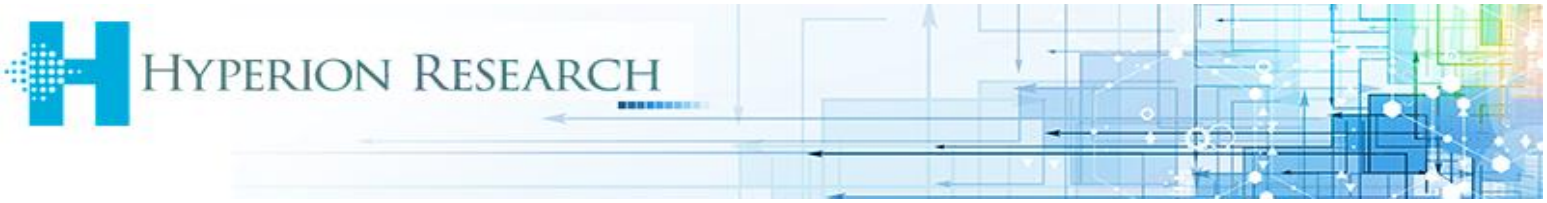

White Paper

The Evolution of HPC Includes Strong Use of Hybrid Cloud

Sponsored by: IBM

Mark Nossokoff and Tom Sorensen
October 2023

HYPERION RESEARCH OPINION

There has been a decided shift in the tone of public discourse relative to utilization of HPC resources in the cloud over the last 12-24 months. Where it previously centered around whether cloud-based resources could reliably and performantly be utilized for complex HPC workloads, it has evolved to understanding which workloads are capable of being migrated to the cloud to meet a user's cost, performance, security, and support requirements. This is inclusive of both traditional HPC modeling and simulation workloads and modern AI/HPDA/LLM workloads.

The most widely adopted framework for leveraging HPC resources in the cloud is in a hybrid environment where a user runs their HPC workloads both on-premises and in the cloud. For purposes of this paper, hybrid refers to each of the following scenarios:

- Sites running jobs both on-premises and in the cloud.
- Application workflows that complete portions of a single job both on-premises and in the cloud.

Many options exist for users to choose from for their hybrid cloud requirements, ranging from large, public cloud service providers (CSPs) serving many markets, to HPC-market specific CSPs, to system vendors managing and provisioning their own private and public clouds. Each has different strengths that may be more applicable to certain workloads than others. Strengths can range from availability of the most recent CPU or GPU technologies, access to their own custom hardware and software, domain area expertise in various business and research areas, and tools that assist users to optimize utilization of HPC cloud resources.

Based on its legacy leadership operating in HPC environments, its breadth of IP across the HPC software stack, its depth of service capabilities, and investments in emerging technologies, IBM and its IBM Cloud HPC is worthy of consideration for users exploring utilization of the cloud for their HPC workloads, particularly if the user is already familiar with IBM's HPC software stack on-premises, including its LSF or Symphony schedulers.

SITUATION OVERVIEW

HPC Cloud Market Snapshot

HPC in the cloud has been experiencing accelerating growth over the last few years. Consumption of HPC public cloud resources is projected to grow to $14 billion in 2027, with several verticals embracing the use of cloud resources more warmly than others. Biosciences and CAE/Manufacturing sites are the largest verticals relative to spending in clouds today. CAE/Manufacturing, Economics/Financial, and Geosciences are expected to see the largest cloud spending growth. See Table 1.

Consumption of HPC public cloud resources is projected to grow to $14 billion in 2027

TABLE 1

HPC Cloud Forecast by Vertical 2021-2027

($M)	2021	2022	2023	2024	2025	2026	2027	CAGR 22-27
Bio-Sciences	$1,439	$1,557	$1,847	$2,097	$2,338	$2,514	$2,821	11.8%
CAE/Manufacturing	$957	$1,289	$1,587	$1,896	$2,242	$2,616	$3,017	22.3%
Chemical Engineering	$128	$154	$189	$224	$264	$306	$267	19.0%
DCC & Distribution	$289	$347	$425	$505	$593	$687	$732	18.9%
Economics/Financial	$315	$398	$512	$620	$745	$882	$1,041	22.9%
EDA	$372	$443	$538	$632	$737	$846	$975	17.8%
Geosciences	$328	$403	$505	$614	$746	$890	$1,069	22.1%
Mechanical Design	$24	$27	$33	$37	$40	$45	$45	13.3%
Defense	$391	$471	$578	$689	$813	$946	$1,091	19.3%
Government Lab	$364	$443	$548	$697	$872	$1,086	$1,351	24.4%
University/Academic	$242	$276	$322	$364	$407	$464	$549	13.9%
Weather	$140	$184	$246	$318	$408	$537	$675	30.8%
Other	$111	$141	$187	$239	$304	$381	$436	28.0%
Total	$5,100	$6,132	$7,516	$8,931	$10,510	$12,197	$14,069	189.1%

