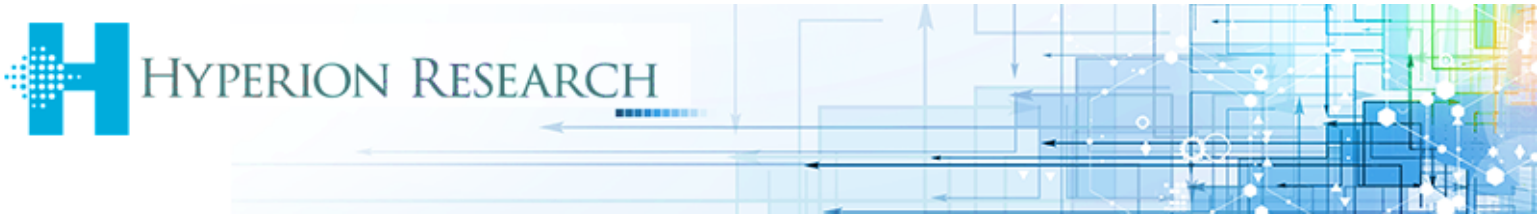


White Paper

Running HPC in the Cloud: The HPCaaS Consumption Model is Here to Stay

Sponsored by: Rescale

Mark Nossokoff and Alex Norton
June 2022

Hyperion Research Opinion

Running HPC workloads is simultaneously becoming more complex and more expensive. The number of users who can benefit from leveraging HPC infrastructure are growing while the skillsets required for deploying, maintaining, and supporting the systems are becoming more scarce. Additionally, virtually all organizations' budgets are being challenged either with shrinking CAPEX (capital expense) and OPEX (operating expense) budgets, or a shifting of CAPEX budget to OPEX. A new business model for accessing and deploying HPC resources would be welcome across the HPC ecosystem.

The public cloud, defined as computing resources provided by a third-party provider to complement or replace on-premises computing, has emerged as a new business model for running HPC workloads. While the cloud does provide elasticity of resources and a pay-as-you-go OPEX cost model, along with access to the latest technological advances, challenges still persist. It can be confusing to identify the appropriate cloud service provider for users' workloads. Once a CSP is selected, getting the application to optimally run in a consistent fashion is no easy task. Determining a predictable and repeatable cost estimate across a variety of workloads can also be troublesome for users.

HPCaaS (HPC as a Service) aims to address these mounting challenges by:

- Making HPC jobs far easier to run and eliminating much of the inherent complexity of HPC.
- Exposing HPC resources for pay-as-you-go OPEX consumption and alleviating underutilization of resources.
- Virtualizing cloud resources to support best-of-breed resources and latest architectures
- Supporting continuous access to all geographies.
- Providing recipes for targeted vertical HPC applications and delivering the appropriate resource profile and predictable cost model matched to the specific HPC workload.

- Assisting scientists, engineers, and researchers in porting their code for cloud resource consumption, thus optimizing workload completion times and time to results.

Rescale is a prime example of an HPCaaS provider delivering what they refer to as HPC Built for the Cloud. Rescale's platform was developed exclusively to support HPC users working to accelerate their engineering innovation and was designed to provide access to optimized cloud-based HPC resources to solve a broad array of scientific and engineering challenges. Rescale provides seamless provisioning of cloud resources, including an extensive library of domain-specific ISV software packages that allows HPC users the flexibility to dynamically meet their technology availability and budgetary requirements.

Note: This page is intentionally blank.

Situation Overview

For purposes of this paper, the following definitions are used:

- HPC Cloud: HPC resources hosted by a cloud service provider provisioned to users on a dynamic, pay-as-you-go basis. Typically, this refers to infrastructure.
- HPCaaS: All HPC resources (e.g., compute instances, storage, applications software) provisioned to users on a dynamic, pay-as-you-go basis, regardless of where the resources are hosted. Resources could be provided by a CSP, collocated at a managed service provider, or housed within a private on-premises datacenter.

HPC Cloud Market Overview

HPC in the cloud, and more recently HPCaaS, has been on the rise for the better part of the last decade, with accelerating growth over the last few years. Consumption of HPC public cloud resources is projected to grow to $9.3B in 2025. See Table 1 for the forecast.

