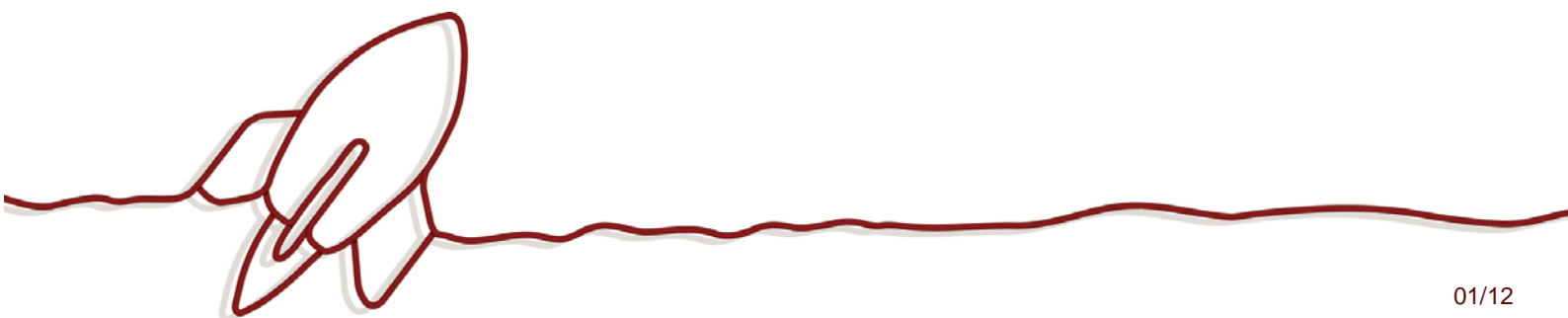


## Trading latency

### Achieving successful latency reduction with analytics

By establishing a holistic view that correlates trading activity against the latency round-trip and its distribution across all components, organisations can target latency reduction with quantified evidence into the key contributors.





## Introduction

In today's highly competitive investment banking landscape, the race to process trades at breakneck speeds has placed latency reduction firmly in the spotlight. But while solutions from faster infrastructure, network links and co-location proximity hosting are grabbing headlines in the market, they only represent part of the potential solution to the latency puzzle. In this introductory white paper, we examine how the advanced statistical modelling and analysis techniques of IT Analytics are helping leading investment banks gain a progressive, holistic level of insight for targeted latency optimisation.

## The need for speed

The latency question and the ensuing "arms race" it has propagated have taken the world of investment banking by storm. Whilst the rapid growth of automated algorithmic trading systems for vanilla workflow has been the primary catalyst for low latency technology spend, recent industry reports suggest that new categories of high frequency trading such as FX and equities are now joining the fray. So why is there such urgency being placed on latency?

Quite simply, it's the old adage of time equalling money. Achieving faster trade and order flows by improving latency throughout the processing chain can have a dramatic impact on the top-line; quotes can be generated more quickly, trading engines can work off data that is fresh and market accurate and avoid being arbitrated, there's faster liaison between trading venues, and pre-trade risk calculations can be executed rapidly. Industry experts TABB Group estimate that if a broker's electronic trading platform is 5 milliseconds behind the competition it could lose at least 1% of its flow – accounting for \$4 million in revenues per millisecond, with up to 10 milliseconds potentially resulting in a 10% drop in revenues.



Whilst some types of algorithmic trading platforms are already achieving latencies in the order of tens of microsecond speeds, leaving very little room for improvement, there is nevertheless significant potential to improve latency where tens of milliseconds or more are evident. Various solutions to the latency problem are being touted by a wide variety of vendors – from remote DMA (direct market access) appliances through new switching architectures to proximity hosting and exchange co-location. But throwing investment at infrastructure and network upgrades alone is not the answer. Latency is a fickle issue and can surface when you least expect it, appearing then disappearing at various points throughout the processing chain at different hours, days and months.

