Special Report

Cloud-based AI Activity for HPC: Widespread but Primarily Exploratory

Tom Sorensen and Bob Sorensen
September 2024

EXECUTIVE SUMMARY

The purpose of this study was to gain a better understanding of the activities and use behaviors of AI users leveraging HPC-centric cloud resources. Key goals included creating a picture of user goals for AI integration, their current and planned methodologies, budget allocations, model lifecycle expectations, and preferred hardware and cloud platforms for their HPC-centric AI endeavors. This study also sought to gain a greater understanding of inferencing activities among organizations currently or planning to integrate AI into their workflow.

Highlights from the study results include:

- Public cloud resources are considered a valuable asset in exploration and integration of AI into HPC or compute-intensive environments.
- Respondent organizations are leveraging a wide range of public cloud offerings.
- There are numerous architectures and device types currently used to meet inferencing needs.
- There is a plethora of desired qualities for the future of cloud computing expected by current and prospective AI users.
- Budgets among respondent organizations are expected to increase to meet training and inference needs, both in the cloud and on-premises.

The survey, which was conducted in July 2024, collected input from 105 survey respondents who indicated current or planned use within the next 12-18 months of AI on public cloud-based resources to support HPC or compute-intensive activities.

- Respondents came from a mix of major sectors: commercial (76%), academic (2%), and government (14%), representing verticals led by computers and related electronics but also including the financial sector, bioscience, advanced manufacturing, and geosciences.

Key Findings

Key Finding #1: Cloud AI users most frequently indicated that they are still exploring AI-centric options both in the cloud and on-premises.

When asked about their AI activities in the cloud, respondents most often indicated that they are currently engaged in exploratory and experimental stages of AI adoption, even those who have already integrated these technologies into some portion of their production environments. Across the board, these respondents demonstrated preference towards the cloud over on-premises options for AI-centric activities.

Key Finding #2: Respondents indicated a wide range of motivating factors for AI integration, most of which are targeted at improving existing processes.

When asked about driving factors and goals for exploring the adoption of AI technology, there was a wide range of desired outcomes. Notably, while respondents are hopeful for new advancements to be made in their field, most indicated expectations that the adoption of AI technologies increases the effectiveness and time to solution for their current HPC or related compute intensive activities.

Respondents most often indicated that they are currently engaged in exploratory and experimental stages of AI adoption.

Key Finding #3: Surveyed organizations are exploring a diverse range of machine learning, deep learning, and Large Language Models (LLMs), with multiple model types being explored or leveraged.

Reflecting the exploratory and generally optimistic expectations of HPC users adopting AI into existing workloads, respondents often reported experimentation and development of multiple AI models and algorithm types. Continuing the trend of further expectations to introduce new models for efficiency and optimization, all algorithm options were selected at relatively high rates for both ML and DL uses.

Key Finding #4: Respondents often indicated using multiple sources for public cloud services, primarily with AWS, Google, and Microsoft Azure.

Respondents generally indicated having a primary cloud provider among the 'big three' (Microsoft Azure, AWS, and Google) for their cloud compute provisioning, but most respondents were using multiple cloud providers to meet their needs. Usually, these secondary cloud services were a small fraction of the compute and budget ratios.

Key Finding #5: AI adoption among respondent organizations may be showing signs of approaching saturation, with only 11% indicating planning to initiate new activities within the next 12-18 months.

The majority of respondents indicated that they have already adopted AI methods into their HPC workflows, many between 12-18 months ago. While the trajectory of AI adoption is still in its early phases, especially when it comes to efficient and optimized integration, organizations with a passing or no relationship with AI are dwindling. This is not a reflection of market saturation in terms of full integration but rather an indication of lowering numbers of new entrants to the market.

Key Finding #6: Generative/LLM training and fine tuning are conducted at almost all respondent sites.

While previous studies have indicated AI in HPC users leveraging commonly available market LLMs, respondents still indicated training their own generative models and LLMs as well as engaging in fine

tuning in-house. Furthermore, many respondents are taking responsibility for the entire lifecycle of an AI model and its integration, including full training, fine tuning, testing, verification, and inference.

Key Finding #7: Certain architectural needs are yet to settle on an industry standard, especially inferencing needs.

When asked about preferred and expected device types, respondents reported that inferencing needs are being met by a range of hardware options including general on-premises devices, inference-specific hardware, laptops/desktops, and edge devices. The spread of these reported preferences indicates that organizations' inference needs are varied, and that industry standard solutions for production inferencing have not yet settled.

