

Sumerian helps US investment bank to optimise capacity and reduce latency by over 75% for critical FX service targeting new market launch

Client challenge

Today's economic environment has driven large-scale change and merger activity across the investment banking community. In order to reclaim a strong position in the newly restructured market, banks have had to act fast and their IT teams have had to respond to change rapidly, delivering highly performant, flexible systems to support the changing business dynamics.

For one such US investment bank, with a multi-billion global revenue stream and operations spanning the globe, the changing market conditions lead to 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 access the ability of the current FX systems to scale and meet the aggressive growth targets, moving quickly to mitigate any associated risk.

Sumerian solution

FX as a business is characterised by massive volumes and extremely tight latency targets (sub 10ms tick to quote). In such a demanding environment having accurate visibility and understanding of capacity and latency across the supporting IT landscape is critical. Recognising the need for fast, accurate information, the investment bank's IT team engaged with Sumerian to analyse both the capacity and latency of the underlying IT infrastructure used to support the FX business.


Sumerian's approach is to capture low level, granular data from across the entire IT environment and use it to create a sophisticated and highly accurate model of the system. In this case, we modelled all aspects of the system including market data interfaces, price creation engines, risk management, spread and precision management, right through to the FIX interface to customers and trade booking on the return leg. Sumerian captured utilisation data from across the 80 server estate - physically located across the US and Europe - and captured log files from all application components. This equated to approximately 500GB of data for just one week's operation. We then modelled the correlations between quote and trade volumes and server utilisation and latency (both end-to-end and per component).

Summary

The challenge

- Complex FX trade application at large US investment bank, driving multi-billion global revenue stream
- FX business becoming more strategic to the bank – double digit growth target over next 3 years
- FX characterised by massive volumes and extremely tight latency targets (sub 10ms tick to quote)
- Concerned about scalability to meet business growth goals and delivering competitive levels of performance

Value delivered

- 
 CTO: Identified footprint of trade on each platform, and quantified existing operating capacity in terms of trades per minute.
- 
 Head of Infrastructure: Identified existing capacity bottlenecks, enabling implementation of change programme to horizontally scale existing application; resulted in headroom growth from 11% to 30% without any further investment.
- 
 Head of Application Development: Identified limiting component for end-to-end latency and relationship between volume and latency – significantly reducing latency by over 75%
- 
 All: Put in place recurring service to repeat analysis on a monthly/quarterly basis. Now expanding to seven additional applications.

Outcome and results

Sumerian's analysis delivered a number of insights and recommendations that were particularly valuable for the team. For example, Sumerian identified initial capacity bottlenecks in two underlying system components which would limit growth to an 11% increase over current quote volumes. One of the immediate actions taken, based on Sumerian's recommendations, 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.

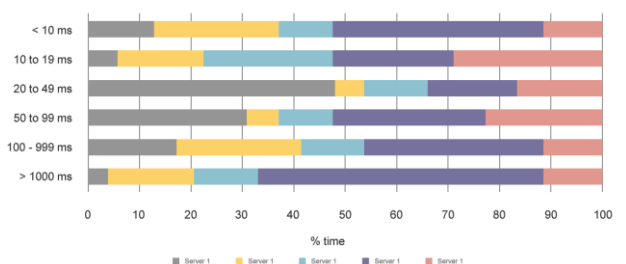


Fig. 1 – Example of Sumerian visualisation for latency response times for reduction optimisation

Also, by quantifying the current end-to-end latency across the system, Sumerian identified bottlenecks and resource constraints that were limiting performance and only enabling the bank to reach its latency targets in 25% of cases. This shortfall, if not addressed, would hamper the bank's competitiveness in achieving its targeted market position. 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.



Fig. 2 – Example of Sumerian visualisation for latency measurements and reduction

Based on Sumerian's analysis and recommendations, 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 actual outcome of these optimisation adjustments, Sumerian ran a second series of analysis to re-measure the end-to-end latency. The results provided conclusive evidence into the success of the applied changes - end-to-end latency had been reduced by over 75%.

Overall, the bank has found Sumerian's analysis so powerful that they have now engaged us to provide a capacity and latency modelling and analysis service for the FX application on an ongoing basis. In addition, the bank is now expanding Sumerian's service to another seven key business applications, enabling the bank to optimise its trading services across the board and achieve a strong competitive market position.

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 or visit our Web site at www.sumerian.com