Special Analysis

HPC Cloud Resources Have Become a Viable Tool for Running Many Large-scale Scientific Research Workloads

Mark Nossokoff, Bob Sorensen, Melissa Riddle, and Earl Joseph
September 2023

HYPERION RESEARCH OPINION

This paper explores the evolution of HPC workloads and how cloud computing has evolved to become a viable tool to complement on-premises capabilities for running large-scale scientific workloads. It looks at the challenges driven by advanced, large-scale scientific workloads, the importance of their resultant research, how cloud-based resources are capable of meeting HPC requirements, and examples of results provided by running large-scale scientific workloads in the cloud.

While there will continue to be a need for on-premises HPC systems for many HPC workloads, cloud-based resources are increasingly being successfully utilized across a wide range of HPC workloads. Some of the factors driving the increased adoption of cloud resources include:

- Access to Larger Scale Computing: For many HPC sites, cloud computing provides the ability to run jobs at a much larger scale than possible on-premises.
- Immediate Access to Resources: Clouds can provide immediate access to resources for unanticipated increases in workload demands.
- Access to New Technologies: Clouds can support fast access to new technologies in both hardware and software, allowing sites to test new technologies before purchasing them.
- Supporting Collaborations: Clouds can provide an easy-to-use platform for broad collaborations, which is simple to use and provides data sharing.
- Sustainability: Clouds can support corporate and regulatory sustainability and energy initiatives.

Hyperion Research projects users will spend over $11B on cloud-based HPC resources in 2026, reflecting a 5-year CAGR of 17.6%. Table 1 summarizes this forecast.

TABLE 1

HPC Worldwide Cloud Forecast: Spending to Run HPC Workloads in Clouds

($M)	2020	2021	2022	2023	2024	2025	2026	CAGR 21-26
Yearly Cloud Spending	$4,300	$5,100	$6,304	$7,369	$8,511	$9,873	$11,453	17.6%

Source: Hyperion Research, 2022

SITUATION ANALYSIS

The Evolving HPC Environment: New Workloads Engender a Spate of New HPC Requirements

The HPC sector, forever exemplified by continual reinvention in technologies, suppliers, markets, workloads, and end user composition, is entering yet another phase in its evolution. The HPC landscape is undergoing significant changes, driven by a combination of the expanded computational demands of traditional HPC workloads amplified by a new host of growing and increasingly critical workloads and related use cases, particularly in AI and big data spaces. These changes are creating major opportunities for both HPC suppliers and end users, affecting computer system design, implementation, and most importantly, use case prospects.

However, in this new environment, HPC end users are facing increasingly complex challenges in correctly characterizing and prioritizing their current and future workload requirements to effectively identify and realize the most suitable HPC solutions to meet those needs. Most recently, the characterization challenges involve properly understanding the options between HPC resources in the cloud and on-premises HPC infrastructure.

In the early days of HPC, key use cases centered on some of the most aggressive modeling and simulation problems that included commercial applications such as automotive crash tests, commercial airframe design, drug discovery, animation, and reservoir extraction procedures. Counterpart national security applications included weapons design, both conventional and nuclear, cyber security research, and weather forecasting for military operation planning.

In this environment, HPC systems suppliers successfully delivered more capability on a near continual basis, fulfilling HPC end user needs for more capable processors, larger memories, faster interconnects, and more performant storage. These improvements, however, came at a certain cost. Typically, HPC vendors focused their development efforts and subsequent product offerings on a relatively limited range of HPC architectures and related system options that were well suited to a narrow class of applications, centered generally on computationally intensive modeling and simulation applications. As a result, HPC users were largely limited to those applications best suited to the HPC architectures of the day. Any significant deviation from those targeted use cases or attempts to introduce new ones ultimately resulted in significant system performance degradation and underutilized resources.

