three thrillers published and translated into five languages by major international publishing houses. He holds a Bachelor of Arts degree in Political Science from the University of Newcastle-upon-Tyne, England.





www.tabbgroup.com

Westborough, MA
+1.508.836.2031

New York
+1.646.722.7800

London
+44.(0)203.207.9397