Source: Hyperion Research, 2023

Recent studies indicate that approximately 60% of all HPC sites surveyed indicated that they are currently using the cloud, with over 70% of respondents in the industry sector employing cloud resources. While that reflects a large percentage of sites leveraging the cloud, it's not fully reflected in the amount of workload runtime users run in the cloud. 29% of workload runtimes were run in the cloud for those sites who indicated using the cloud, while 37% of the surveyed industry sites' workload runtimes were run there. In addition, there is ample room for growth.

Drivers and Barriers to Adoption of HPC Cloud Resources

Drivers for HPC Cloud Resource Adoption

Users consistently cite several items that are driving their increasing utilization of HPC resources in the cloud:

- **Cost effectiveness:** Users whose workloads run more periodically or infrequently without stringent completion-time requirements and without heavy data movement needs often find running workloads in the cloud more cost effective than on-premises.
- **Extra capacity for surge workloads**: Many HPC users have access to adequate on-premises HPC infrastructure. Still, there are times when the resources aren't immediately available for critical modeling/simulation runs, and resources are immediately available in the cloud.
- **Access to resources not available on-premises:** In many cases, users are now keeping on-premises machines up to 5 years, often causing them to miss a generation or two of the latest technology. Cloud resources allow them to access the latest innovations for critical workloads, as well as evaluate new technology before acquiring it on-site.

Barriers for HPC Cloud Resource Adoption

As you may expect, users have also been consistent with items that are barriers for them to migrate their HPC workloads to the cloud. The top barriers cited are:

- **Cost effectiveness:** Interestingly, cost effectiveness is also cited as the top barrier, in addition to the top driver. While cloud utilization may look cost-effective when running smaller, highly parallel jobs and test cases in the cloud, users have found costs from running their workloads at scale in the cloud far exceed what the test case projected. In large part, costs include many things like unexpected charges for data movement within and data egress from the cloud. Often the higher costs are driven by reduced performance when running larger tightly-coupled and highly computational jobs in the cloud.
- **Security:** While this item is seen to be a decreasing limitation for adoption of HPC cloud services, it does remain a concern. Many users view their data as critical IP to their businesses and are concerned about bad actors gaining access to it, as well as potential inadvertent access in multi-tenant environments. A related issue is also compliance with regulations or corporate requirements and for certain data to remain in certain geographic areas.
- **Data locality:** Data creation is occurring in a wide variety of places, including at the edge (e.g., smart city sensors, real-time data capture during high performance auto racing, large arrays of astronomy antennae, manufacturing line visualization quality control) and in multiple locations. Running the computing analysis near the data creation points, as opposed to moving the data to computing elements on-premises or in the cloud, may be more cost effective and more performant than moving data to the compute resources in the cloud.

Trends in HPC Cloud Services and Payment Models

Broad Range of Service and Models

There are multiple types of services available to users for cloud-based HPC resources. The common element between the various types of services is a consumption-based, "pay as you go" OPEX business model. Various examples include:

- **Full service CSPs:** Cloud service providers offering a wide range of performance capabilities across many different markets to numerous different companies.
- **Industry-focused:** Cloud providers offering resources focused on a specific industry or vertical, such as energy or media and entertainment.
- **Capability-focused:** Cloud providers delivering an optimized set of resources for several markets, such an HPC cloud supporting multiple industry verticals.

- **Private clouds:** Infrastructure placed on-premises at a single organization but managed by the system provider and billed based on consumption OPEX, as opposed to up-front CAPEX.

Each of the above can also support various service models, ranging from bare metal systems which a single tenant user can deploy with their own preferred operating environment to fully virtualized or containerized with the CSP providing everything, including operating environment, application code licensing, domain-area expertise and assistance.

Hybrid Cloud, Multi-cloud, and Native Cloud

Utilization of hybrid cloud provides the best of both on-premises and cloud worlds. It affords users the capability to augment their on-premises HPC capabilities with specific resource requirements in the cloud with the lowest possible risk.

