



WHITE PAPER SERIES

*The Goal Posts are Widening:
Addressing the ever increasing need for performance*

by Nils Undén, cto, SunGard Front Arena

The Goal Posts are Widening:

Addressing the ever increasing need for performance



ABOUT THE AUTHOR

Nils Undén, cto, SunGard Front Arena.

Prior to joining SunGard Nils held the position of VP Application for Frameworkx. Before Frameworkx he was director of business architecture, at IMI, responsible for establishing and enforcing a consistent product architecture across the suite of products in IMI's portfolio. He has also worked for the Swedish arm of both IBM and Logica as a consultant and architect.

Financial Institutions want more accurate calculations delivered at a quicker rate for more complex instruments and in greater volumes than ever before. And they want this all achieved through one, high performing, cross-asset system. Nils Undén explains how new technology and techniques from other industries and markets can be applied to financial technology systems in order to address these performance issues.

System performance is not a new issue in any market but for financial institutions it has gradually become a highly topical and pressing concern. In part this is due to the nature of financial institutions and their competitive nature. The systems are always run to their limit and the more performance is improved, the more is asked of them. It is akin to the owner of a Ferrari. The cars are always fast but the drivers always want them to be faster.

In other ways the concern stems from current market conditions. Volumes have increased massively of late thanks to the advancements in electronic trading and fervent consolidation. At the same time instruments have become more complex and more varied, resulting in portfolios of increasing size, variation and sophistication. Furthermore the pace and competitiveness of the current market means that calculations not only have to be more accurate than ever before, they also have to be delivered within an ever-shortening timeframe, for instance within market making where times are measured in microseconds.

For vendors that are supplying large scale

financial systems the issue of performance is paramount – to deliver solutions for highly computational and intensive applications; to keep up with the continually escalating trade volumes; and to provide consistently accurate pricing for all possible contracts.

Calculation speed has become the defining property for any trading and risk management system but due to size and complexity of the average trading portfolio and the pedestrian nature of traditional numerical methods such as Monte Carlo simulation, vendors have had to reinvent their offerings.

For example, a typical portfolio will contain a substantial number of positions and a cross-asset mix of financial instruments each with their own associated financial model ranging from slow to fast in terms of calculation rate. As for the environment in which these instruments are traded in, the constantly changing values of stock and the need to recalculate any portfolio analytics accordingly creates a significant computational challenge.

Vendors have also had to rethink the architecture behind their systems.

Financial stand-alone silo-based systems in favour of much larger cross-asset trading and risk systems. This trend means that systems now have a much wider variety of users than ever before – traders, risk managers, heads of dealing, electronic exchange market-makers, retail and institutional sales teams, mid-office staff – each of which have their own specific demands on the one system.

Traders, for example, are demanding increased trading capacity and also pricing and risk analytics that can demonstrate “just-in-time” responsiveness. For those trading in more structured and exotic instruments and relying on numerically-based, OTC contracts, there is the need to perform accurate pricing, conduct their own analysis and also structure new offerings.

For the heads of desks, chief traders and risk managers, their principal need is to be able to view several aggregated portfolios, all updated in as close to real-time as possible. This will give them an enterprise-wide view of trading activity and the accompanying risk exposures, allowing them to take swift action against specific traders should the need arise.

Market makers on electronic exchanges and particularly those dealing in derivatives, now have to provide thousands of quotes for the same instruments to cater for the growing demands of individual traders. But with the advent of high frequency traders all scouring the market for pricing anomalies - market makers have to provide these prices with the absolute lowest latency to avoid being victimized by arbitrage trading.

For users in the middle office, the requirements are all about scale as they now have to produce more advanced and more numerous reports covering compliance, risk and profit & loss – all on a scale never experienced before.

While the urgency of and dependency on these requirements is welcome news for those vendors able to offer these solutions, the task of delivering speedy, scalable and accurate offerings to a range of different users all

with different requirements brings with it a pressure that can only be addressed through the use of new technology and techniques.

