Financial Markets INSIGHTS
The Realization Group

Powering Up:

Handling high-performance computing to boost alpha and risk management

The complexities surrounding the Banking, Finance and Insurance sector today have led to a significant growth in the use of grid computing and high-performance computing (HPC) for computationally-intensive tasks. These are many and varied, and include areas such as derivative pricing, risk analytics, quantitative modelling, portfolio optimisation, and bank stress testing.

In this article, Mike O'Hara and Dan Barnes look at the areas where HPC and grid are being used in financial markets, and hear from ***Alastair Houston*** *of Nvidia,* ***Andrew Jones*** *of NAG,* ***Robin Mess*** *of big xyt,* ***Leon Lobo*** *of the National Physical Laboratory and Verne Global's* ***Stef Weegels*** *and* ***Lewis Tucker,*** *about the key considerations that firms should take into account when putting together the necessary infrastructure to support their computationally-intensive needs.*

Introduction

The intensity of risk calculation within Europe's banks is reaching fever pitch. Local, regional and international regulators are imposing increasingly quantitative measures to ensure risk is being adequately managed.

For banks, regulatory pressure means a more focussed modelling of risk, on a more frequent basis, using externally imposed ratios and measures. But this creates a technical challenge. Running a significantly higher volume of calculations is more complex, and investing in the necessary software and hardware can be expensive.

Although providing regulators with the most accurate reports is not a profitable enterprise, using more nuanced risk models can nevertheless make capital requirements less onerous, thus reducing total costs for an enterprise. However, the outlay and support overheads for software, skilled personnel, servers and the data centres in which they are applied all need to be well managed.

In this low-yield environment, the use of enhanced data analytics can make a difference to the top line as well as the bottom line. Many firms are seeking to become data-led businesses by taking their growing data sets and finding points in the market to seek alpha. To optimise the expenditure for this increased data analytical capability, where firms are now running numbers across a huge number of compute cores in order to get the figures they need, they must address how they implement the infrastructure that underpins it.

Risk & Regulation

The financial crisis was a failure of risk management. It led regulators to redress the balance by requiring banks to perform increasingly stringent stress tests and adopt more risk averse models, by computing a more detailed picture of their activity under capital adequacy regulations, from Basel III to the upcoming Fundamental Review of the Trading Book (FRTB).

"You've got the Comprehensive Capital Analysis and Review (CCAR), the stress testing framework introduced by the Federal Reserve in the US; you've got the Basel Committee's FRTB going live in 2019; and even the upcoming MiFID II will require firms to run a far higher volume of data processing than ever before, to better assess how they are trading and to report to their client and competent authorities," notes **Stef Weegels**, Business Development Director at Verne Global.

Under Basel III, banks had to raise their capital thresholds, increasing Tier 1 capital up from 4% to 6%, while it added various measures to balance out risk – capital buffers and leverage ratios – which ensure banks take a more granular insight into their risk exposures than they have needed before. The standardised approaches to operational risk and credit risk have been subject to revision since 2014 and the on-going review of capital requirements is reason enough for banks to look for operational flexibility as they establish their risk management infrastructure.

The FRTB will require banks to use either an internal model to assess their risk, which sets charges and fees in order to tackle issues like default risk, expected shortfalls and non-modelled risks; or they can adopt the standardised approach which increases the capital charges they face around default risk and also the fees for using more complex instruments. For either approach, the level of detail required to comply is far higher than has been needed before, with a requirement to analyse risk exposure from across the firm right down to trading desk level.

" Regulators have been applying increasing pressure to drive the banks to a better understanding of their risk, so that they're able to run intraday or real-time risk, rather than as a batch overnight process."

Alastair Houston, Nvidia

The intense level of data processing necessary to comply with these new regulations has resulted in a big push for the use of high-performance computing (HPC).

"Regulators have been applying increasing pressure through Basel III and some of the Dodd-Frank regulations in the US to drive the banks to a better understanding of their risk, so that they're able to run intraday or real-time risk, rather than as a batch overnight process," says **Alastair Houston**, Business Development Manager for Financial Services at graphic processing unit (GPU) specialist Nvidia.

"For the banks that want to run their own internal model approach rather than take the standardised approach, that will, according to a number of our partners and customers, require a step function increase in the amount of computing they need to run", he adds. Anecdotally he reports hearing that up to 30 times existing computation power might be needed.

Search for Alpha

Larger data sets and more powerful compute resources are not only being directed towards risk and regulation. As the volume of accessible data increases, so too does the potential for generating alpha. Quantitative analysts are being deployed across buy- and sell-side trading operations to find new trading opportunities and to automate low-level trading where possible.

