White Paper

Banks Must Expand Their Grid Capacity Fast and Seamlessly or Suffer the Consequences

Sponsored by: IBM

Madhumitha Sathish Peter Rutten
January 2024

EXECUTIVE SUMMARY

Financial institutions are confronted with many factors that influence their ability to operate successfully and profitably, including heightened capital adequacy regulations, escalating geopolitical tensions, global economic fluctuations, mounting interest rates, surging inflation, increasing cost of operations, fewer revenue avenues, sustainability issues, and stringent reporting requirements. Because of these disruptive factors, banks have an increasingly urgent need for high-volume, real-time data processing based on a variety of grid workloads for financial services. But for line-of-business (LOB) staff, having unimpeded access to grid computing for real-time, performance-intensive data processing that is performant, reliable, secure, compliant, and cost efficient is not a trivial undertaking. What is more, providing such comprehensive grid computing is becoming increasingly difficult in an organization's own datacenter.

The financial sector is also subjected to a constant stream of new government policies in the form of new or expanded regulations and tax policies. For example, regulatory entities such as the Financial Conduct Authority (FCA), Securities and Exchange Commission (SEC), the Monetary Authority of Singapore (MAS), or the Financial Industry Regulatory Authority (FINRA) have imposed significantly stricter mandates for establishing rigorous and robust transaction monitoring systems in foreign exchange (FX), fixed income, and equities activities, among others. These regulations and monitoring systems also apply directly to the data being leveraged in grid workloads, further complicating their processing.

As a result, there is sustained pressure on margins within the financial industry, with many banks taking a hit on their return on equity as they struggle to adjust their grid computing capabilities to these external factors. The sections that follow provide an in-depth look at how innovations in grid computing can alleviate these pressures. This paper also looks at IBM's Cloud HPC (high-performance computing) grid offering, which IDC believes meets the many demanding requirements for grid computing in the financial sector.

January 2024, IDC #US51765124

SITUATION OVERVIEW

Data

The Need for Real-Time Data Processing

There is a growing need for managing and processing ever larger volumes of real-time data in the industry to make increasingly more accurate trading decisions. For example, LOB staff at all banks are under pressure to improve their real-time data gathering, normalizing, and distributing mechanisms; risk management; and data analytics to enable them to engage in higher-risk, higher-margin trades. Therefore, the need for large-scale risk modeling is growing, leading to a proliferation of financial services grid workloads such as mortgage analytics, actuarial analysis, model backtesting, value at risk (VaR), and Monte Carlo.

At the same time, the complexities of tackling these different grid workload patterns are increasing. Grid applications and workloads require different scheduling paradigms with different requirements for compute, scalability, cost, and time to results. Recently, this complexity has been further exacerbated by the infusion of AI into predictive computing, including the development of AI data sets. While grid computing is essential for doing business in the financial sector, it is also expensive, not just when in use but also, and especially, when not fully in use outside of peak demand. Banks are under pressure to implement operational efficiencies with their grid computing to reduce costs.

Data Management Challenges

Data management costs continue to increase as operations become more complex because data is gathered from a multitude of sources such as banks, exchanges, fintech data aggregators, and central counterparty (CCP) clearing houses. In addition, the banks' legacy systems hold large amounts of data that need to be converted to specific formats before it can be processed with modern analytical tools. Data gathering, normalizing, and distributing from all these sources are imperative for ensuring full visibility on the market. It is also critical for enabling informed decision-making by a wide range of stakeholders — internal clients, external clients, and regulators — to whom the data needs to be distributed at varying delivery frequencies and in a variety of formats. What is more, all of this data is highly sensitive and must be protected, even as unforeseen events roil the industry. New workforce dynamics caused by the COVID-19 pandemic, for example, with many financial sector employees suddenly working from home, have exacerbated banks' cybersecurity vulnerabilities and forced them to be even more prepared for unexpected challenges to their security stance.

Regulatory and Compliance

Challenges with Regulatory Requirements

Stringent rules and operating under great scrutiny are the new normal for all banks. With the latest updates on the Basel III Endgame (stricter bank capital requirements aimed at ensuring the stability of big banks) and the proposed July 2025 deadline for compliance, the banks have very little time to reengineer their strategies for better business outcomes and also leverage this opportunity to modernize their capital infrastructure by investing in new technologies and processes. The need for extreme resilience to regulatory changes is typically high in the industry, as banks are constantly subjected to new regulatory changes, which in some instances even means they must be prepared to lose their autonomy to some extent. This growing complexity of the regulatory landscape has led to margin pressures.