### The importance of a holistic view

To address latency methodically IT organisations need to gain an improved visibility into the trading functions and processes they support. By establishing a holistic view that correlates trading activity against the latency round-trip and its distribution across all

components in the trade/quote processing chain, IT organisations can target latency reduction with quantified evidence into the key contributors. To gain this view it is imperative to examine all the workings and interdependencies that comprise the trading environment. At a high level, this will often include the following common components:



**Market data sources** providing input that you will make decisions on



**Algorithms** that will decide on pricing, hedging, risk, valuations and so on



**Exchange or direct customer connectivity** where you will provide quotes or place orders



**A collection of network components** that connect all of these together

When thinking about where to focus attention, many organisations will focus first on the network links. But what should really be considered is the round-trip or end-to-end latency through all of these components – the time elapsed between getting data and taking action on it. For example, for LAN links, latency is typically measured in the order of tens of microseconds; improvements made here will only shave off a few microseconds from your end-to-end time. On the WAN, you might be looking at milliseconds or tens of milliseconds, depending on the distances covered; however, any potential improvements offered here will be limited by pure physics – nothing can go faster than the speed of light. For example, fibre optics can achieve speeds in the region of 200,000 kilometres per second - this means that every hundred kilometres you have to travel is going to add at least half a millisecond to your latency.

**“ Whether the application is a one-box algorithmic platform, or a multi-stage flow, application code can potentially add latency in the order of milliseconds or tens of milliseconds.**

So with latency already at a near optimised state on the LAN, there's little scope for improvement, and on the WAN the potential has a hard, physical limit. This leaves one remaining area to examine, and it is one that is often overlooked – the application code. Whether the application is a one-box algorithmic platform, or a multi-stage flow application, application code can potentially add latency in the order of milliseconds or tens of milliseconds. And, as it's typically all executing on a small set of servers in a confined geographic location, the physical limits tend to be less of a problem.

Therefore, by taking an approach that measures the relative contribution of each application component in the end-to-end, round-trip chain, organisations can manage latency holistically with quantified, informed understanding on where to invest and make any necessary improvements. Such measurements are now possible with the advent of sophisticated data mining and analytical





techniques. And it is with these advances in mind that IT Analytics is providing the necessary degree of accuracy and quantified evidence to target effective latency reduction optimisation.

### Collaborating with a specialist provider

Although organisations can attempt to undertake an analytical approach to latency reduction in-house using various software tools, there are a number of skill, process and resource challenges that can often make it difficult to set-up and maintain on an ongoing basis. By collaborating with a fully independent analytics service provider like Sumerian, organisations can take advantage of the dedicated skills, investment and expertise that has been built up across a broad range of clients and challenges.

### The IT Analytics approach

Sumerian's Analytics offers a unique approach to latency reduction by delivering quantified, holistic insight using real data. By combining the large volumes of existing data generated by trading environments with advanced statistical modelling and human expertise, Sumerian's approach supplies organisations with advanced reporting and deep "actionable insight" into the latency issues resident in their critical trading platforms - helping both infrastructure and application teams to concentrate on taking positive, practical action to drive business advantage. It enables organisations to answer questions which, although vital to optimised trading platforms, traditionally go unanswered or, at best, answered without the full facts:

- **What is the typical round-trip latency of a quote or trade?**
- **What happens to latencies during busy periods such as MPC or non farm payroll announcements?**
- **Which component(s) are contributing most to latency?**

An analytics approach (see Fig. 1) works by connecting and relating layers of granular systems data that exist within trading environment from low-level infrastructure components right up to business processes, trade/quote transactions and end-users – to make the IT supply/business demand relationships tangible and quantifiable. Using data captured from the trading environment, the first step in the approach is the creation of a baseline model.

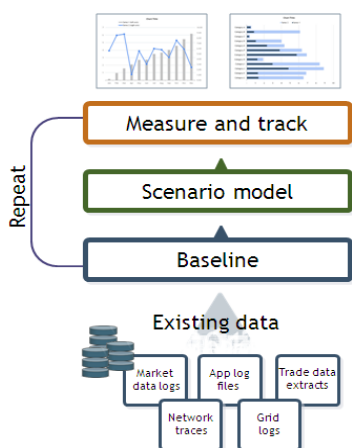


Fig.1 – Sumerian approach to holistic latency management and optimisation

### Creating a baseline

The baseline model is achieved by capturing and mapping existing data such as application log files and infrastructure utilisation data from underlying components that comprise the trading application/service. These metrics provide data dimensions for the volumes of trades being processed, the time spent on each component and the server/stage/network utilisation. The captured

## // Focussing on averages is not useful – it doesn't help to know your average is 10ms if 5% of transactions take over 100ms...

data is then analysed by Sumerian to uncover the precise levels of latency resident within each step of the processing flow.

### Understanding end-to-end latency

To fully understand your application latency, it is necessary to address a number of dimensions. Firstly, what is the statistical distribution of end-to-end latency? Focussing on averages is not useful – it doesn't help to know your average is 10 ms if 5% of transactions take over 100 ms. Analytics uses an explicitly statistical approach to latency analysis. By determining the latency of every single quote, order or trade, analytics can easily determine what percentage of business transactions falls within any given latency band, thereby providing visibility of the all-important latency tail (see Fig. 2).