According to analyst firm Hedge Fund Research, equity 'quantitative directional' strategies saw inflows of US$4.7 billion in 2016, while funds running 'fundamental value' strategies saw outflows of US$17 billion and 'fundamental growth' strategies experienced US$11 billion in outflows. Likewise amongst macro strategies, investors allocated over US$6 billion of capital to quantitative, trend-following commodity trading advisor strategies, and redeemed almost US$6.0 billion from fundamental, discretionary strategies that year. As part of this search for alpha, some hedge funds are using dedicated HPC resources, including GPUs, which allow them to parallelise processing tasks to run faster and more efficiently. Using HPC in order to gain a competitive quantitative advantage in this way can offer increased investment returns and support for attracting client investment. Hedge funds can deploy HPC to a greater or lesser extent, to try to identify signals in market data or back test strategies, for example.

Workloads

Whether for risk or alpha generation, the technology needed to handle these challenges depends on the workloads being run. A large compute platform needs some way of spreading the work over a distributed machine.

Andrew Jones, Vice President for HPC Consulting at NAG, says, "There are two classes of workloads; what we call 'embarrassingly parallel', where you could split the workload into a million pieces and none of them have any interaction with each other. For these types of workload you can sensibly use almost any parallel processing interface layer, such as Spark or Hadoop."

Jones explains that the other class is 'closely coupled' workloads, where the million pieces do need to talk to each other on a regular basis throughout the communication - possibly many thousands of times per second. Here, two elements come into play.

"Firstly, your hardware has to have a strong interconnect between the different compute nodes so they can handle all of that data communication, and secondly, you need to choose a parallel programming method that makes it easy for you to deal with that coordination," Jones says. "So you probably wouldn't be using things like Hadoop or Spark, you'd probably be moving more towards traditional scientific tools like MPI (Message Passing Interface)".

The infrastructure needed to run HPC can be complex, so implementing HPC capabilities in a cost efficient and flexible way is crucial. Jones says, "When companies are looking to put in a compute capability, they've really got two choices about it; the first is which technology to use, which comes down to things like GPUs vs. traditional processors. The other side is the delivery mechanism, assessing in-house capability versus managed services, versus cloud."

Determining how the software will interact with hardware is an essential element of this. Selecting the hardware to deploy has a direct impact on both capabilities and cost.

" Using GPUs to run certain types of applications can be more efficient than traditional-based processors at certain types of tasks, typically ones that have a lot of readily accessible parallelism."

Andrew Jones, NAG

"Using GPUs to run certain types of applications can be more efficient than traditional-based processors at certain types of tasks, typically ones that have a lot of readily accessible parallelism", says Jones.

Houston agrees. "The GPU is more of a throughput device, better served for risk computation, massive datasets and so forth". He cites one bank reporting a seven-times return on investment from their GPU work relative to CPU for a particular class of application that is parallel in nature.

Infrastructure & Resources

Developing the expertise to support HPC applications in-house can be challenging, and migrating functionality between different hardware needs to be managed carefully, as it can increase the expense of a project.

The use of third party resources can be invaluable in helping to keep costs down, especially where development of in-house expertise proves challenging.

Jones says "On the skills side of things, there is very much a shortage, spanning all of the parts of the spectrum. There's not only a shortage of people who understand the business aspect of HPC, but also who understand how to programme HPC infrastructures, whether GPU or other types of HPC."

Third-party hosting of both hardware and software via smart relationships with cloud and datacentre providers can not only bring these additional skills into the project, which are otherwise hard to acquire, but help to control costs too.

" Based on conversations with many banks, we've learned that many of them are going through exercises to look at current costs within their IT infrastructure. Where the data centre is located is a key factor."

Stef Weegels, Verne Global

Weegels says, "Based on conversations with many banks, we've learned that many of them are going through exercises to look at current costs within their IT infrastructure. Where the data centre is located is a key factor. For example, TCO analysis in a recent white paper by Citihub Consulting suggests that for wholesale requirements, Iceland may be as much as 50% cheaper over a seven-year term when compared with London and popular North American locations, such as New Jersey and Illinois."

Placing trading engines physically close to exchanges has long been a requirement in order to reduce latency in the trade lifecycle. As the community has become sensitive to reducing costs, however they are increasingly looking at HPC capabilities away from expensive co-located trading environments for latency-tolerant business functions. One such area is the research and testing of algorithms, which, under MiFID II, will impact a far greater number of firms.