LOBs at the banks are forced to learn about, assess, and react to the dynamism in the regulatory environment to remain compliant. To this end, they need predictive modeling to assess a situation early on and have a set of potential business-favorable amendments in place to keep risk low. This area of investment is significantly capital intensive and requires technology expertise to redefine, analyze, improve, and control the processes that can deliver cost efficiencies to doing business and at the same time ensure compliance. A portion of profits is therefore reinvested for compliance. Technology and related services are the areas that these investments are being channeled into — as a result, fewer funds are available for business innovation, greater agility, and improved efficiency.

Challenges with Compliance and Risk Management

Banks need to make important decisions about the extent to which they should automate compliance processes to run a seamless compliance practice while ensuring that they are fully accountable with their reports. These considerations extend to questions around the benefits of growing in-house skill sets for compliance and risk management versus outsourcing some of the same.

Technology

Challenges with Managing Legacy Technology Assets

CIOs are experiencing challenges with technology management as they source newer technologies from a variety of vendors to modernize their operations. Meanwhile, their legacy infrastructure continues to demand significant capital investment and staffing resources to keep it running.

Born-in-the-cloud fintech sector companies do not have to deal with the burden of maintaining legacy infrastructure and, therefore, have an edge over the older, larger organizations with regard to allocating resources to innovation and optimizing for flexibility and agility. This competitive pressure forces the large banks to work toward more agility.

The Growing Adoption of Digital and Quantum Technologies

Banking operations have been going through profound changes as banks have embraced digital transformation and made customer centricity a priority. The growing adoption of digital technologies, with an increased infusion of traditional predictive AI and generative AI, and initial experiments with quantum computing have further intensified the competition for already strained compute, labor, and budget resources within the organization.

Fluctuating Demands for Grid Computing Capacity

Several critical use cases in the banking sector require elastic and scalable grid computing, a form of distributed computing composed of many networked, loosely coupled computers acting together to perform large tasks. IDC considers grid computing as a category of performance-intensive computing (PIC), which requires different, more performant, and typically more parallelized computing and storage infrastructure, as well as high-throughput computing software scheduling that orchestrates millions of tasks with varying priorities and deadlines.

Examples of use cases that require grid computing include high-frequency trading, algorithmic trading, end-of-day and intraday risk management, developing and maintaining electronic trading platforms, enhancing security solutions, integrating AI into risk assessment, developing and leveraging customer insights, and building and executing complex financial models. Whether it is big data and analytics for predictive analytics, blockchain technology, or generative AI, all these critical workloads put significant

pressure on an organization's distributed infrastructure and, in turn, the grid, which is a subset of that infrastructure.

There is also a surging need for banks to model customer behavior and leverage new PIC technologies such as HPC, AI, big data, analytics, and even quantum or quantum-inspired solutions to predict risk and trading patterns. These technologies require vast infrastructure resources as well as IT, data science, AI engineering, and quantum development skills. Furthermore, accurate and expert financial modeling solutions are imperative for banks to be able to better respond to other forms of disruption, including wars and geopolitical tensions, financial crises, market volatility due to elections, and new financial regulations.

Finally, PIC server processors and accelerators used in grid computing are becoming increasingly power hungry. The rising power density per rack has resulted in increased energy consumption and higher energy costs. Power-hungry PIC infrastructure can run afoul of increasingly mandatory sustainability policies, thus putting pressure on banks to optimize, if not reduce, their energy to keep costs down and stakeholders happy. What is more, many older bank datacenters are not designed to support the power profiles of modern HPC infrastructure.

An important part of optimizing banking applications for performance is fine-tuning them on the server processors of the grid environment that they run on, but such processor-based optimization can also become an impediment to moving the applications to another environment — for example, the cloud — as the applications will then have to be recompiled as well as tuned or optimized to run equally well in the new environment.

Challenges with Grid Bursting and the Public Cloud

To bring more grid capacity into the organization, banks often use grid bursting — moving some of the infrastructure capacity needed to the cloud to achieve a hybrid computing model. However, grid bursting to a cloud comes with its own challenges because few cloud providers act as a one-stop shop for raising and resolving all possible technical glitches and providing effective and timely troubleshooting support. Typically, banks find themselves dealing with several vendors that are involved in the grid bursting implementation for resolving issues. This adds to the complexity of operating the cloud portion of a grid and can severely prolong the time it takes to solve technical problems — time that banks do not have. Periodic service changes by the cloud service providers, and frequent changes to the APIs, add to these difficulties with multiple vendors that are responsible for a hybrid HPC grid. In a typical bank business, such glitches and time delays in resolving issues can cause severe trade losses, P&L woes, and regulatory noncompliance.