Table 1

HPC Cloud Forecast 2019-2025

($M)	2019	2020	2021	2022	2023	2024	2025	CAGR '20–'25
NEW 2021 Cloud Forecast	3,910	4,300	5,100	6,300	7,150	8,100	9,300	16.70%
HPC Broader Market Forecast*	26,979	27,283	29,383	34,121	37,378	40,015	39,867	7.90%

Source: Hyperion Research, August 2021

To handle the increased interest in HPC cloud resources, cloud services providers (CSPs) have made dedicated additions in personnel and offerings to address the specific needs of HPC users. For the first time, recent Hyperion Research studies indicate that cloud offerings are starting to erode the growth 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 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 datacenters.

HPC sites and their users look to optimize their compute resource pool from several perspectives, including cost, performance, and capabilities. Certain HPC workloads may see competitive cost-per-performance if run in the cloud, especially those that are highly parallelized and do not require strict time to solution requirements. These workloads can be run in a more price-performant environment on the cloud, leveraging inexpensive compute instances. Additionally, AI and HPDA (High Performance Data Analytics) workloads, leveraging data sets that are already stored in the cloud, can minimize data movement costs, which can contribute significantly not only to the cost of running HPC workloads in the cloud, but also to their HPC carbon footprint. Lastly, users are looking to the cloud to evaluate specialized architectures for use with specific workloads without having to commit to acquiring those resources with limited CAPEX funds.

On the other hand, workloads that rely heavily on specific system tuning or unique hardware and/or software packages may make more sense to keep on-premises. Very large HPC jobs, with high communication requirements, are also often more cost effective on on-premises systems.

Data security and associated protocols for the cloud computing area are also a critical consideration. CSPs have made concerted efforts to address the necessary certifications for data storage and movement in, out, and within the cloud. Decisions must be made from the perspective of an optimization problem, with on-premises and cloud computing resources representing two groups that can each handle specific subsets of scientists', engineers', and researchers' HPC workloads.

Anatomy of the HPC Cloud

Resources required to support users' complex, varied, and dynamic HPC workloads span a broad set of areas, each of which also have a wide spectrum of options. These are summarized in Table 2.

Table 2

HPC Resource Requirements and Considerations

HPC Resource	Considerations
Compute	- Type: CPU, GPU - Performance levels, cache sizes, core counts - Memory: amount and type of Interconnect: standard or accelerated
Storage	- Capacity - Performance: flash, HDD and HDDs - Retention: durable/persistent or ephemeral/temporal - Access frequency: hot, cold, archive - Access method: file, block, object
Networking	- Standard ethernet - High-performance network (e.g., InfiniBand, OmniPath, accelerated ethernet)
File System	- Parallel file system (e.g., Lustre, Spectrum Scale, BeeGFS) - Scaleout NAS (e.g., NFS, OneFS)
Applications	- Choice between ISV providers

Table 2

HPC Resource Requirements and Considerations

HPC Resource	Considerations
	- Breadth of domain-specific ISV packages - Licensing model (bring you own or provided by CSP) - Typically file or object data types
Technical support	- Infrastructure: guidance on selecting the appropriate compute instances and storage capabilities for specific applications - Domain-specific: support on porting codes and tools from on-premises HPC infrastructure
CSPs	- Single CSP vs. flexibility based on requirements of the HPC application
Service Level Agreements (SLAs)	- Guarantees or commitments to any type of service level (e.g., performance, workload completion, uptime) - "Every unit item" or price differences associated with varying degree of SLA

Source: Hyperion Research, March 2022

Determining the appropriate cloud resource operating environment to optimally match their complex HPC and AI workloads can be a challenging task for users. At an aggregate (and even individual) level, traditional CSPs have hundreds of compute instances available for users to choose from. Variety, diversity, and choice is a value up to a certain point (which varies by users' skillsets and knowledge base), but it can become confusing and even overwhelming. This can be particularly true for users who are new to HPC operating environments, typically those realizing they need that level of resource to support data-intensive AI-related workloads for their scientific, engineering, or business needs.

