HYPERION RESEARCH

Technology Spotlight

HPC and the Cloud – A Strong and Maturing Relationship

Sponsored by: Amazon Web Services and NVIDIA

Alex Norton, Mark Nossokoff, and Earl Joseph
November 2022

HYPERION RESEARCH OPINION

High performance computing (HPC) in the cloud has become a standard approach for most organizations consuming HPC resources, with many sites deploying HPC compute capabilities in a hybrid fashion, leveraging both on-premises systems and cloud instances. Until recently, the cloud was used sparingly by HPC sites, employed mainly to address surges in workloads that were highly parallelized, and usually not for most critical production workloads. Over the past few years, extensive amounts of time, money, and effort have been invested by the cloud service providers (CSPs) and users alike, to create solutions and develop expertise that can address the varied and complex needs of HPC workloads in the cloud. The requirements and demands that this community of scientists, engineers, and researchers place on the overall HPC ecosystem are driven by both traditional modeling and simulation workloads, as well as the emergent artificial intelligence (AI) workloads, inclusive of machine learning (ML).

By leveraging cloud computing resources, HPC sites are often looking to reduce capital expenditures for large on-premises clusters and avoid being locked into one generation and one type of technology. Taking advantage of cloud resources, users pay only for the type of technology required at any point in time and only for the duration of its consumption, alleviating underutilization of an equivalent fixed-cost on-premises asset. Further seeking to reduce their overall HPC cost expenditures, users can dynamically optimize their cloud infrastructure to match their workloads' requirements with the help of cloud service providers' tools and expertise.

Recent studies by Hyperion Research indicate nearly every HPC site is either already using the cloud or investigating the cloud to address their HPC workloads. The cloud market is expected to grow at a rate faster than any segment of the on-premises server market, and ultimately become the second largest portion of the broader HPC market behind technical servers, from a revenue perspective. The rest of the broader HPC market tracked includes add-on storage solutions, middleware, applications, and maintenance services.

In recognizing the market need and seeking to address the challenges that on-premises HPC users are facing, AWS and NVIDIA have collaborated to provide performant and easy to use solutions that enable HPC users to unlock the potential of their workloads. Long queue times can be minimized or eliminated for HPC users running certain workloads on AWS' large scale, flexible HPC cloud services. Users can also explore new tools and applications with the help and guidance of experts at AWS and NVIDIA, which in turn allows these scientists, engineers, and researchers to devote more time towards exploring new use cases and solutions to existing problems within their areas of domain expertise.

Note: This page is intentionally blank.

SITUATION OVERVIEW

HPC Cloud Market Overview

Given the shift in HPC buying behavior to the cloud, Hyperion Research expects the cloud market for HPC resources to grow at more than twice the pace of the on-premises server market. Projecting out to 2026, the cloud market for HPC workloads is forecast to reach $11.5 billion USD, growing to nearly half the size of the on-premises HPC server market. While each segment in the HPC ecosystem is adopting cloud resources, a few key sectors and application areas are fueling this high growth:

Hyperion Research expects the cloud market for HPC resources to grow at more than twice the pace of the on-premises server market.

- Artificial Intelligence (AI) and other data-intensive workloads are being run at a higher rate in the cloud due to immediate capacity of diverse types of storage, the wider availability of GPUs (for sites without large GPU installations), the access to public data sets, and the expertise offered by cloud service providers of running AI workloads with HPC.
- The manufacturing and financial services industry (FSI) sectors both are expected to outpace the overall cloud market in growth over the forecast period as manufacturers continuously iterate new designs and reduce time to market while financial institutions use HPC simulations to manage risk.
- The Government sector, specifically government labs like Oak Ridge National Laboratory in the United States or Leibniz Supercomputing Center in Germany, are expected to aggressively increase their cloud usage over the forecast period as they work to understand how and where the cloud can optimize their compute infrastructure.
- Sites that have smaller primary HPC systems, such as those in the workgroup or departmental segments, have shown strong adoption of cloud resources for HPC jobs over the past few years.

This HPC cloud forecast represents what users are spending to run their HPC workloads in the cloud, as opposed to what cloud service providers (CSPs) are spending on HPC infrastructure. This encompasses all HPC resources (e.g., compute, storage, networking, file systems, application licenses, and services).

The table below shows 2020 and 2021 actual HPC user spending in the cloud, alongside the cloud forecast from 2022 to 2026. Notably, the cloud market is predicted to break $10 billion USD by 2026. For 2022, Hyperion Research expects the market to grow 23% YOY, reaching $6.3 billion USD. The five-year CAGR is 17.6%.