While there are existing grid-based financial applications, transitioning them to cloud computing is not always a straightforward task. Furthermore, due to a range of privacy and data ownership concerns, some public clouds can be unsuitable for the financial sector. Therefore, an HPC grid deployed as an on-premises private cloud, with a seamless transition between on premises and cloud whenever required, would be the logical choice. It is recommended that the on-premises and cloud portions are serviced by the same vendor for instant troubleshooting, which is a critical requirement for the financial sector.

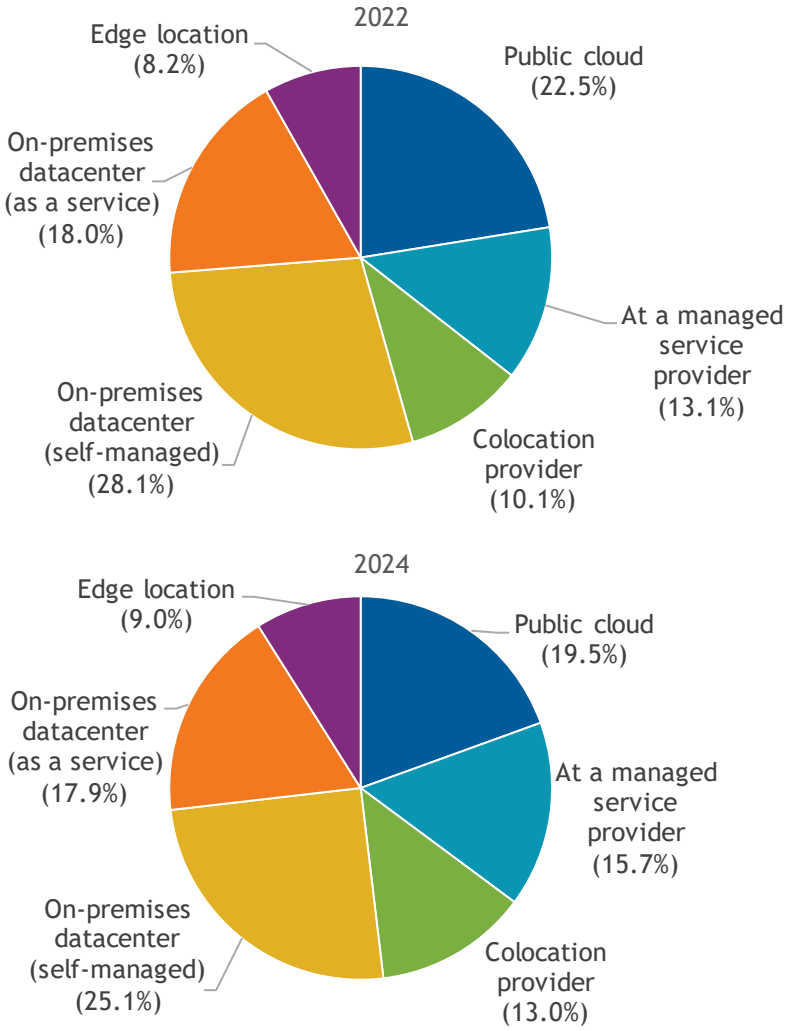
As the performance-intensive compute requirements of the top-tier banks expand, banks are complaining about a growing supply/demand mismatch of available cores at the public cloud service providers to serve them. As a result, banks continue to struggle with issues of scalability, cost, and whether to adopt dedicated versus public cloud. An IDC survey of HPC end-user sentiments

conducted in 2022 shows that the financial services sector does not consider the public cloud a preferred deployment scenario for HPC. Data security constraints, the cost of additional cloud data storage and compute, and data compliance requirements are cited as the top reasons by financial services organizations for deploying a smaller percentage of their HPC environment in the public cloud. The on-premises datacenter was the most preferred deployment scenario (see Figure 1).

FIGURE 1

HPC Environments in the Financial Services Industry

Q. *What percentage of your organization's HPC environment is deployed in the following locations now and will be deployed in the following locations in 24 months?*



n = 149

Source: IDC's *HPC Survey,* August 2022

Another complication is that it can take several years to certify a public cloud due to the complexities of integrating the public cloud with the organization's HPC environment. IDC's *HPC Survey* revealed that more than 45% of the respondents cite integration of the cloud with the organization's HPC environment as complex, and a majority of the respondents attribute this to cost, performance, data security, latency, and data compliance issues. A lack of expertise to implement such integrations also plays a major role, as does the fact that the traditional governance process associated with public cloud is not only expensive but also time-consuming – it can take 12–36 months for approval depending on the size of the firm.