Summary and Next Steps

The public cloud AI user behaviors and expectations in this survey are reflective of broader market sentiments regarding use patterns, current place in integration processes, and willingness to contribute monetary and labor efforts to their endeavors. This widespread adoption is enabled by a robust and frequently updated set of offerings from public cloud providers, as well as continued excitement and mounting use cases for AI integration into like environments.

While cloud offerings fill the same general role of on-premises resources in terms of meeting compute needs, many users leverage the lower long-term commitment, high architecture diversity, and update frequency of CSPs to explore new methodologies and configurations. While budget commitment to AI-capable cloud resources is high, it can often be used as a means of testing compute configurations as integration is optimized.

Note: All numbers in this document may not be exact due to rounding.

Note: All monetary values shown in US dollars unless specified otherwise.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

IN THIS STUDY

The intent of this study was to gain a better understanding of cloud resources in the development, integration, and production use of AI algorithms. For the purpose of this study, AI activity in the cloud was limited to those workloads that directly supported existing HPC or compute-intensive workloads within the respondent organization. Respondents represented organizations that used a hybrid of on-premises and cloud resources. This effort took a broad-based approach to AI that included the major categories of machine leaning, deep learning and large language models.

Research Approach

This study is based on an independent survey of organizations currently involved in cloud-based AI integration or production use or planning integration or production use within the next 12-18 months.

Key study goals included:

- Describing the current and anticipated benefits and goals for AI integration.
- Assessing the current AI types and models within groups currently or planning to leverage AI.
- Understanding the computing landscape used in AI integration and usage within organizations that commonly use HPC.
- Characterizing the interest and expectations of AI in the cloud among HPC users currently or planning to leverage AI methodologies within existing workflows.
- Exploring the details of cloud offerings, CSPs, and specific architecture types most commonly used in enabling HPC workloads.
- Identifying how inference needs are met among these users.

Note that a more detailed description of respondent demographics can be found in the Appendix.

SURVEY RESULTS

This survey sought to gain a greater understanding of the relationship between HPC-capable cloud resources and organizational AI integration in its varied stages, including:

- Understanding the scope, expectations, and goals of organizations leveraging public cloud resources for AI in HPC.
- Classifying the model types, methods, and algorithms currently and prospectively in use among these organizations.
- Profiling the purchasing patterns and provisioning behaviors of AI for HPC users.
- Illustrating the architectural landscape of compute needs and how they are met using a diverse set of devices, cloud resources, and on-premises server appliances.
- Predicting and quantifying the current and expected budget and compute needs to meet projected goals.

Characterizing Goals, Expectations, and Activities for AI in the Cloud

The survey results demonstrate a diverse range of current and planned AI in the cloud activity to support HPC endeavors: many users reported continued engagement in exploration of the technology, even while integrating or fully leveraging AI for production processes. Table 1 summarizes the wide

range of efforts currently under way for both cloud-based and on-premises AI efforts. Taken as a whole, respondents are exploring a wide range of HPC-centric AI activity that spans AI technology development, exploration, hardware and software options, and even workload production activities.

Table 1

Current Level of Cloud-based HPC AI Activity

Activity	% Selected
Exploring the range of potential cloud-based AI performance enhancements	64%
Exploring the range of potential on-premises AI performance enhancements	50%
Running production level AI-enabled workloads in the cloud	43%
Testing/assessing cloud-based AI-integrated workload performance	39%
Reaching out to cloud service providers for hardware and software information	35%
Procuring cloud access to necessary AI software	34%
Standing up limited cloud-based AI-integrated pilot programs	31%
Reaching out to AI hardware and software suppliers for information	30%
Procuring cloud access to necessary AI hardware	30%
Running production level AI-enabled workloads on-premises	30%
Passively monitoring AI technology developments	27%
Porting AI capability into existing cloud-based HPC workloads	27%
Testing/assessing on-premises AI-integrated workload performance	26%
Procuring AI-based hardware for on-premises use	25%
Standing up limited on-premises AI-integrated pilot programs	22%
Procuring AI-based software for on-premises use	20%
Porting AI capability into existing on-premises code for use in the cloud	20%
No current activity	1%
Don’t know/ Not sure	1%
Other (please specify)	1%