New technology and techniques

In order to attain a higher level of performance, systems must be more efficient in their use of processing power. It is not simply a case of relying on Moore’s Law and hoping that necessity will be the mother of invention. Successful vendors will have realised that it is about working smarter by eliminating system redundancies or unnecessary calculations and using new system configurations to maximise processing power.

The critical performance areas are twofold – calculation speed and scalable system architecture. Consequently the new technologies that can boost performance fit into one of these two categories, as shown below.

Faster calculations

Lazy evaluation - This fundamental concept ensures that no calculations are performed unless absolutely required. Calculating a theoretical value and displaying it in a user interface is a costly exercise, particularly as it needs to be updated and recalculated whenever the underlying asset changes its value.

However these values are often hidden and remain unobserved - such is the pace of market changes. Therefore, rather than create a costly dependency chain between real and theoretical values, lazy evaluation will update values only when explicitly called for. For example, an open but not visible application or an application part (e.g. a calculated cell) will not be updated until brought into view.

Update merging – This is another technique used to save on processing and calculating power. It ensures that only the most recent update applies if a single parameter is used. For example, should several consecutive updates of the same price take place before a recalculation can be performed only the latest update will be used, eliminating the

prospect of a queue of pending updates. This is a particularly valuable technique when the market is rapidly escalating, ensuring that users do not fall behind in their calculations.

Smart greeks – The clever use of pre-calculated greeks can yield substantial improvements in portfolio performance in exchange for a nominal fee. For numerical models such as finite difference and Monte Carlo simulations, resulting greeks can be calculated at the same time as theoretical values and can then be delivered back to users as cached parameters to be used as parameters for other calculations.

Multi-threading – This is a technique primarily used to spread smaller calculations across several, local computing cores. This technique has become a necessity given the reduced cost of standard, multi-core CPU boards. Consequently, a large scale trading system should make its core-functions ‘thread-safe’ so that a function will automatically be allocated to a processor within a local multi-processor environment.

Grid computing – ‘Gridding’ is a relatively mature technology in today’s terms having been around and in use for over five years now. Its benefits are very clear in that users can offload the heaviest of their calculations to a grid network therefore making a significant contribution to both portfolio start-up and recalculation times.

Typically a calculation is sent to the grid when it is evaluated to take more than around 20 Ms. When declared as grid-enabled, a calculation will automatically be distributed to the grid network if such is connected – else local hardware is used. A number of banks have implemented their own grid infrastructure, separate from the trading system that they use. By offering a trading system with its own grid infrastructure available through an application processing interface (API) or through ready-to-use adaptors that work with existing grid infrastructures, users should be able to get greater benefit from their ‘gridding’ with a lower total cost of ownership.

Remote sheets – The objective for remote sheets is to alleviate memory and CPU load on clients by offloading large portfolio calculations to a service cluster located on server-side, typical examples are P&L portfolios and desk risk portfolios as those used by heads of desks, chief traders and risk managers. These are then sent back to the clients as view-only reports displayed in the running application. This remote view looks and feels exactly as a locally calculated risk view would do and has the added benefit that they can be shared across multiple end users via their clients.

Scalable architecture

Multi-tier architecture with database cache

– A middle tier component can be used to cache data from the system’s database tier to the client tier; this avoids the need for the clients to go all the way to the database tier to access data.

Lazy writes and scalable distribution

– Servers can employ a lazy write technique for writing price updates to the database. The server sends price updates to subscribing clients first, and then writes them to the database in batches. This ensures immediate access to prices. To maximize the number of concurrent client connections, you can run several data servers in parallel.

64 bit architecture – This ensures that the system uses the best available hardware, enabling servers and clients to access more memory and improving performance by reducing latency. Servers are able to manage all business data through a larger cache, therefore improving the capacity for the system to handle larger volumes of trades and instruments.