On the other hand, SaaS-like solutions just take a period of 3–6 months to be approved by the line of business through the third-party risk assessment process. Quick approval by the line of business is a major advantage for banks, all of which need faster access to suitable and sufficient infrastructure to optimally run their workloads and sustain their competitive position. They are already under pressure to operate with smaller margins, and therefore the growing need for risk modeling and the related infrastructure access needs to be addressed in the shortest possible time.

The Opportunities with Grid Computing as a Service

Reengineering the organization's technology architecture is challenging, but adopting new digital technologies for remaining competitive is more necessary than ever. Even as margins are challenged by the regulatory environment and by the need for PIC technology, banks continue to invest in high-end applications and security, as these can stand out as differentiators in their businesses.

In summary, this means that banks can greatly benefit from faster solutions; instant elasticity, built-in security, and seamless scalability; and a one-stop shop for all their technology needs. Grid computing made available as software as a service by a service provider that owns the entire infrastructure and software stack and provides single-vendor technical support can resolve many of the abovementioned challenges.

CONSIDERING IBM CLOUD HPC

The IBM Cloud HPC offering addresses many of the challenges laid out in the previous sections. The service includes an IBM-owned and fully integrated stack; the scheduler, storage, and encryption software; the computing infrastructure (IBM Cloud); and HPC expertise. By being automatically provisioned and expanded with a variety of IBM solutions, IBM Cloud HPC enables fast provisioning and seamless operation of a hybrid grid. Based on IBM Spectrum Symphony, the HPC Grid resource and workload scheduler software has been serving the financial services industry for more than 20 years and is considered to be the "gold standard" for grid computing in financial services.

To deliver IBM Cloud HPC, IBM Symphony is integrated with IBM Storage Scale, IBM Security Guardium, IBM Cloud Infrastructure, IBM Consulting, IBM performance engineering, and IBM support. In more detail:

- **IBM Spectrum Symphony** software delivers enterprise-class management for running compute-intensive and data-intensive distributed applications on a scalable, shared grid. It accelerates dozens of parallel applications for faster results and better utilization of all available resources. IBM Spectrum Symphony provides support for a hybrid HPC cloud, enabling workloads to be forwarded to multiple clouds and data to be automatically staged to or from the cloud. Further, resources that are consumed in the cloud can be automatically scaled up and down based on workload demands and scheduling policies. The resource-

sharing model in IBM Spectrum Symphony helps reduce infrastructure expenses and management costs. It enables deployment of multiple heterogeneous applications on the same shared grid. At the same time, line-of-business ownership can be preserved while delivering service-level guarantees. Dynamic HPC cloud support enables organizations to intelligently use cloud resources based on workload demand, with support for all major cloud providers. IBM Spectrum Symphony is available for automated deployment on IBM Cloud.

- **IBM Storage Scale** is an enterprise-grade high-performance file system (HPFS) that delivers scalable capacity and performance to handle demanding data analytics, content repositories, and HPC workloads. The IBM Storage Scale architecture allows it to handle tens of thousands of clients, billions of files, and petabytes of data written and retrieved as files or objects with low latency. Optionally, IBM Aspera can be used for high-speed data movement between on premises and the cloud, using the Fast and Secure Protocol (FASP).
- **IBM Security Guardium** is a family of data security software in the IBM Security portfolio that can adapt as the threat environment changes, providing complete visibility, compliance, and protection throughout the data security life cycle.
- **IBM Security and Compliance Center** is an integrated solutions suite to define policy as code, implement controls for secure data and workload deployments, and assess security and compliance posture across hybrid, multicloud environments.
- **IBM Cloud with IBM's Virtual Private Cloud (VPC)** enables businesses to create their own configurations for compute instances, high-performance storage, and networking like public gateways, load balancers, and routers.

THE BENEFITS OF IBM CLOUD HPC

Faster Time to Market

IBM Cloud HPC improves time to market with HPC software solutions to configure, deploy, and burst powerful HPC workloads with complete automation and an integrated user experience. Since IBM manages the environment for the client with agreed SLAs, time to market is accelerated and the approval process significantly shortened. This managed services option makes the experience SaaS like for the organization, as the responsibility rests on IBM. It allows LOBs to get into and out of new markets faster with significantly reduced time to procure infrastructure for emerging business opportunities. IBM Cloud partners with a client organization to run through its third-party risk assessment and tune the applications and cloud HPC environment for better performance, efficiency, cost clarity, and predictability.