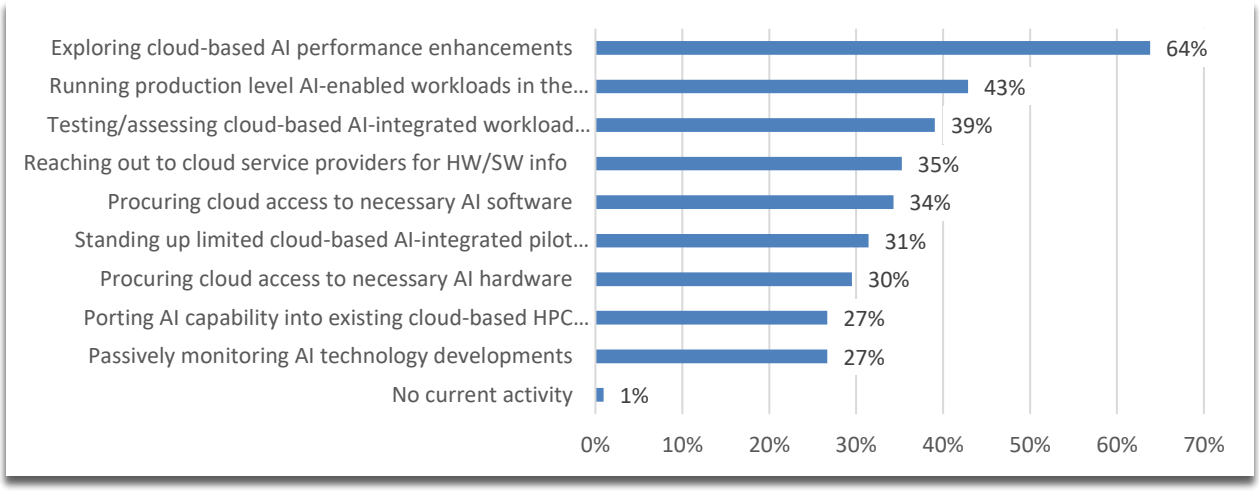
N = 105, Respondents could select all options that apply. Those who selected “Other” specified classified work

Source: Hyperion Research, 2024

Figure 1, seen below, focuses on these responses for cloud usage alone. Organizations report being in various stages of integration lifecycle, commonly running production-level workloads while simultaneously testing for improvements, exploring new options, and provisioning new or more application appropriate resources.

FIGURE 1

Current HPC-related AI Activities in the Cloud



N = 103

Respondents could select all that apply

Source: Hyperion Research, 2024

Table 2, below, categorizes the anticipated benefits of integrated AI capabilities supported by cloud computing. Attracting, developing, and/or maintaining a more effective workforce was the most often selected at 58%, with automated coding or application modernization narrowly behind at 55%. General automation was selected at only a rate of 25%, next to workforce scheduling enhancement at 26%

- Improving data analytics and developing new insights were all at or below 1% in reported uses.
- Directly influencing customer experience was not reported as a popular use goal compared with internal activities such as enhancing current activities, speeding existing processes, or enhancing workflow.

Table 2

Main Benefits of Cloud-Based HPC AI Capabilities

	Currently
Attract, develop, and/or maintain a more effective workforce	58%
Automate coding or applications modernization	55%
Boost existing compute intensive application performance	34%
Create new content with generative AI	31%

Table 2

Main Benefits of Cloud-Based HPC AI Capabilities

	Currently
Drive business process optimization	30%
Drive customer-related recommendation strategies	27%
Enhance computer-vision capabilities	26%
Enhance workforce scheduling optimization	26%
Improve robotics/automation capabilities	25%
Improve customer service capabilities	22%
Personalize customer experiences	17%
Use expert systems to develop new insights in R&D	1%

N = 105

Respondents could select multiple options.

Source: Hyperion Research, 2024

Exploring AI Business and Operation Drivers

Table 3 arrays the most important business drivers for AI efforts among surveyed organizations. Competitive advantage (61%), driving innovation (57%), increasing revenue (55%) and improving research capabilities (55%) stood as the most common drivers while realizing cost savings (47%) and reducing time to market (42%) were slightly less prioritized. Outlying on the bottom, enhancing employee retention (19%) was the least selected.

Table 3

Business Drivers Most Important for AI Efforts

	Currently
Achieving competitive advantage	61%
Driving innovation	57%
Increasing revenue	55%

	Currently
Improving research capabilities	55%
Enhancing business process efficiencies	49%
Realizing cost savings	47%
Reducing time-to-market	42%
Enhancing employee retention	19%
Don't know/Not Sure	1%

N = 105

Respondents could select multiple options.

Source: Hyperion Research, 2024

Table 4 characterizes those operational goals considered most important by integrating AI capability into an organization's base of existing cloud-based workloads. While a faster time to solution for key HPC-based scientific or engineering solutions sits on top at 56%, over half of respondents indicated that their goals include providing new insights or knowledge not possible on traditional systems (55%).