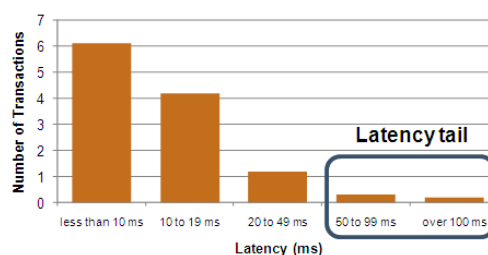


Fig.2 – Sumerian visualisation for latency against number of trades/transactions – indicating latency tail

Secondly, as well as knowing end-to-end latency, it's necessary to know how this breaks down across each application component, in order to determine where to focus attention. By building up a picture based on data from each application component (see Fig. 3), analytics identifies the relative and absolute contributions of each link in the chain.

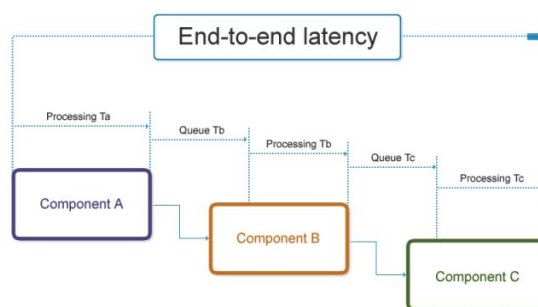


Fig.3 – Calculating end-to-end latency for each component in the trade/quote flow

Crucially, once these two dimensions are in place, the final step is to combine them. Analytics does exactly this, providing valuable visibility into how much latency each component contributes under different end-to-end conditions (see Fig. 4). This critical insight enables organisations to identify which components perform most poorly when end-to-end latency is stuck firmly in the tail.



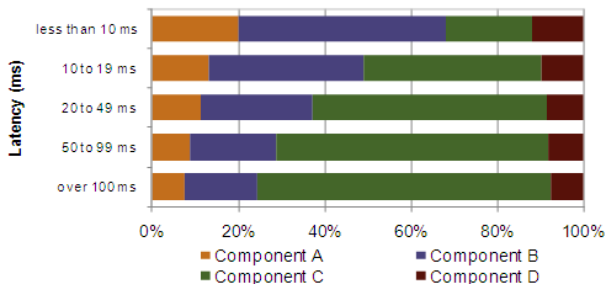


Fig.4 – Sumerian visualisation for latency round-trip distribution across each component in the trading flow

And it doesn't stop there. By combining the latency data with business volumes and identifying a relationship between these, it becomes possible to predict the impact of increasing volumes on your latency. In effect, this approach enables the certification of latency within specified business volume bounds.

Using the analytics approach described above, the following case study illustrates how its powerful capabilities can be applied to targeted latency reduction.

#### Case study 1 – Sumerian helps US investment bank to optimise capacity and reduce latency by over 75% for critical FX service

A US investment bank, with a multi-billion global revenue stream, took a strategic decision to set a double digit growth target for the FX side of their business. This meant the IT team had to rapidly assess the ability of the current FX systems to scale and meet the aggressive growth targets, moving quickly to mitigate any associated risk. FX as a business is characterised by massive volumes and extremely tight latency targets (sub 10ms tick to quote). In such a demanding environment understanding of latency and capacity across the supporting IT landscape is critical.

Sumerian began by capturing granular data from across the application to create a baseline model of the system, which included market data interfaces, price creation engines, risk management, spread and precision management, right through to the FX interface to customers and trade booking on the return leg. The captured utilisation data, which equated to approximately 500 GB, was taken from an 80 server estate located across the US and Europe.

Sumerian's analysis delivered a number of recommendations that were particularly valuable for the team. By quantifying the distribution of end-to-end latency, Sumerian identified bottlenecks that were limiting performance and only achieving latency targets in 25% of cases. Additionally, the footprint of trades on each platform was identified, quantifying the existing operating capacity in terms of trades per minute, providing a useful business-focussed performance metric. Also identified were capacity bottlenecks in two underlying system components which would limit growth to an 11% increase over current quote volumes. One of the actions taken here involved changing the load balancing across a group of servers. This resulted in sufficient headroom being made available to support 30% growth in quote volume, without the need for any additional investment. In conjunction, the bank implemented an action plan which included rearchitecting and redeveloping one of the FX application's key components, which was responsible for up to 95% of the overall latency.