Robin Mess, CEO of high-performance analytics provider big xyt, says, "Over the past few years these requirements have spread across from larger market makers and electronic liquidity providers to smaller trading firms."

The availability of analytics, data-management and testing capabilities enables such firms to be successful if they have a viable trading strategy.

" We see hedge funds and trading firms using highly sophisticated strategies driven by heavy data analytics and HPC - including novel approaches like artificial intelligence and machine learning."

Robin Mess, big xyt

Mess says, "We see hedge funds and trading firms using highly sophisticated strategies driven by heavy data analytics and HPC - including novel approaches like artificial intelligence and machine learning. In order to generate alpha, to optimize execution strategies and to reduce costs, more firms rely on capabilities from third parties enabling them to access high-quality market data, to test algos and to continuously optimize them. This accelerates their core business and results in a competitive advantage."

Synchronisation

To conform with MiFID II regulations, not only do algorithms need to be thoroughly tested to ensure they don't disrupt the market, but trading systems also need to be synchronised, which demands that there is traceability back to Coordinated Universal Time (UTC).

"If people are running their algo testing environments in non-latency sensitive environments, they still have to have that accuracy and granularity around time stamps on the data," says **Leon Lobo**, Strategic Business Development Manager for Time and Frequency at the National Physical Laboratory.

Additionally, for HPC to be used effectively for cross-trade and post-trade functions, there is also a need to time-synchronise and link non-latency sensitive applications with those latency sensitive applications such as trading engines. If firms are using distributed HPC solutions to rapidly transfer data in a coordinated way, critical timing becomes important.

" Where distributed data is being used for systems not directly linked to order or trade flow, UTC-traceable time stamps need to be used to synchronise."

Leon Lobo, National Physical Laboratory

When using distributed systems, the user will need to be able to trace back a sequence of updates. Where HPC is used to process high volumes of data at high speed the process is made more challenging.

"With trading systems, for anything to do with order flow or executions it is necessary to trace timestamps to determine what happened, when," explains Lobo. "Where distributed data is being used for systems not directly linked to order or trade flow, UTC-traceable time stamps need to be used to synchronise."

Conclusion

The application of HPC offers real advantages to financial services firms across functions, from risk management to alpha generation, but in building the capability, it is clear that flexibility is needed. Understanding how demand may change - and the implication that has for cost and complexity - requires insight that most firms will not naturally have.

" We have clients with variable demand, which means they may have increased computing requirements at specific times. We have delivered a burstable cloud for them."

Lewis Tucker, Verne Global

"We have clients with variable demand, which means they may have increased compute requirements at specific times," says **Lewis Tucker,** Enterprise Solution Architect at Verne Global. "We have delivered a burstable cloud for them, so 90% of their HPC load is in their normal, traditional private cloud, but if there's a regulatory change on the horizon, for example, we have a flexible HPC solution for whatever duration they require."

Partnering with the right firms for every element – chips, servers, service management and skills – can create an ecosystem for HPC that works across functions and can grow with the business and customer demand.

By working with such trusted partners in a more consultative or collaborative way, HPC provision can be tailored specially for a business, yet with the mutualisation of cost, to help firms comply with ever growing risk and regulatory demands, while giving them a platform to remain competitive.

For more information on the companies mentioned in this article, visit:

www.nvidia.co.uk
www.nag.co.uk
www.big-xyt.com
www.npl.co.uk
www.verneglobal.com

Financial Markets Insights from The Realization Group, is a series of interviews with thought leaders in financial and capital markets. The purpose of the series is to provide exclusive insights into industry developments, through in-depth conversations with C-level executives and key experts from banks, exchanges, vendors and other firms within the financial markets ecosystem. For more information, please visit **www.financialmarketsinsights.com**

Other topics in the series:

Building the future of finance DOWNLOAD

A qualitative approach to Best Execution DOWNLOAD

Under surveillance -
Getting to grips with the new Market Abuse Regulation DOWNLOAD

Agility in Clearing DOWNLOAD

Voice collaboration - what is the future in Financial Services? DOWNLOAD

The Voice of the Future? Flexible trading, lower cost, more value.

New initiatives in compliance technology DOWNLOAD

Digital transformation - when business & technology go hand in hand DOWNLOAD

Are your mobile apps up to the test? DOWNLOAD

The Realization Group is a full service marketing and business development services company specialising in the capital markets. Our team contains industry practitioners from both the trading and post trade disciplines and we have expertise equally in the on-exchange and OTC trading environments. We apply our comprehensive set of marketing programs and wide-ranging media and business networks to complement the business development requirements of our client organisations.