- This reflects not only the growth of trust in the technology but an expansion of operational consideration among these organizations beyond enhancing and speeding existing research paradigms.
- Similar to faster time to solution for key HPC-based science and engineering projects, a greater efficiency on existing HPC-based workloads was a commonly stated goal at 48%.

Table 4

Most Important Operational Goals Envisioned by Integrating AI Capability into Existing Cloud-based Workloads

	% Selected
Faster time to solution for key HPC-based scientific or engineering solutions	56%
Providing new insights or knowledge not possible on traditional systems by themselves	55%
Greater efficiency on existing HPC-based scientific and engineering workloads	48%
Increasing revenue	42%
Greater computational fidelity on key HPC-based scientific and engineering workloads	39%

Table 4

Most Important Operational Goals Envisioned by Integrating AI Capability into Existing Cloud-based Workloads

	% Selected
Opening new lines of HPC-based scientific and engineering research	38%
Lowering overall HPC hardware costs	31%
Opening new lines of vertical-specific end uses	31%
Lowering overall HPC software costs	30%
Opening new lines of multiple vertical end uses	29%
Lowering overall HPC power requirements	18%
Reducing HPC staffing requirements	13%
Other (please specify)	1%

N = 105

Respondents could select multiple options

Source: Hyperion Research, 2024

According to Table 5, 24% of respondent organizations are already spending over $1M in the cloud on AI related activities per year, and 22% spend between $1 million and $10 million. There appears to be a wide range of current budgetary commitment to support cloud-based AI activities, with 24% spending less than $50,000. However, for many organizations, AI-based cloud activities are inextricably linked to other compute intensive activities and cloud infrastructure, and it may not be clear which budgetary commitments are exclusively for AI-related efforts.

- Indeed, 10% of respondents indicated that they were not tracking AI cloud spending separately and another 10% were unsure.

Table 5

Current Annual Spending for AI-related Work in the Cloud

	% Selected
Under $25,000	9%
$25,000 to less than $50,000	13%

Table 5

Current Annual Spending for AI-related Work in the Cloud

	% Selected
$50,000 to less than $100,000	9%
$100,000 to less than $250,000	10%
$250,000 to less than $500,000	7%
$500,000 to less than $1 million	10%
$1 million to less than $1.5 million	8%
$1.5 million to less than $2.5 million	6%
$2.5 million to less than $5 million	5%
$5 million to less than $10 million	3%
More than $10 million	2%
We do not track AI cloud spending separately	10%
Don't know/ Not sure	10%

N = 105

Source: Hyperion Research, 2024

AI Development Areas: ML, DL, and LLMs

Table 6, below, arrays a simple breakdown of AI types currently being used in public clouds by respondent organizations. Most respondents indicated exploring multiple regions of AI methodologies, with a combination of ML and DL being most common at 85%, and LLMs second at 57%. This is reflective of the experimental and exploratory behaviors of users even after production phases have been initiated.

Table 6

Planned and Current AI Types Using Public Clouds

	% Selected
Large Language Models (LLMs)	57%
Machine Learning	53%

Table 6

Planned and Current AI Types Using Public Clouds

	% Selected
Deep Learning	32%

N = 105

Respondents could select multiple answers

Source: Hyperion Research, 2024

Table 7 arrays the machine learning algorithm types being used by the 53% of respondents who indicated they are leveraging ML for their compute intensive workloads. Classification algorithms that many used to identify, sort, or clean data, were shown the most interest at a rate of 67%. Neural networks, which can carry out prescriptive types of generation, was not far behind at 64%.

- Public cloud resources allow for speed and ease of piloting potentially scalable AI models and methods. Users are more likely to explore numerous model and algorithm types with entree to remotely accessible and available cloud resources as opposed to on-premises devices.

Table 7

Machine Learning Types Planned or Currently Used on Cloud AI Platforms

	% Selected
Classification algorithms	67%
Neural networks	64%
Regression algorithms	57%
Reinforcement learning	46%
K-means clustering	42%
Self-supervised learning	41%
Random forest algorithms	39%
Probabilistic clustering	36%
K-nearest neighbor	30%
Hierarchical clustering	29%

Table 7

Machine Learning Types Planned or Currently Used on Cloud AI Platforms

	% Selected
Naive Bayes classifying	22%
Not Applicable	1%

N = 102

Respondents could select multiple answers

Source: Hyperion Research, 2024

Table 8 arrays the deep learning algorithm types used by respondent organizations. Convolutional neural networks (67%), recurrent neural networks (64%), and generative adversarial networks (57%) were the most commonly selected types. As with machine learning types, respondents indicated a general willingness to employ various kinds of model configurations, with even the least represented algorithm type being reported by over 20% of respondents.