- The new HPC users can require additional support and expertise from several perspectives:
- Identifying the appropriate operating environment for their particular workloads
- Simplified expense management via a one-stop-shop for compute, storage, and ISV application licensing
- Guidance on optimizing their home-grown applications for the cloud
- Direction on whether one CSP or compute architecture is better suited to their workloads than another

HPC Cloud Consumption Models

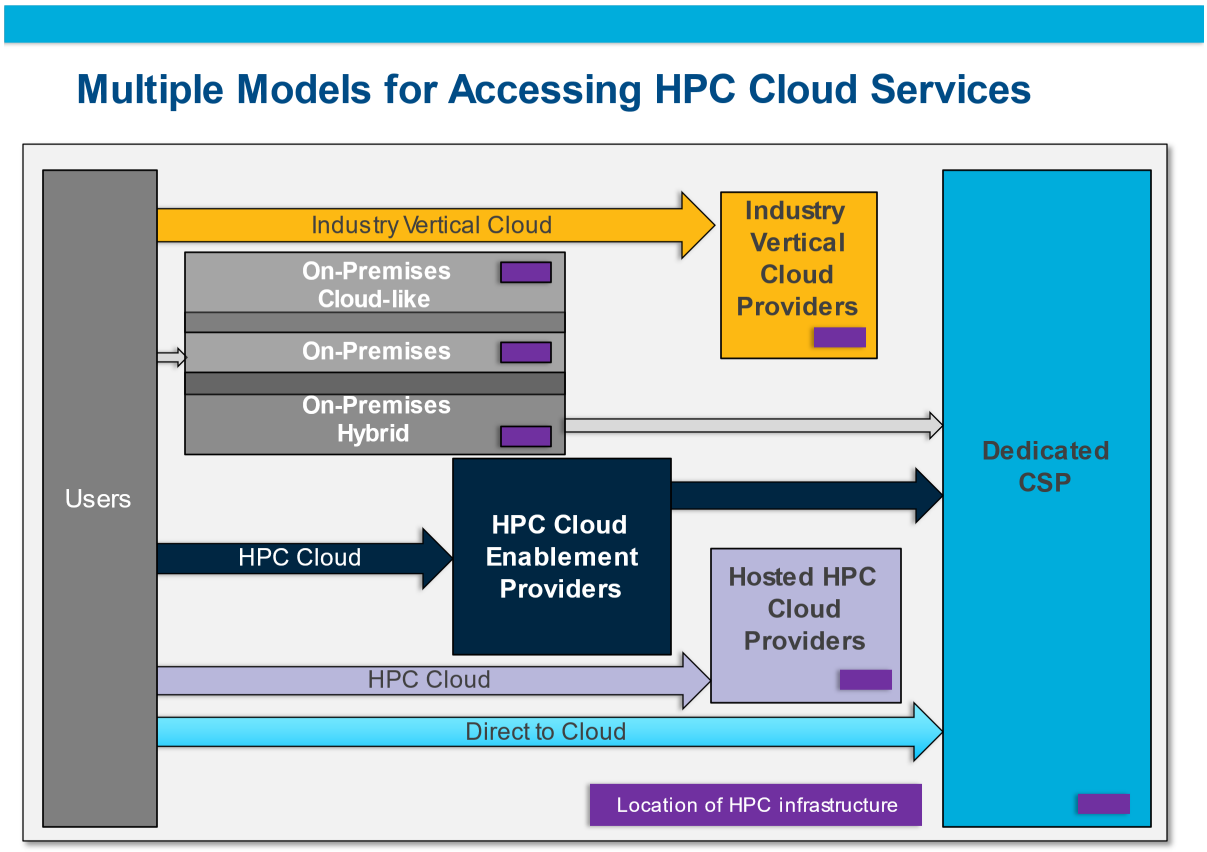
There is tremendous diversity in the breadth of HPC cloud resources available to users. On one end of the spectrum, a user can bring their entire operating environment (operating system, application code, data) to the cloud to run on "bare metal" accessed directly on a primary CSP's cloud. On the

other end of the spectrum, a user can access a complete turnkey HPC cloud solution that includes all required infrastructure, operating environment, application code, and sometimes even the data.