The long standing HPC paradigm described above is rapidly reaching the end of its useful life. In its place, the sector is increasingly being driven by the base of HPC users and their requirements for new HPC use cases, operational requirements, related system architectures, and access models. This new paradigm is marked by:

- More varied and complex hardware options
- More heterogenous design in both system architecture and associated software
- More targeted capabilities for unique computational requirements of any specific HPC/AI workload

Likewise, new opportunities in HPC use cases are broadening to include big data analysis and AI/ML/DL/LLM, while still experiencing significant growth in traditional modeling and simulation jobs, some energized by the inclusion of AI techniques. Equally notable are additional emerging uses cases based on the growing interest in HPC at the edge and real-time HPC requirements. These use cases

will also be influenced by the availability of new accelerator options such as AI-centric GPUs, FPGAs, and nascent quantum capabilities.

Within this new environment, configuration decisions for HPC end users are becoming increasingly more complex, not only for configuring dedicated on-premises facilities but for those seeking to optimize the varied range of HPC/AI in the cloud options including both hybrid and cloud-only configurations. Considerations for HPC designers now include a growing need to balance CPU and GPU configurations, mixed and low precision data formats, memory options such as DDR, HBM, and flash, interconnect schemes to connect nodes of different types and capabilities and flexible storage systems capable of managing the explosive growth in data size, location, and formats.

This changing ecosystem will be marked by several major notional shifts within the HPC sector at large, many due to the various opportunities and challenges surrounding the increased presence and variety of mature cloud-based capability for HPC workloads. Indeed, since their inception CSPs have been managing complex interactions across heterogeneous compute environments. Likewise, CSPs have been investing heavily to offer features that are easier to use, more HPC-orientated, more frequently updated, more easily scalable, more responsive to unanticipated shifts in workload composition, and more varied in hardware and software options than those found in a traditional on-premises counterpart.

Additional considerations for HPC users will include tradeoffs concerning sustainability, energy, and carbon footprint impacts between on-premises and cloud alternatives, in-house versus third party support for domain or vertical-specific application development or operational expertise, and long-term considerations of optimizing budgets.

CSPs may become the computing choice for many new HPC entrants that are attracted to the opportunities enabled by the spate of new HPC uses cases, particularly those embracing AI-based opportunities looking for low barrier-to-entry options with a more flexible pay-as-you-go financial commitment.

Cloud Capabilities for Large-scale Scientific Workloads

The expansion of CSPs' HPC capabilities across areas important to scientists and researchers has made them more capable of running a broader range of scientific workloads. The conversation about employing cloud-based resources has turned from an adversarial “cloud vs. on-premises” debate to a dialogue about how to leverage both approaches to provide scientists and researchers with an optimal balance of performance, scale, cost, features, and skillsets to drive continued scientific advancement and discovery.

Multiple recent Hyperion Research studies have revealed several consistent motivations identified by users for employing 3rd party cloud HPC resources to run a broad range of HPC workloads, including those for large-scale scientific research:

- Access to Larger Scale Computing: Clouds may provide the ability to run jobs at a much larger scale than possible on-premises, supporting potentially major increases in performance, shorter turnaround job completion times, and the ability to do science at a much larger scale.
- Immediate Access to Resources: Clouds can provide immediate access to temporary resources for unanticipated increases in workload demands.
- Access to New Technologies: Clouds can support fast access to new technologies in both hardware and software and allows sites to test new technologies before purchasing them.

- **Supporting Collaborations:** Clouds can provide a platform for broad collaborations, which is easy to use and provides secure data sharing.
- **Sustainability:** Clouds can help support corporate and regulatory sustainability and energy efficiency initiatives.

Access to Larger Scale Computing

Many HPC researchers and sites have HPC and AI problems that are dramatically larger than what can be resolved with their existing resources. Some desire to use sizable data sets, others would like results completed faster, and many need to run more complex or higher resolution problems. CSPs offer the ability to scale their problems and workloads to a much larger size and complexity than what may be available users on-premises.