Table 8

Deep Learning Types Planned or Currently Used on Cloud AI Platforms

	% Selected
Convolutional neural network	67%
Recurrent neural network	64%
Generative adversarial network	57%
Long short term memory network	46%
Multilayer perceptron	42%
Autoencoder	41%
Self-organizing map	39%
Deep Q-learning	36%
Deep belief network	30%

Table 8

Deep Learning Types Planned or Currently Used on Cloud AI Platforms

	% Selected
Radial basis function network	29%
Restricted Boltzmann machine	22%

N = 105

Respondents could select multiple answers

Source: Hyperion Research, 2024

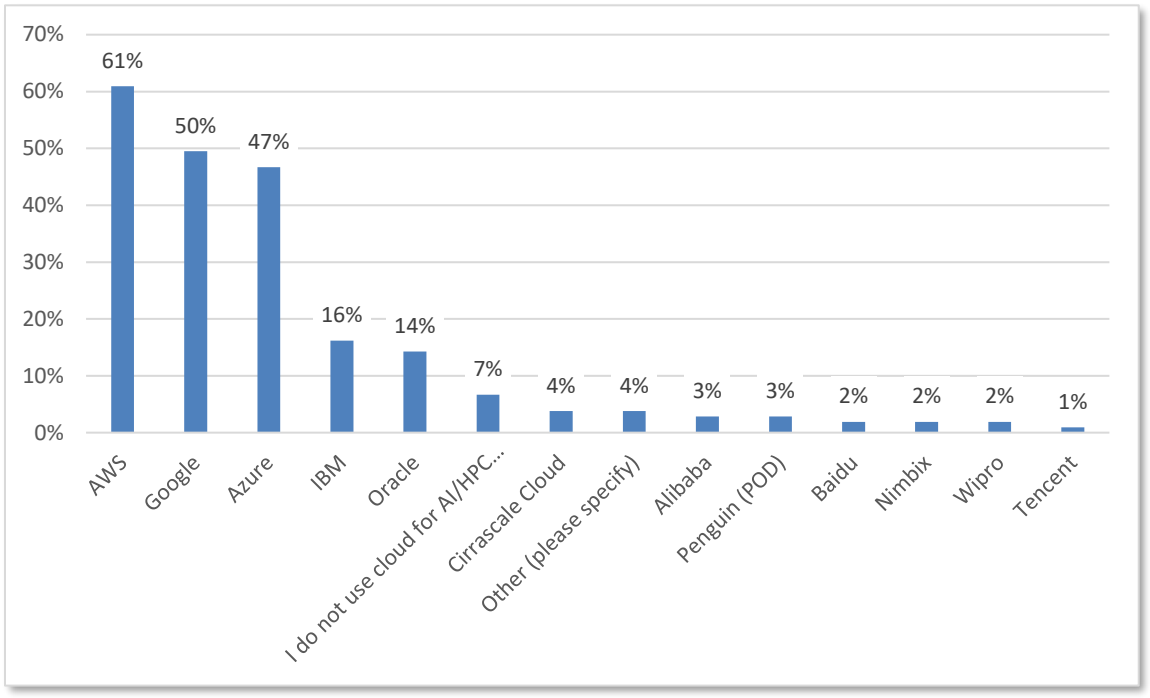
Cloud Provider Considerations

Figure 2 lists major public cloud providers and their use rates and preferences as indicated by the respondent organizations. The three main players were AWS (61%), Google (50%), and Azure (47%). IBM (16%) and Oracle (14%) were second-tier selections, with ten other CSPs in the mix, albeit all less than 4%.

- Respondents often indicated leveraging multiple cloud providers simultaneously, most likely for reasons such as application specific benefits, cost-savings efforts, or pre-existing infrastructure.
- Others mentioned included OpenAI, Lambda, and Digital Ocean.