Figure 1 provides a visual depiction of the various consumption models to access HPC cloud resources. Each model provides users (shown on the left) varying levels of physical access, technical support, licensing terms, and budgetary impacts for each method of cloud resource consumption (displayed in their various forms across the graphic).

FIGURE 1

HPC Cloud Consumption Models



Source: Hyperion Research, 2022

On-Premises Cloud and Hybrid Cloud

Until recently, HPC infrastructure located on-premises at an organization's datacenter was purchased outright by the organization from the CAPEX budget. Lease arrangements were more common at larger sites but in either case they were generally treated as fixed costs. This is still true for those sites that primarily run workloads more appropriately run on-premises, with no intention of running HPC jobs in the cloud.

More recently, vendors have developed alternative business models whereby the HPC infrastructure is still located at the organization's datacenter, but the cost is structured in a pay-as-you-go fashion based either on a metric (e.g., capacity for storage products), or service level (e.g., CPU hours used, performance or availability). This cloud-like purchasing of traditional on-premises infrastructure is structured as an organization's OPEX and may be dynamic, based on utilization or SLA. (Service Level Agreement).

Many on-premises datacenters also employ a hybrid cloud operating environment that takes advantage of dynamic use of HPC cloud resources, most often in a "Direct to Cloud" model, in addition to running jobs on their installed servers.

Industry Vertical Clouds

Some users in specialized industries don't require the breadth of service offerings available from the major CSPs, and the major CSPs may not be able to offer the depth of domain-specific technical systems and support these users require. These users often turn to industry vertical cloud service providers. Industry vertical cloud providers provide the domain-specific technical depth and understanding of their users and tailor the infrastructure they host to the users' highly specialized workloads. Some are based on a specific application set. Industry vertical cloud providers are prevalent in the energy and manufacturing sectors.

Many of these providers entered the market in a specific vertical and are now expanding their industry coverage in an effort to expand their market coverage and grow their businesses.

HPC Cloud Enablement Providers

HPC cloud enablement providers provide a bridge between users and cloud resources. They make using clouds much easier and faster for end users, including training for how to best use clouds. Automation is a key element and value-add for HPC cloud enablement providers. Cloud automation can also bring a level of management and control. For example, cloud enablement providers can maximize both engineering productivity and workload performance while minimizing IT risks in security and compliance as well as cost overruns.

HPC cloud enablement providers are highly skilled in the needs of the users, often in a specific set of verticals. They offer a broad range of popular HPC application software and can provide the domain-specific technical support and tools across the diverse HPC user base. Unburdened by large CAPEX demands hosting and managing the HPC infrastructure, these providers can focus their investments on offering their users the widest range of domain-specific expertise while supporting them with best-of-breed HPC resources available across the array of dedicated CSPs. In some cases, and this is true of other consumption models, as well, HPC cloud enablement providers may offer a complete "black box" HPC resource experience either as a base feature or with varying degrees of integrations and SLA to choose from.

Hosted HPC Clouds

Hosted HPC cloud providers provide cloud services focused on running HPC workloads while at the same time owning, managing, and supporting the HPC infrastructure. Hosting the infrastructure allows these providers to optimize it for complex HPC workloads. The types and scale of workloads

these providers can support are limited by the budget available to invest in the expensive HPC infrastructure.

Direct to Cloud

Sometimes referred to as a do-it-yourself (DIY) HPC cloud, this model affords users a wide range of flexibility as to the level of help and support they require. The DIY model allows users who are deeply skilled in both HPC and domain-specific areas to tailor their HPC cloud experience to their specific needs.

While the dedicated CSPs initially entered the HPC cloud market to drive greater utilization of their general purpose cloud infrastructure, they have recognized the business opportunity for supporting HPC workloads. The CSPs are broadening their compute and storage services to address the advanced needs of HPC users, while also adding domain-specific technical support staff to help users transition their existing on-premises applications in the cloud, as well as enable new users to develop cloud-native applications.