Multi-cloud provides even greater flexibility. Different CSPs have different strengths, and the ability to have portable code that can run on multiple CSP platforms allows users the capability to select best-of-breed for a wide range of workloads. Organizations with concentration risk concerns may also find multi-cloud solutions of interest.

Native cloud deployments are growing, particularly with newer startups and technologies. New businesses have limited budgets for capital investments and the cloud provides the capability for business to grow until they reach a scale where it may be feasible to bring resources on-premises.

Emerging Technologies

As new technologies and innovations occur, users want to evaluate the technology prior to deciding what to procure on-premises. Many times, CSPs are first to market with these technologies (e.g., having the most recent GPUs), and may also receive priority access to new technologies in supply-constrained environments.

Still other emerging technologies, such as AI and quantum computing, require very high capital investments that make it economically unfeasible for users to bring on-premises for evaluation or code development. Vendors developing these advanced technologies may also only initially provide access via cloud-based capabilities.

The IBM Cloud HPC Approach and Benefits to Hybrid HPC Cloud Services

As the landscape of cloud resource products expands and diversifies, users are seeking tools to more easily and optimally use and manage those resources in an easy, flexible manner to achieve faster time to results for their HPC-based engineering jobs and research. The IBM Cloud HPC is an example of a flexible, cost-effective, and highly-scaling cloud resource, and is uniquely suited for occasional bursting, hybrid computing, or a full cloud native experience. It is powered by an impressive wide-ranging portfolio of IBM IP to choose from, including:

- Software stack: LSF, Symphony, OpenShift, Storage Scale
- Virtual, containerized, and bare metal provisioning
- x86 instances, GPUs and IBM Power
- Security tools
- Managed services

The integration of these solution components and services enables users with tools for scheduling, containerizing, storage management, and more all while remaining in one cohesive workspace. The ability to quickly and easily access this suite of tools can provide considerable ease-of-use and increase time-to-solution.

Properly Matched Performance Requirements

One major consideration for users is properly matching the performance requirements of their workload with available cloud resources. Workload management platforms like IBM Spectrum LSF are specifically designed for HPC deployments, provisioning and configuring compute resources with optimization and scaling in mind. With auto scaling and the ability to automatically add and remove nodes based on workload specification, there are also considerable cost benefits in using this kind of advanced workload management platform.

LSF is a leading commercially-adopted workload management platform within the HPC community. For current users of LSF in their on-premises environment, migrating their workloads to IBM Cloud HPC should be almost seamless.

Cost Management

On the topic of cost management, studies indicate that organizations waste around 32% of the spending that goes into cloud services. Paying for unused time, overprovisioning, and under-optimized scheduling can end up costing significant capital to users and organizations. Continuous awareness and reactiveness to shifting compute, storage, and database configurations for cloud applications is nearly impossible for users without some sort of automated optimization tool. These tools, such as IBM Turbonomic, which generates actions that optimize cloud VMs, volumes, database instances, paths, Kubernetes, even discounts and RI utilization and coverage, leverage AI to fluidly adapt to changing workload needs. Tools like these, when combined with the consumption-based pricing of the IBM Cloud HPC allow for fully supported workloads and burst periods with a significantly reduced risk of overpaying and overprovisioning. Furthermore, IBM Cloud HPC does not charge for data traffic within or between IBM-Cloud data centers, providing an additional measure of cost effectiveness for IBM Cloud HPC clients.

Accelerated Time to Results

Provisioning, cost-management, and scheduling tools integrated within a cloud computing environment can ultimately provide users with an expedited time-to-solution by deploying HPC clusters in significantly reduced time, scaling with on-demand capacity on the cloud, and managing logistical issues associated with on-boarding and compliance.

The ability to quickly provision and deploy intensive HPC workloads without deep infrastructure knowledge lightens the load for users looking for a quick, powerful, and easy solution. Accelerated time to results can be considered a force multiplier in optimization and production. Not only does it reap the benefits of a project completed early, but it can create time and room for more projects, user time, and budget freedom.

Security

In the landscape of advanced cloud compute services, security is mission critical and consistently cited by users as a concern for them as they consider migrating HPC jobs to the cloud. Hybridizing a traditionally on-premises workflow to include cloud resources or entirely migrating to a cloud environment can be a demanding and risky undertaking for an organization.

To help alleviate this risk and allow for smooth transitions and hybridizing, IBM offers Cloud HPC Infrastructure as a Service (IaaS) for building HPC environments using IBM's Virtual Private Cloud (VPC). This type of service enables users to create their own configuration with diverse compute instance type options including the highest levels of security and encryption with FIPS 140-2 Level 4. IBM Security Guardium, a suite of data security tools, offers end-to-end compliance, risk, automation, and monitoring to all users within this ecosystem. With integrated threat management, data protection, user identity access management, and multi-cloud support, IBM Cloud HPC equips users with security tools worthy of any enterprise or production application.