FIGURE 2

Current Cloud Providers for Overall AI/HPC Compute Intensive Workloads



N = 105

Respondents could select all that apply

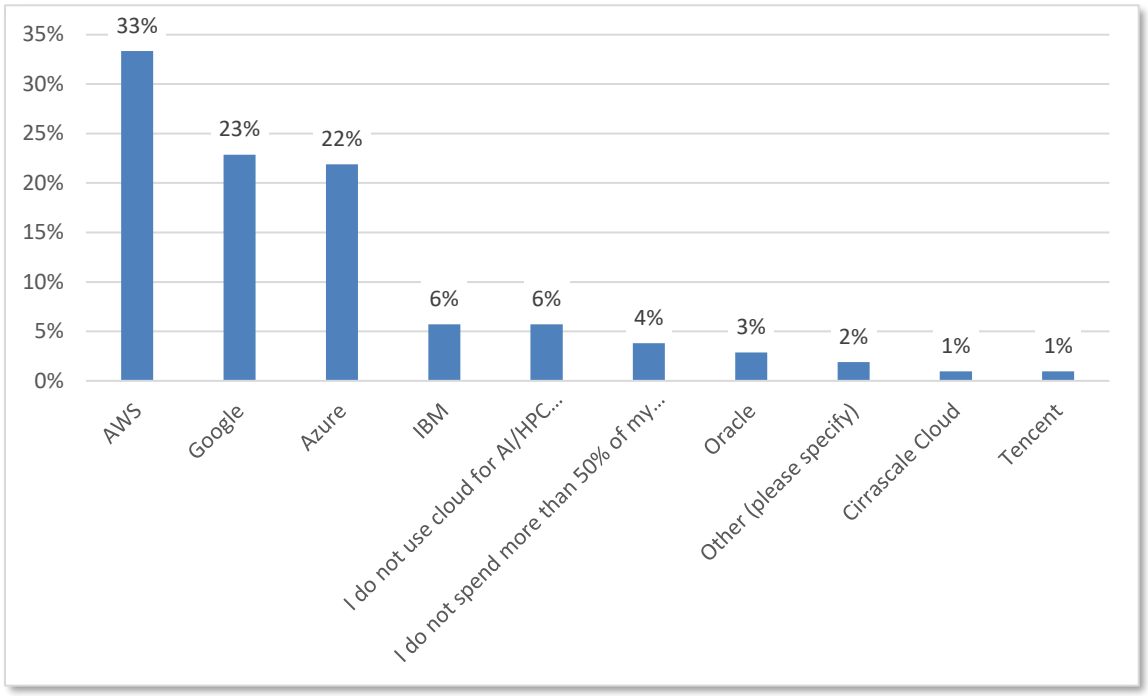
Source: Hyperion Research, 2024

As Figure 3 demonstrates, when asked about their primary CSP for the HPC-centric AI workloads, most respondents selected AWS (33%), Google (23%), or Azure (22%). These three main CSPs, all US-based, account for 78% of cloud-based HPC-centric AI work.

- The lack of commitment to Chinese CSPs, Alibaba, Baidu, and Tencent, likely is due more to the limited participation of Chinese and related Asia/Pacific respondents in this survey than any general lack of interest for their particular services.
- Only a small percentage (4%) do not commit the majority of their workload to a single CSP.

FIGURE 3

Primary Cloud Provider for AI/HPC Workloads



N = 105

Source: Hyperion Research, 2024

Table 9 quantifies the recency of usage for AI in the cloud to support compute intensive workloads. As expected, most users are already using public cloud AI options. Nearly 80% of users indicated they have already begun efforts, and 60% indicated having started more than 6 months ago. Only 11% plan to wait six months or more, perhaps an indication of maturation of the technology, and that 'new players' will be less common as integration processes continue their cycle.

Table 9

Recency of Public Cloud AI Utilization for HPC Support

	% Selected
Have not started yet	9%
In the last 3 months	7%
In the last 4-6 months	11%
In the last 7-12 months	21%

Table 9

Recency of Public Cloud AI Utilization for HPC Support

	% Selected
In the last 13-18 months	39%
In the next six months	4%
In the next 7 to 12 months	5%
In the next 13-18 months	2%
Don't know/Not sure	3%