TABLE 1

HPC Cloud Forecast 2020-2026

($M)	2020	2021	2022	2023	2024	2025	2026	CAGR 2021-2026
2022 Cloud Forecast	4,300	5,100	6,304	7,369	8,511	9,873	11,453	17.6%

Source: Hyperion Research, 2022

To support the increased interest in HPC cloud resources, CSPs have made dedicated additions in personnel and service offerings to address the specific needs of HPC users. Recent Hyperion Research studies indicate that cloud offerings are starting to erode the growth of on-premises HPC infrastructure investments. Although not pervasive yet, some HPC organizations are moving their HPC work completely to the cloud. A much larger group of users are moving part of their on-premises HPC budgets to support running applications in the cloud, either delaying future on-premises procurements in favor of the cloud or reducing the size of future procurements and using the residual budget for cloud computing.

Prior to 2021, cloud computing was treated primarily as complementary to on-premises computing spending, namely for burst or surge capabilities from users to address spikes in application runs during specific times. Now, cloud computing is becoming a critical computing environment for many HPC practitioners.

THE STATUS OF CLOUD MIGRATION FOR HPC

Cloud computing for HPC has emerged as a strong solution to address some of the consistent issues HPC users are facing worldwide. Many of these issues have become driving forces behind cloud migration from on-premises solutions. CSPs have taken note of these limitations of on-premises HPC computing and addressed those issues with cloud offerings for HPC.

One of the most important drivers to cloud is the need to run workloads immediately. For many HPC sites, queue times to gain access to the necessary scale and technology on a central on-premises HPC system can be long (e.g., weeks, and sometimes even months). As a result, engineers and researchers risk missing their deadlines or being idle. This can have economic impacts, including inefficient use of users' time, lost revenue, or cost savings delays due to increased time to realizing results of the users' work.

Having immediate access to services to run workloads immediately is emerging as primary driver for migration to HPC clouds.

In addition to waiting in long queues, users are also limited by the scale of their central on-premises machines. On-premises HPC clusters are finite and have a maximum scale and a fixed type of technology. The cloud, however, can provide far more resources at a moment's notice, allowing users to scale out as they wish on the most performant technology available for their workloads. For example, as users adopt more AI workloads and require accelerator technology or a different solution than their on-premises HPC cluster, they can look to the cloud for a variety of processor and accelerator options, as well as the necessary elasticity to scale their workloads as needed.

On the personnel side, many HPC sites are experiencing a diminishing number of experts available to employ. The HPC ecosystem is lacking a talent pool large enough to accommodate its expanding nature. Further, many HPC experts are retiring, leaving many HPC sites without the necessary expertise to deploy HPC workloads. The CSPs have taken notice and are actively hiring both HPC experts and domain experts to address the needs of established and emerging HPC engineers. CSP experts are also developing solutions and tools to aid in running optimized HPC and AI workloads.

Cost has always been a hotly debated topic in the context of on-premises versus cloud for HPC. For some workloads, the cloud can be more costly than an on-premises solution, especially for workloads that drive the design of a large HPC cluster. On the other hand, many workloads can be run cost-effectively in the cloud. The core of many cost debates between the cloud and on-premises lies in the budgeting and payment structure of utilizing the cloud resources, and the comparison between capital expenditure and operating expenditure.

The core of many cost debates between the cloud and on-premises lies in the budgeting and payment structure of utilizing the cloud resources, and the comparison between capital expenditure and operating expenditure.

Fixed cost budgeting (sometimes referred to as capital expenditures or CAPEX) is typical of many HPC sites worldwide. These sites make large HPC cluster acquisitions on a four-to-five-year cycle, designing the system well ahead of time and optimizing around the anticipated workload distribution for the projected time period of that system's lifespan. Fixed cost system purchases can lead to more cost-effective computing when considering that some sites do not bear the cost of certain operating cost components, like electricity, power, and physical facilities. In such cases, the cost per job run can be driven extremely low for highly utilized machines. On the flip side, by making a large capital purchase for a system that will last four to six years, the site is committing to the technology available at the time of purchase. The implication being that for the lifespan of the system, the site will either need to upgrade to take advantage of newer technology (e.g., improve efficiency, access the most advanced functionality) or the site will need to buy another system.

Running HPC workloads in the cloud, however, provides a pay-as-you-go business structure. A user pays for compute resources, storage resources, services, and other tools in the cloud in a per-use manner, paying only for what is being used at that moment. For those workloads that have sufficiently long run-times with demanding data capacity and movement requirements, a cloud-based variable cost model may result in a higher cost than an equivalent on-premises system. Conversely, the pay-as-you-go model may be more cost effective for workloads that exhibit shorter run times and less demanding data requirements than when the on-premises solution isn't fully utilized. On top of its elasticity and non-committal nature, a cloud-based pay-as-you-go structure also enables users to always have access to the most up-to-date technology without sizeable up-front expenditures.

An additional element to factor into the evaluation of whether to adopt a fixed cost or pay-as-you-go approach for HPC resources is a realistic assessment of system utilization. When demand for on-premises HPC resources exceeds what is available, users are either burdened with long queue times or could find relief in leveraging cloud-based resources for those periods of time. Conversely, prolonged periods of underutilization of on-premises resources suggests one of two things: either that they were over-provisioned and a smaller-scale system would have been more appropriate, or adopting a cloud-native approach (e.g., little-to-no on-premises HPC resources) could be more cost-effective.