Running HPC/AI workloads at a larger scale can provide performance results that redefine how their research is conducted. For example, taking a problem normally requiring over a day to run and having the capability to run it in under an hour can allow the researcher to try many times more options, leading to a significant improvement in the rate of their research and more scientific breakthroughs. In other cases, having the greater scale allows solving problems that couldn't even be considered without having the scale of computing.

Using a cloud for running larger problems can be a major advantage for sites that only need to run very large jobs part of the time.

In many centers, there are an extensive number of jobs being run at a given point in time and this can cause many jobs to take a long time waiting in line to be completed. For time-critical jobs, the use of clouds can speed up the wait time and provide results more quickly.

- Existing on-premises infrastructures are often being pushed to their limits, balancing how many jobs can be handled and how quickly they can be completed. Job completion times are determined by an array of factors including size and complexity of the workloads, amount of data required for each workload and queue depth of the number of jobs competing to utilize the infrastructure, among others.

- The scale of available resources directly impacts how quickly a job will be completed and how large the job can be. This is true of both traditional scientific research modeling and simulation workloads and newer training and inferencing of large-scale AI models.

CSPs have greatly improved and built out their high-performance capability instances and services. In certain cases, CSPs have achieved this by working closely with, and learning from, sites with leadership HPC capabilities. While many large-scale scientific workloads fit dramatically better on-premises, many today can perform well when run in the cloud, often with the help of domain area experts employed by the CSPs to optimize researchers' codes. The performance provided by HPC cloud resources is largely a result of the very large scale at which CSPs can operate. CSPs have the ability to offer a broad variety of resource options at a scale not likely available to mainstream on-premises data centers.

Immediate Access to Resources

Often a site has an HPC or AI computing job that requires immediate access to different types of hardware, software or scale. There is a need to run these jobs quickly, but there is not time nor budgets to conduct an acquisition of new resources. Clouds can help in these situations with an ability to quickly provide access to the needed computing resources.

In addition, depending on the type of research being performed or the cyclical nature of the scientific research process, there are times when on-premises resources are fully utilized and important jobs are being held in the queue, waiting for a large-enough scale of resources to become available to run the job. If this happens with enough frequency, the time to science and discovery can be greatly delayed. Having immediate access to large-scale HPC capabilities in the cloud for known times of cyclical oversubscription of on-premises resources or times of unanticipated increases in workload demands can greatly accelerate time to results of the research.

Access to New Technologies

The variety of computing hardware and software has grown dramatically over the last few years, and few sites have a broad mix of all of the different technologies available today. Many sites haven't yet decided which new technologies should be added to their existing systems, and often have to wait 3 to 5 years before making each technology upgrade. Clouds can provide fast access to new technologies and offer a pay-as-you-go pricing model, allowing sites to use and explore new technologies before having to go through the acquisition process.

- Due to CSP's scale, they often have access to the most recently available hardware and software innovations much sooner than many on-premises sites.
- New and emerging technologies such as quantum computing, application-specific acceleration capabilities, and networking innovations are often first introduced and available in the cloud.
- Even if on-premises data centers were provided access at the same time as CSPs, they may not know what they want to procure, have the budget for it, or have the expertise required to immediately leverage it.

Emerging use cases such as generative AI are driving the need for new, advanced, and application-specific hardware and architectures. These new architectures are yet to be fully proven and can be cost-prohibitive. Evaluating the new technology in the cloud affords scientists and researchers to prove its capabilities, define areas where codes may need to be optimized, and assess vendor strengths well ahead of needing to expend large capital expenditures for on-premises investments. In some cases, the evaluation process may identify best practices for preserving and continuing on-premises investments while simultaneously exploiting HPC in the cloud to accelerate time to discovery and expand the amount of discovery scientists and researchers are capable of.