N = 105

Source: Hyperion Research, 2024

AI Use Rates in the Cloud

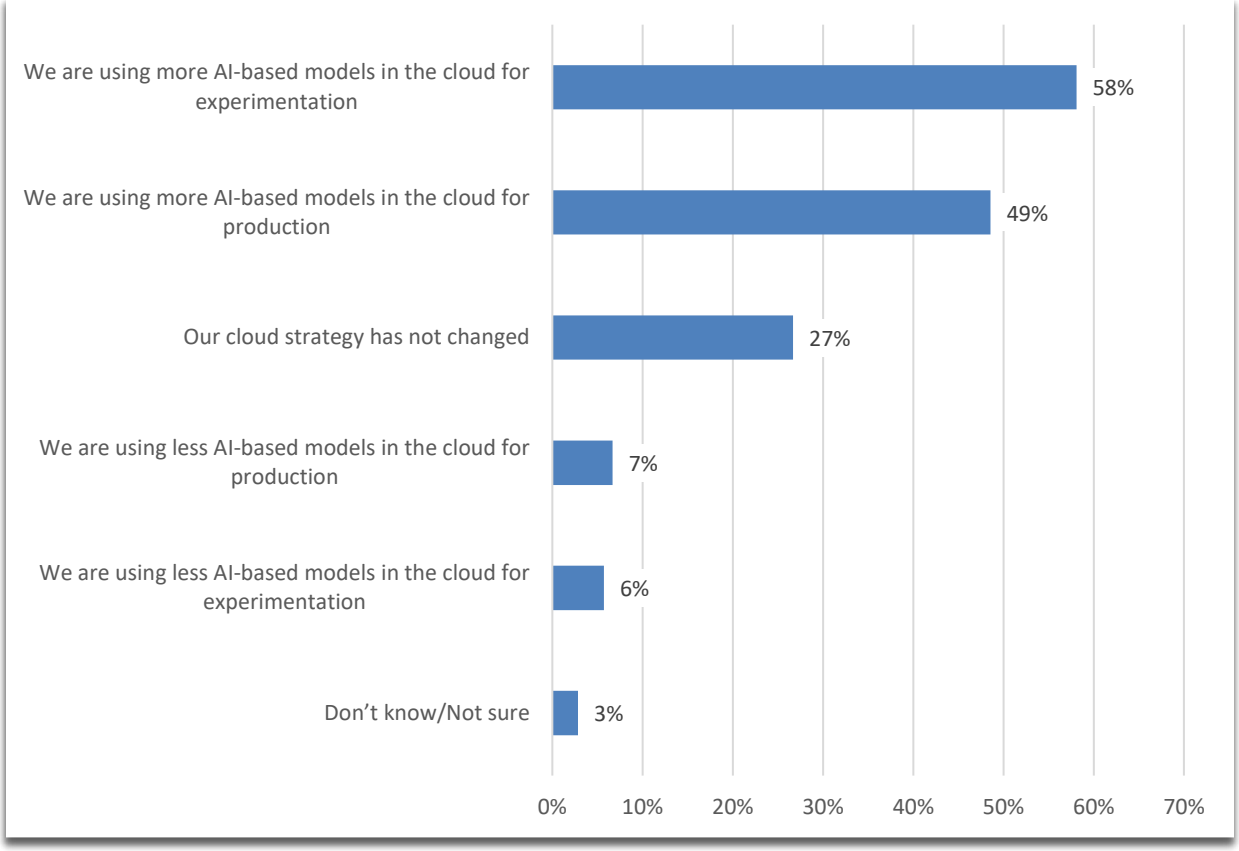
Respondents indicated high levels of importance for cloud resources in the development and integration of AI-based solutions. Cloud environments are seen as beneficial especially for experimentation. That said, these resources are used for all stages of workloads, from development of models to training, testing, and inference. Despite cloud offerings being codified in manners suggesting their intended use, such as CPU or memory-centric, users indicated using a wide variety of instance types or resource configurations to support their needs.

Figure 4 arrays the decision influence based on new or emerging AI based models. Only 27% of respondents indicated that new developments in AI-based models have left their cloud strategy unchanged, while 58% indicated that they are now using more cloud resources for AI experimentation and 49% using more cloud for production.

- Only 13% of respondents indicated that they are using the cloud less for AI activities.

FIGURE 4

How AI-based Models Have Influenced Cloud Strategy



N = 105

Respondents could select all options that apply.

Source: Hyperion Research, 2024

Table 10 demonstrates the importance of AI-based models in relation to current cloud-based HPC activities. Nearly 90% of respondents indicated that these AI-based models were somewhat or very important to their current cloud-based activities, with 49% reporting it as very important.

- These AI-based models in the cloud have become a pivotal part of organizations leveraging HPC for production, science, and engineering within the overall roadmap for testing, integrating, and scaling new technologies.

Table 10

Importance of AI-based Models to Current Cloud-based HPC Activities

	% Selected
Very important	49%
Somewhat important	39%
Neither important nor unimportant	7%
Somewhat unimportant	2%
Very unimportant	2%
Don't know/Not sure	2%

N = 105

Source: Hyperion Research, 2024

Table 11 further reinforces that organizations currently leveraging HPC or compute intensive methods for their research, production, and science often have considerable investment in cloud services. 44% of the respondents surveyed indicated their overall computational workload being 50% or more in the cloud.