To quantify the outcome of these adjustments, Sumerian ran a second series of analysis to remeasure the end-to-end latency. The results provided conclusive evidence into the success of the applied changes – round-trip latency had been reduced by over 75%. Overall, the bank found Sumerian's analysis so powerful that they have now engaged us to provide latency and capacity analysis service for the FX application on an ongoing basis – with another service key business applications also being added.

### Identifying key contributors and clearing bottlenecks

The results from the latency distribution analysis provide quantified evidence to pinpoint which components are contributing most to latency and at what levels. The causes of latency will differ across organisations and will be contingent on the particular trading environment/application under examination. However, common causes might include sub-optimal application code, imbalanced hardware, and capacity headroom issues. After the necessary period of reconfiguration and redevelopment, the original latency analysis should be repeated to quantify the actual outcome of the changes taken. This enables teams to categorically understand whether the adjustments made have been successful - and if not, undertake further action to correct. Unlike rudimentary methods of tracking such progress, analytics' ability to quantify the results and their relative impact in business terms is especially powerful for demonstrating value and justifying future investment.

### Scenario modelling latency

From understanding current latency and gaining insight to reduce it, scenario modelling techniques can also be applied to gain precise answers for forecasting growth requirements or making predictions around the impact of architectural or capacity changes under consideration. The advantage of scenario modelling is that it can safely and accurately answer any number of "what if?" questions around your latency reduction objectives – essentially providing a virtual testing environment in which to trial ideas before implementation. For example, scenario modelling can help organisations to understand the impact that a 40% increase in trade volumes would have on latency, and where bottlenecks are likely to occur.



The following case study provides a real world example of how such scenario modelling was used to model the impact of relocating latency sensitive applications to an outsourced HPC grid.

#### Case study 2 – Sumerian helps large UK investment bank to relocate latency sensitive applications to more cost-effective datacentre

The trading arm of a large UK bank was planning to move their HPC grid out of its current City of London offices to more cost effective premises outside the Greater London perimeter. The core trading application infrastructure was to remain in London, with the grid engines being outsourced. The bank had worked with Sumerian on previous analysis projects, and had found its advanced analytical capabilities to be advantageous in de-risking such large-scale change.

Sumerian captured packet traces from the existing network along with data from the grid's task reports, to build a model of the grid's current work loads. Sumerian was then able to forecast what the impact of additional latency would be between the outsourced grid engines and the core application infrastructure. The results of the analysis revealed that most applications would not experience any serious performance degradations with the additional latency; however, one application had a performance profile which placed architectural restrictions on the model to prevent data traversing multiple offsite locations.

The Sumerian analysis enabled the bank to plan which applications could be outsourced to the new grid location, and take appropriate action to keep the latency sensitive ones within shorter proximity. As a result, Sumerian enabled the bank to de-risk the project and ensure services were not impacted by the additional latency introduced.



## Ongoing monitoring and targeted investment

For latency reduction to be truly successful and keep the business competitive, it is important to maintain a running view of trading performance against specified target criteria, such as number of trades processed per minute. Trading behaviour and demand will fluctuate over time and there will be unforeseen changes that can't be predicted. Therefore, keeping a running view of latency is key to maintaining progress and exploiting optimisation opportunities. With Sumerian's repeatable analysis and ongoing tracking, trading applications can be constantly measured against key business criteria and KPIs, with trends and changes in trading performance proactively identified, allowing appropriate action to be taken by teams. Using analytics on a continual basis supports both the IT organisation and the business with a unified understanding into trading performance, providing quantified reporting and identifying areas where latency improvements will deliver the best ROI, all the while ensuring risk and costs are kept firmly in check.



## Staying competitive, growing business

The latency question and the resulting “arms race” has propelled latency reduction to the top of investment banking technology spend. However, the immediate assumption that upgrading infrastructure hardware and the network will solve latency issues is often misplaced. Instead, gaining quantified, holistic visibility into the end-to-end latency resident in trading platforms, to identify the true causes of latency bottlenecks *before* making investment decisions is an imperative in today's highly competitive landscape.

By using analytics to gain deep visibility into latency and its impact on trading throughput, organisations can focus attention on ensuring trading applications are continually aligned and optimised to serve the best interests of the bank - increasing competitiveness and supporting sustained future growth.

### More information

For further information on Sumerian or to arrange a demonstration of our services, contact us on 0141 229 7580, e-mail us at [info@sumerian.com](mailto:info@sumerian.com) or visit our Web site at [www.sumerian.com](http://www.sumerian.com)