Governance Support

By owning the entire stack, IBM can coordinate the release of newer versions of the IBM Cloud stack and libraries that are suitable for the infrastructure. This makes consumption of the integrated software stack seamless and significantly more time effective within the stringent guardrails of cloud governance. Top-tier firms that find regulatory overhead a major restraint for migrating to the cloud can reap the benefits of an HPC service made up of IBM-owned infrastructure, software, and HPC expertise team.

As a cloud provider to large enterprises in the financial industry with sensitive data and regulatory requirements, IBM offers *IBM Cloud for Financial Services,* which is focused on governance and controls, with built-in security measures that streamline the risk management process. The offering helps organizations reduce risk and expedite the adoption of cloud services, even for their most

sensitive workloads. IBM offers a hybrid cloud approach that complements and future-proofs organizations' on-premises investments. It also offers a seamless cloud-bursting experience during workload peaks. Finally, the service helps speed up innovation with built-in controls to address security and compliance needs and improves agility, thanks to various hybrid cloud deployment options.

IBM Cloud emerged as the Trust Leader in the latest *IDC Trust Perception Index* (source: *Analysis of the Future of Trust Survey,* February 2022), which measures the four foundations of trust: security, privacy, compliance, and environmental, social, and corporate governance (ESG). Respondents valued information and guidance on security, privacy, and compliance adherence as IBM Cloud's biggest strength in addition to good governance, best practices, accountability, and responsibility. The latter was the result of IBM's commitment to increasing the ratio of value- and ethics-based decisions versus cost-based decisions.

Latest Processors and Accelerators for PIC Workloads

IBM offers a complete portfolio of integrated HPC solutions for hybrid cloud that provides flexibility to manage compute-intensive workloads on premises and/or in the cloud, using 4th Gen Intel Xeon Scalable (Sapphire Rapids) processors. IBM Cloud HPC also uses the latest GPU accelerators to power HPC and AI with bare metal and virtual servers, allowing banks to mix and match compute resources to meet different business or application needs.

Decades of Financial Domain-Specific Experience and HPC Grid Expertise

In today's business world, in which organizations are subjected to a variety of disruptive forces such as geopolitical tensions, global trade disputes, energy shortages or price fluctuations, environmental disasters, political extremism, destabilizing elections, and other business-impacting events, having immediate access to PIC infrastructure for modeling is critical. IBM Cloud HPC enables such fast access with a SaaS-like experience through its managed services offering model. IBM Cloud HPC can provide fast access to a very large number of cores while delivering:

- End-to-end provisioning and support of the infrastructure
- A complete technology stack of CPU, memory, storage, I/O, network, OS, integrated software
- Access to trained human resources with deep expertise for running HPC workloads (IBM has provided HPC expertise for more than three decades and is also a pioneer in AI and quantum computing, which are increasingly relevant for the financial industry.)

Zero Application Migration, IBM-Owned Integrated Hardware and Software Stack, and Single Point of Contact for Support

Today, IBM Symphony runs in the datacenters of many banks. By adding IBM Cloud HPC for their grid capacity needs, organizations can take advantage of a seamless transition on IBM Symphony from one deployment scenario to the other – the customized IBM Cloud HPC platform enables easy application migration. By offering HPC expertise to tune an organization's applications, single-vendor support for the entire IBM Cloud HPC stack, and one point of contact to triage and resolve issues, the offering removes the uncertainty of timely resolution of technical glitches that can occur when multiple vendors are involved. Many businesses consider this their biggest pain point, as an unresolved technical problem can expose them to risk that is measured in millions of dollars, especially near market closing time.

CONCLUSION

Capital markets are mission critical and time sensitive, intraday risk management must be fault tolerant and secure, and the infrastructure, software, and technology blocks that deliver grid computing must be always reliably available. This is true for most performance-intensive computing workloads, whether emerging workloads such as AI, ML, and LLM or well-established workloads such as Monte Carlo; all of them are extremely time, security, and latency sensitive. These workloads benefit from being hosted by a single vendor that owns and integrates the entire stack, provides all the aspects of support, and is accountable for the performance, availability, speed, security, resilience, and redundancy across geographies. IDC believes that IBM's Cloud HPC offering provides exactly that, meeting many of the demanding requirements of the financial industry (especially if the end user is already using IBM scheduler software Symphony and/or LSF) while greatly reducing the risks associated with grid computing.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com