Managed Services

The process of migrating enterprise and production workflows to the cloud is not simply about technical specification and system requirements. Users can find themselves facing difficult privacy and security questions or jumping through time consuming and unexpected compliance hoops during on-boarding. IBM Cloud HPC can include IBM managed services to offload much of this onus onto IBM as the service provider, allowing users to focus on their priority: the application. IBM Cloud for Financial Services is one offering within this ecosystem with the express function of helping clients mitigate risk and accelerate cloud adoption for the most sensitive workloads. Compliance, adherence to security standards, and the assurance of responsible data management can become complex problems for users ultimately resulting in massive time and value loss. Cloud service providers with the capabilities to shoulder these responsibilities can, in many cases, liberate users from these complex, costly pitfalls.

IBM Cloud HPC Customer User Case: Cadence Design Systems, Inc.

Cadence Design Systems, Inc. recently began using IBM Cloud HPC to support development of chip and system design software. Chip and system design of this kind requires innovative solutions, powerful compute resources, and advanced security support. The semiconductor design and manufacturing industry has been challenged in recent years by increasingly complex designs, demanding requirements to lower silicon power consumption and increase global competitive investment, among other pressures. In response and with support from the US Chips Act, the industry is making efforts to deliver secure, timely, and reliable results as a matter of global importance within the entire ecosystem. Tarak Ray, VP and CIO at Cadence, has this to say about leveraging IBM Cloud HPC as part of a multi-cloud, hybrid environment with IBM Spectrum LSF as the scheduler: "The IBM cloud and services has allowed us to achieve, high-compute utilization, which lets us more efficiently utilize our cloud budget and streamline our computational workload."

"The IBM cloud and services has allowed us to achieve, high-compute utilization, which lets us more efficiently utilize our cloud budget and streamline our computational workload."

This holistic modernization of HPC workload management encompasses not only provisioning compute resources but can support virtually any step in the lifecycle of an HPC workload from compliance to storage offerings. This assures users the ability to see improved time-to-solution, cost-reduction, and performance improvements wherever support is most needed. The improved time-to-solution and streamlined workload management reported by Cadence Design Systems is a shared experience among enterprises leveraging IBM Cloud HPC.

FUTURE OUTLOOK

Successfully running applications with modern HPC is about much more than raw compute power. Not only is there a diversifying landscape of hardware and machine configurations available for optimal efficacy, there are mounting infrastructure, policy, and expertise factors that contribute to user requirements. Remotely accessible cloud resources have made sourcing diverse tools easier, but with this availability has come many challenges for users.

Even when application types can be optimally accommodated by one of the many configurations offered by modern CSPs, there are often other requirements necessary to prevent mismanagement of access, unexpected costs, and unanticipated schedule impacts (or mismatched requirements). All of these issues can be compounded in the process of migrating traditionally on-premises workloads to a hybrid cloud or native cloud environment. Avoiding these pitfalls, staying up-to-date on IT developments, knowing how and when to scale, all while conducting advanced scientific and engineering applications, is an incredibly taxing set of demands on users.

When users are able to leverage cohesive and robust hybrid cloud services via platforms like the IBM Cloud HPC, many of the most difficult demands are lightened. With many HPC users looking to a hybrid cloud compute environment, the topics of migration, remote access, and data management have been receiving a lot of attention. These issues, while normally trivial or routine in an on-premises environment, can cause tremendous headaches for new cloud users. Hybrid cloud via IBM Cloud HPC offers managed services that can avoid these pernicious situations before they result in stoppages and lost time.

The adoption of hybrid cloud resources can provide users with access to almost any scale of resources they require, technical support to utilize those resources, flexibility in sourcing from best-in-class CSP infrastructure, and a consistent, predictable OPEX cost structure. Hybrid clouds can be a viable complement and, in some cases, an alternative to traditional on-premises HPC infrastructure deployments for many users and HPC workloads. Users should continuously evaluate their HPC resource needs and determine the appropriate balance between their workload requirements, the infrastructure required to run them, and available budgets. With high-performance computing resources and a set of infrastructure tools and services, users of IBM Cloud HPC can be equipped with what is necessary to keep optimization high, overpaying low, and allow flexibility to adapt to shifting application needs.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com