In summary, HPC users are facing a number of challenges today with their on-premises solutions:

- Not having the scale to address growing workload demand by HPC sites
- A shortage of expertise to effectively deploy workloads on optimal technologies
- Long queue times and inadequate scale to complete simulations in a short amount of time
- Lack of access to the latest and greatest hardware solutions in on-premises clusters
- Large capital investments for on-premises systems

AWS and NVIDIA are aiming to address these challenges with their collaboration, leveraging their collective technology and expertise to enhance HPC capabilities for the sector.

HOW AWS AND NVIDIA ARE ENABLING HPC WORKLOADS IN THE CLOUD

With the goal of assisting HPC users to migrate their workloads to the cloud, AWS and NVIDIA have worked together to develop solutions and tools that aid in the migration of HPC workloads to the cloud. First and foremost, AWS and NVIDIA have teamed to develop an infrastructure with the latest advanced technology available to address HPC workloads. Elements of this technology include:

- Amazon Elastic Compute Cloud (Amazon EC2) instances powered by a broad range of NVIDIA GPUs
- High-performance file systems
- High-throughput networking
- Software SDKs

This portfolio of AWS HPC services and NVIDIA software presents a highly capable entry point for HPC engineers to begin their journey. GPU adoption in the HPC ecosystem has been on the rise for several years, especially as both AI workloads grow and as traditional HPC modeling and simulation workloads are modernized to take advantage of GPUs. Benefits targeted by AWS and NVIDIA state-of-the-art services include scale that allows users to run workloads on whatever amount of resource is required, as well as flexibility and choice in available technologies.

Building on their respective areas of expertise, AWS and NVIDIA have collaborated to develop the necessary tools to enable users to run their HPC workloads easier and more effectively. These tools allow users to extract performance from the latest NVIDIA GPUs and pull together the necessary pieces of the workload cycle to run the jobs at the right scale.

By utilizing the services, tools, and expertise that AWS and NVIDIA bring to the table, many users can run their HPC workloads:

- On-demand without a long queue time
- At the optimal scale on the latest technology
- Without requiring the HPC expertise they may otherwise have had to develop themselves

The cloud-based pay-as-you-go model of AWS allows users to pay for what is used, track more closely their spending on compute resources, and optimize budgets around specific workloads or time-constraints. Users can purchase the capacity they need for the duration of their workload but are not committed to a long-term purchase of technology.

A key enabler for cloud adoption in the HPC space is the availability of tools and expertise available from the providers, especially as HPC users adopt new AI applications in their workflow. Both AWS and NVIDIA have collaborated to bring software and toolkits to help improve the skills of HPC users across the span of their workflows. NVIDIA has developed SDKs that help users tackle the challenges of deploying HPC workloads across GPU technologies, optimizing applications to run on the NVIDIA platform.

The collaboration of AWS and NVIDIA ultimately intends to provide a performant, elastic, and flexible platform to run HPC workloads at scale on the most effective technology. In certain cases, these solutions may rival, and at times out-perform, on-premises solutions in cost, performance, and time to solution. Adding in the expertise and services provided by AWS and NVIDIA, the collaboration strives to guide users with best practices in migrating their HPC workloads to the cloud in a seamless and expedient fashion.

FUTURE OUTLOOK

The HPC ecosystem is undergoing a fundamental shift where the cloud has become a crucial part of HPC compute resources, rather than an additional resource to be used only in limited situations. The CSPs have heavily invested to bring solutions to their services that remove some of the larger friction points that are present in on-premises HPC. CSPs now offer HPC users the ability to scale their workloads, both traditional modeling and simulation, as well as AI, as far as needed on their platforms. These workloads can also leverage both the state of the art in terms of new technology as well as the expertise, tools, and services provided by CSPs. Looking forward to the next few years of the HPC ecosystem, nearly every HPC user site will be investigating how to take advantage of what the cloud can offer, especially as the cloud has become more cost effective for many HPC workloads.

Working together, AWS and NVIDIA are aiming to deliver a scalable, performant, and easy to use platform to accelerate and run all HPC workloads. Users can take advantage of the sheer scale of the global AWS cloud infrastructure, running workloads on the latest NVIDIA GPUs, and can do so with ease by using the tools and expertise provided by AWS and NVIDIA.

As users look to investigate future resources, it is important to interact with providers who understand how best to run workloads on the cloud, investigate which technologies and tools are available, and optimize budget around getting more work done in a shorter amount of time. While not every workload will make sense to run on the cloud, AWS and NVIDIA are working to address the needs of HPC sites worldwide.

According to Hyperion Research analyst Alex Norton, *"The future is bright for cloud computing for HPC workloads as providers like AWS and NVIDIA look to bring solutions to address the wide variety of needs that HPC users face."*

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
and www.hpcuserforum.com
