HYPERION RESEARCH

Special Analysis

2022 HPC End Users Perspectives on Use of Public/External Clouds for HPC Workloads, Trends, and Drivers

Mark Nossokoff and Earl Joseph
February 2023

HYPERION RESEARCH OPINION

While barriers to cloud adoption persist in some areas, many have been overcome, and users are moving beyond surge and experimentation use cases to an increasing number of production workloads. The cloud market for HPC continues to grow at a strong pace as cloud service providers (CSPs) address the needs and concerns of HPC users as well as the on-going education of users on how to optimize HPC workloads for the cloud. Insights into the critical factors driving these and other trends are detailed in the 2022 iteration of Hyperion Research's annual MCS end users' study, 2022 HPC *Multi-Client Study: Use of Public/External Clouds for HPC Workloads, Trends, and Drivers*. Key findings from the report are summarized in this document.

Hyperion Research conducts an annual Multi-Client Study (MCS) to measure and track key trends across the spectrum of the HPC market. The latest iteration of the MCS encompassed 181 HPC end-user sites with 3,830 HPC systems. Reports produced as a result of the study span:

- AI and HPDA Usage and Future Technology Trends
- Vertical/Application Workload Areas and Technical Computing System Software and Middleware
- Use of Public/External Clouds for HPC Workloads, Trends, and Drivers
- Processors, Coprocessors/Accelerators, and HPC Budgets
- Trends and Forecasts in HPC Storage and Interconnects

Running HPC Workloads in the Cloud

Cloud computing continues to garner high interest in the HPC space for a variety of reasons:

- Users are publicly sharing their successes running HPC workloads in the cloud.
- CSPs are improving their HPC resource capabilities and assistance.
- The cost-effectiveness of using the cloud for HPC workloads has improved.
- The question of “why not go to the cloud?” has become widespread in the IT space overall.
- What used to be a complementary solution for auxiliary or non-essential workloads has become a crucial compute resource for production HPC workloads at many sites.

SELECT KEY FINDINGS

Key findings of this report include data on HPC public cloud usage, spending, and HPC user plans for future utilization of HPC resources in the cloud.

Cloud Adoption is Now Impacting On-Premises HPC Spending

Over 40% of respondents who utilize HPC cloud resources indicated their on-premises budget would be impacted (e.g., reduced or delayed), albeit in most cases by a relatively small amount at this time.

Nearly half of the respondents who employ HPC cloud resources indicated their on-premises HPC budget will be impacted by the cloud budget. Of those who shared what that impact would be, half indicated they would take some of their on-premises HPC budget and directly shift it to HPC cloud resources.

Of sites that indicated employing HPC cloud resources would delay procuring a new on-premises system, the average delay specified was between one and three years.

Compute and Storage Consume Over 80% of Respondents' Cloud Budgets

Almost 60% of cloud budgets go towards compute instances. Storage is next largest consumer of cloud budgets with approximately 23% going towards persistent storage and 10% going towards ephemeral storage. The balance goes towards software licenses and other expenses.

AWS is the Most Common Primary CSP for HPC Users

60% of sites surveyed spend at least half of their cloud budget on AWS and 75% of sites surveyed indicated using AWS for some portion of their cloud-based workloads.

More Sites are Employing Multiple CSPs

Among the cloud users surveyed, each site employed an average of two CSPs with some utilizing up to four. Reasons for employing a multi-cloud strategy include:

- Gaining access to specialized hardware only available on one CSP platform, like the TPU from Google or the Trainium processor from AWS.
- Accessing specific tools or services.
- Leveraging a dataset hosted or already in a specific cloud platform.
- Collaborating with other researchers who utilize different platforms.

The Majority of Cloud Instances are Reserved

HPC cloud users report that most of their cloud compute instances today (64%) are reserved. This is part of the overall trend among cloud users transitioning from occasional burst workloads to more of a long-term infrastructure view.

Many HPC cloud instances (43%) are accelerated, echoing the on-premises rise of GPUs in HPC. Additionally, respondents report nearly a quarter (23%) of their HPC public cloud spending is now dedicated to persistent storage.

FUTURE OUTLOOK

Running HPC workloads in the cloud has become a more commonplace occurrence in the HPC ecosystem compared to the past few years. Ultimately, users will continue to reevaluate their evolving workflows to investigate how best to optimize and distribute their compute budget across platforms and providers. Some workloads may fit well in on-premises systems while others may be cost-effective and more performant when run in a cloud environment. Looking forward, users will have to both understand the needs of their workflow as well as communicate with their providers to identify what can be done with a varied set of compute resources across cloud and on-premises.

HPC sites will have to evaluate their future compute resources similar to an optimization problem with many variables, including cost, time to solution, data, hardware requirements, skillsets, and others. Cloud computing for HPC will continue to be a significant component in future HPC deployments and technology roadmaps, competing directly with other solutions to address critical HPC applications.

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