Additionally, many HPC sites are keeping their systems longer, often for 4 to 5 years. This makes it difficult to have access to new hardware that may appear after their on-premises system is designed and procured. Using clouds for certain jobs can provide access to new hardware before the next on-premises system upgrade cycle. Also, due to the scale of CSPs operations, CSPs may be given preference by vendors on allocation of the most recent technology in supply-constrained environments, causing delays in availability of the technology on-premises for smaller-scale customers.

Supporting Collaborations

In many cases, collaboration between scientists and researchers is vital in achieving the important outcomes being researched. Providing the capability of more researchers to have access to the most up-to-date datasets in their entirety can promote:

- Acceleration of scientific discovery.
- Production of an increased amount of research.
- Advancements in the quality of insights and research being produced.

Increasing collaboration requires fast and easy access to shared data and information, which CSPs can provide. These open datasets typically consist of contributions from public sector and government collaborators, which can then be subsequently shared with and accessible by both public and commercial researchers.

One requirement to enable broad global collaborations is user confidence in strong, reliable security and the protection of their IP and data. CSPs tend to frequently apply security updates and maintain high levels of security throughout their operations. By maintaining and evolving security capabilities, CSPs are able to support global access, data sharing, and collaboration for users with sensitive research and information.

Sustainability

Sustainability, energy efficiency, and reducing HPC's carbon footprint impact on the environment is a global concern and top-of-mind for many organizations as they strategically plan their future HPC resource requirements. CSPs have been leaders in the effort to drive large-scale data centers towards carbon neutrality. Migrating large-scale workloads to CSPs that have demonstrated continued results to carbon-reduction goals can both satisfy organizational mandates for achieving their internal goals, as well as preclude potentially costly investments to on-premises facilities to achieve those goals.

Exemplars of Successful Large-scale Scientific Research Workloads Utilizing HPC Cloud Resources

There are numerous examples of successful utilization of HPC cloud resources for large-scale scientific and research workloads across a broad range of domain areas (e.g. computational fluid dynamics [CFD], protein folding, classical weather modeling). Four are highlighted below.

Moderna

Project Summary

Moderna completed the sequence for their coronavirus vaccine (in partnership with NIH) in just two days after receiving the virus's genetic sequence by leveraging AI, ML, and parallelizing previously sequential drug development processes. This greatly accelerated the process to vaccine production, testing, approval, and distribution, ultimately saving lives.

Key Cloud Benefits: Running at a larger scale, access to new technologies, and cost savings

The infrastructure flexibility of cloud HPC to incorporate GPUs or high memory strictly for the relevant stages of drug development was critical to both speed and cost. Moderna also frequently used spot

instances and auto scaling to access large clusters for short periods of time to produce results in hours rather than days. Overall, Moderna states that using the cloud saved them 90% of their compute cost compared to on-premises procurements of similar scale.

Moderna also credits their CSP for enabling integration of the AI algorithms used to optimize each possible molecule, automate logistics decisions, and perform quality control data analysis with improved speed and accuracy. Cloud HPC also benefited this project by helping to integrate structures across teams and embedding analytics.

Schrödinger

Project Summary

Schrödinger modeled the interactions between a potential drug candidate and its protein targets. This simulation is repeated for thousands of potential drugs to determine which candidates are most promising to synthesize and test in the lab. Schrödinger previously used on-premises data centers but found their workload sizes were not consistent as they required 100,000+ processors for only a few days each month to perform each batch of drug selection.

Key Cloud Benefits: Running at a larger scale, immediate access to resources, access to new technologies, and cost savings

HPC cloud usage was particularly well-suited for this workload because of the unevenness in the processing power needed throughout the month. Running these workloads in the cloud was more cost-effective than procuring a large on-premises system that would be fully utilized sporadically and sit idle the remaining time. In addition, using the cloud gave access to a near-infinite volume of available processing power that allowed Schrödinger to scale up beyond what could have been accomplished on-premises. With their CSP, they were able to periodically provision 100,000 GPUs as needed. Being able to fit each iteration into a single GPU card also improved the stability and effectiveness of the simulations compared to what they had experienced on-premises.