HPCaaS and Rescale

HPCaaS is best embodied by the HPC Cloud Enablement Providers. These providers:

- Make using clouds seamless (including multi-cloud and multi-workload use cases)
- Deliver highly skilled support relative to the advanced infrastructure needs of the users
- Offer a broad range of popular HPC application software
- Provide the domain-specific technical support and tools across a diverse HPC user base
- Match the users' specific HPC applications' requirements with the most appropriate hosted cloud service provider

Rescale is a prime example of an HPCaaS provider. Their platform is designed to provide scientists, engineers, and researchers access to optimized cloud-based HPC operating environments. Rescale's seamless, black-box provisioning of the best-in-class HPC cloud resources available in the market is augmented with access to an extensive library of domain-specific ISV software packages, including the requisite domain-specific technical support. Additionally, Rescale can also automate customer’s own cloud infrastructure tenants for those that do not prefer the black-box approach.

Rescale has a diverse customer base representing a broad range HPC applications and workloads. Two example customers who have turned to Rescale for their HPCaaS needs are: Vertical Aerospace and NOV.

Vertical Aerospace

Vertical Aerospace is a pioneer in designing urban electric vertical takeoff and landing (eVTOL) taxis. Adapting cutting-edge digital R&D techniques from automotive and aerospace sectors, the Vertical Aerospace engineers take advantage of simulation and high performance computing (HPC) to improve their vehicles’ aerodynamics, battery performance, and rotor drivetrain efficiency.

Vertical Aerospace established several goals in determining their HPC requirements:

- Support a variety of compute-intensive workloads and R&D simulation applications
- Dramatically accelerate delivery of results
- Manage scale, volume, and workload effectively
- Deliver a cost-efficient solution for our R&D workloads

Adopting Rescale's HPCaaS yielded the following benefits to Vertical Aerospace:

- Transition HPC operations into the cloud without incurring large CAPEX costs
- Support for a wide variety of workloads
- Ability to monitor and adjust the cost-performance of their HPC operating environment
- Improve efficiency of each workload by finding the optimal architectural configurations and maximize utilization of software licenses
- Strengthen system governance with more robust access controls

NOV

NOV, a global leader in oil and gas and renewable energy, uses advanced computer-aided engineering (CAE) simulation to design and test new technologies, pumps, regulators, and drill heads. NOV faced increasing engineering delays from their existing HPC resources, with growing wait times and backlogs of IT support caused by their strong business growth and new R&D initiatives.

Adopting Rescale HPCaaS, including deploying HPC application software from Abaqus, Ansys, and Star-CCM+, NOV was able to achieve the following outcomes:

- Full software license utilization and optimized licensing costs
- 95% reduction in cloud HPC deployment time
- 80%+ decrease in upfront HPC costs and reduced overall operational costs

Future Outlook

Until recently, scientists and engineers who required the technical resources available from powerful high-performance computing (HPC) systems had to be either fortunate enough to work for an organization who could afford one of those systems or compete for access on time-shared resources from a leading academic institution. Even then, the available HPC system may not have supported the latest state-of-the-art technological innovations to provide the timely or extensive results desired

by their scientists and engineers. It is an expensive proposition for any IT organization, even a leadership HPC datacenter, to keep up with the most advanced technologies and solutions.

Infrastructure needs are not the only areas requiring on-going, sustainable investment to stay abreast of the latest trends. Talent is required to architect, design, select, and deploy the systems. Service, maintenance, and support skills are needed to keep the systems running smoothly. Technical experts are required to assist their engineers and scientists to write and run their codes for optimal performance and timely, accurate results.

An HPCaaS business model 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. HPCaaS can be a viable complement and, in some cases, an alternative to traditional on-premises HPC infrastructure deployments for many users and HPC workloads.

Rescale, being an early mover in deploying an HPCaaS consumption model, is well-positioned to continue to innovate within this business model. Key factors to its early success include:

- Developing a team that deeply understands the needs of end users
- Partnerships with the leading dedicated CSPs
- Breadth of tools to monitor resource and application utilization
- A wide array of support for and agreements with the leading HPC ISV application vendors
- Skilled technical support for both HPC infrastructure and domain-specific areas

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.

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 and vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide 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.hpcuserforum.com and
www.HyperionResearch.com