Server-side aggregation of positions – By employing a concept of position/trade aggregation, the number of trades in a system can be kept at a manageable level. Trades are aggregated in night-batches so that clients can start next days trading with a small trading footprint and with their P&L and other values aggregated and readily available for position monitoring and simulation.

Server-side quoting - This is essential in order to have minimum response times on any market making activity. Servers are fed with vectors of pre-calculated values of re-quote instructions from clients on large amounts of derivatives, activated by any market price update of the underlying – giving market makers and traders an important edge in extremely quick reactions to changing market conditions.

Scalable throughput - For optimal performance and to be able to handle massive throughputs of trades into the database, systems should employ a clustered approach to servers which are able to support parallel processing of transactions and messages and guarantee much-needed market access throughput to any market.

Applying these techniques to financial systems. The common concern with any new technology, particularly when applied to mission critical applications such as trading systems, is that it will be unproven. So how can these concerns be allayed in this context?

Although some of the technology may be new to trading systems, developments such as grid computing, remote sheets, multi-threading and 64 bit architecture have been established for some time in other industries. The growth of grid computing vendors such as Data Synapse and Platform Symphony is a testimony to the acceptance of and demand for this technology.

The economics of today's technology market also serve to make a compelling case for grid computing, multi-threading, 64 bit architecture and other approaches based on using multiple central processing units (CPUs) to spread the processing burden.

Moore's Law may state that processing power doubles every two years but bandwidth availability is currently doubling every nine to eighteen months, meaning it is more economical to add more CPUs at their current capacity rather than waiting for more powerful ones to develop. Users need not be overly

concerned about the location of these extra CPUs and whether or not they are on the same motherboard.

A number of users may have experimented with 'gridding' or multi-threading and developed their own internal applications based on the technology. For a number of vendors this technology has now become more standardised and more accessible and these kinds of features can, for the first time, be formally included within vendors' offerings.

This is the case with SunGard's Front Arena, a front to back trading and risk system designed to serve the needs of three distinct user groups – pricing and risk managers; institutional sales and online or retail clients.

By employing all of the above techniques, Front Arena is able to maximise efficiency by avoiding unnecessary processing, distributing tasks across the entire system and employing scalable architecture on the server side. The aggregate effect of this deployment is truly superior system performance.

Approximately 1,000 transactions per second can be processed by Front Arena through a buy or sell-side FIX connection, while twice this amount of transactions – 2,000 per second – can be processed through the internal market. Thanks to a millisecond response time, over 320 trades per second can be entered into the Front Arena database and 12,000 exchange-based price updates can be distributed every second.

The value of these statistics is supported by the experience of early adopters of Front Arena's new approach. A leading German bank has been able to handle 150,000 trades per day using Front Arena. Meanwhile an internet broker has been able to serve 150,000 registered users and 10,000 concurrent users with the system.

Ensuring maximum performance is a constant task and vendors should constantly be looking at new technologies that can help to improve system's efficiency.

About Front Arena

A trading solution serving a range of financial institutions, SunGard's Front Arena provides straight-through processing by integrating sales and distribution functions, trading capabilities, and risk management. Institutional asset managers and brokers, traders, and market makers use Front Arena to trade equities, fixed-income, interest rate derivatives, and credit. Front Arena's components can be configured to meet a variety of trading needs and scaled to support operations ranging from local to global.

For more information, visit www.sungard.com/frontarena

About SunGard

With annual revenue of \$5 billion, sunGard is a global leader in software and processing solutions for financial services, higher education and the public sector. SunGard also helps information-dependent enterprises of all types to ensure the continuity of their business. sunGard serves more than 25,000 customers in more than 50 countries, including the world's 50 largest financial services companies.

Visit SunGard at www.sungard.com.

©2008 SunGard

SunGard, SunGard Adaptiv and the SunGard logo are trademarks or registered trademarks of SunGard Data Systems Inc. or its subsidiaries in the U.S. and other countries. All other trade names are trademarks or registered trademarks of their respective holders.