ExxonMobil/XTO Energy

Project Summary

XTO Energy collected, pre-processed, and analyzed data on their oil wells such as temperatures, pressures, and flow rates. Data was collected from each well through IoT and then datamined through cloud HPC offerings to produce business insights into the operations and maintenance of existing wells as well as prospects for future drilling.

Key Cloud Benefits: Collaboration & data sharing, access to external cloud experts, and access to new technologies

XTO decided to use HPC cloud for this workload to take advantage of the benefits of IoT offerings more fully. Since the data from each oil well is collected via IoT and natively stored in the cloud, their integrated approach with HPC cloud computing made their solution more efficient and reduced downtime compared to on-premises computing. Since data is easier to move within cloud environments, this also improved access to data by their geographically distributed engineers.

XTO also cited ease of use and technical support from their CSP as a major benefit. They had a small team of five people working on this project, so support from experts at their CSP allowed them to grow their computational abilities manageably and sustainably. Similarly, XTO plans to take advantage of

cloud tools and services to expand their current computing environment and to rely on their CSP's expertise to incorporate AI and machine learning.

Royal Hollaway, University of London

Project Summary

Royal Hollaway, University of London, researchers analyzed 3D microtomography images of rocks to determine their pore volume and capacity for carbon storage. These factors determine a potential reservoir's suitability to absorb carbon dioxide from the atmosphere and potentially reduce the effects of climate change.

Key Cloud Benefits: Running at a larger scale providing faster turnaround times, immediate access to resources, access to new technologies, and cost savings

Moving this workload to the cloud significantly reduced time to solution by enabling researchers to select the amount of memory and threads needed in a manner not possible on their on-premises system. On their local system, each iteration could be completed in just under three minutes (2:49), but due to the large number of iterations, this resulted in nearly two days runtime (47 hours) per simulation. Using the cloud reduced by more than half the iteration time to just over a minute (1:12 per iteration), and the simulation time to less than a day (20 hours per simulation). Also noted is the speed-up was accomplished at a lower-than-expected cost.

FUTURE OUTLOOK

Achieving the greatest success for migrating and running large-scale scientific and research applications in the cloud is largely predicated on "working with" as opposed to an "instead of" on-premises infrastructure. As each HPC site is unique, so are their deployment models for HPC resources. Ultimately, choosing where an HPC job can and should be run is based on a myriad of factors, each of which needs to be carefully evaluated, especially for large-scale applications.

Cloud providers have invested heavily to improve their solutions to address large and complex HPC and AI challenges. Additionally, they have added HPC technical expertise to assist in designing better systems and to improve end user support. In the areas of AI and LLMs, they have made strong investments in developing custom hardware including complex, workload-specific processors to help improve their capabilities. They have also recognized that HPC-class system designs and architectures are required to run the growing number of new AI large language models.

Furthermore, CSPs have been targeting traditional modelling and simulation workloads in their system designs. They have been expanding and improving their cloud based HPC resources to address the challenging issues encountered by many scientists, researchers, and engineers. These challenges include:

- Ability to run jobs at a much larger scale than possible on-premises.
- Immediate access to temporary resources for unanticipated increases in workload demands.
- Having a mechanism to leverage new technologies for emerging use cases.
- Facilitation for broad collaboration on important scientific challenges.
- Supporting corporate and regulatory sustainability and energy efficiency initiatives.

As the requirements of these demanding users have increased to solve ever-more complex scientific and engineering challenges, many users have successfully turned to the cloud to augment their current on-premises infrastructure.

These investments by the CSPs have enabled their HPC resources to evolve to a more mature state to support a broad range of HPC workloads, including large-scale scientific applications. In many cases, the cloud has emerged to be a viable tool to complement existing and future on-premises infrastructure to optimize and accelerate scientists' and researchers' time to discovery and expand the universe of outcomes that they are capable of addressing.

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