- Only 2% aren't using the cloud today.
- Very few respondents (1%) indicated 100% cloud use, likely a result of the targeted user base.
- The largest portion of respondents who had less than half of their computational workload in the cloud was the 20%-29% range at 18% reporting.

Table 11

Percentage of Overall Computational Workload On-Prem vs Cloud

	% Selected
None in the cloud	2%
1%-9% in the cloud	15%
10%-19% in the cloud	6%
20%-29% in the cloud	18%
30%-39% in the cloud	6%

Table 11

Percentage of Overall Computational Workload On-Prem vs Cloud

	% Selected
40%-49% in the cloud	6%
50%-59% in the cloud	9%
60%-69% in the cloud	5%
70%-79% in the cloud	13%
80%-89% in the cloud	12%
90%-99% in the cloud	4%
100% in the cloud	1%
Don't know/Not sure	4%

N = 105

Source: Hyperion Research, 2024

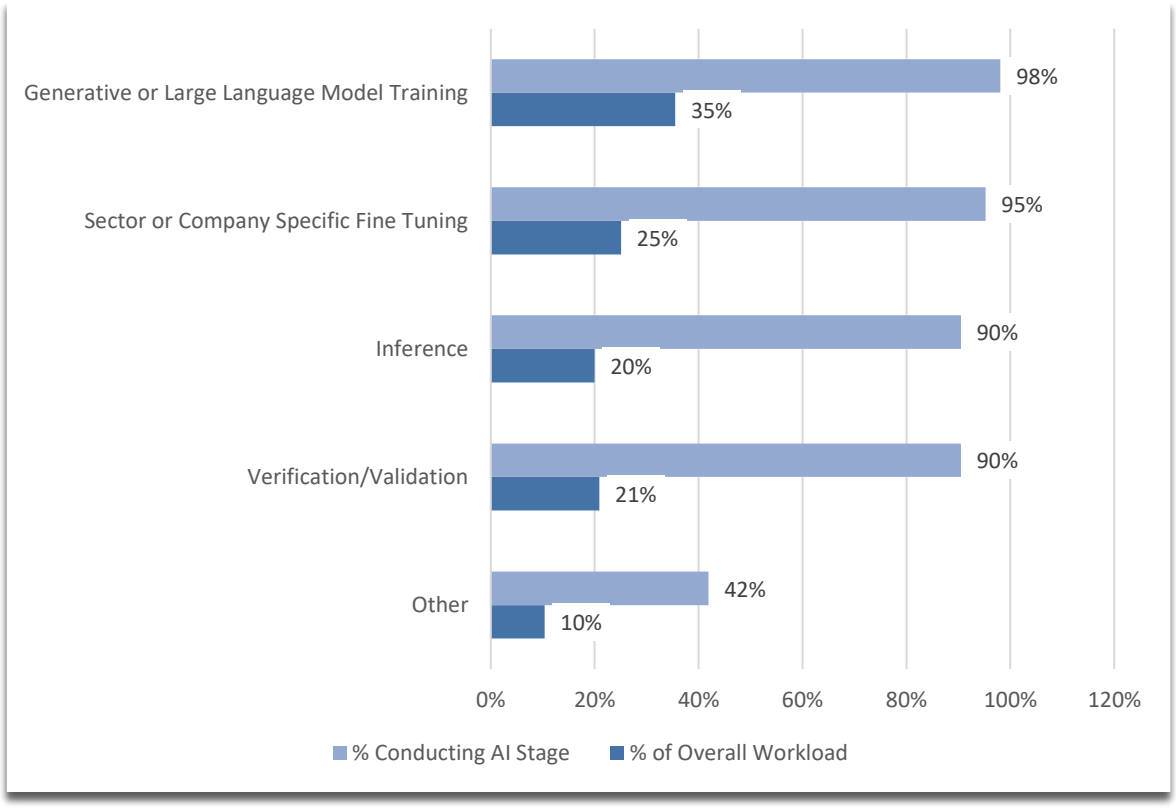
Figure 5 reveals the number of respondents who carried out any given stage of AI workload (light blue) and what percentage of their workload commitment goes into that stage (darker blue). For example, generative AI or LLM training is conducted at almost all respondent sites (98%), and those respondents indicated that 35% of their overall AI workload was committed to LLM training.

- Likewise, fine tuning was conducted at 95% of respondent organizations, consuming one quarter of total AI runtime for those respondents.
- The vast majority of respondents indicated carrying out all elements of AI workload stages (Gen AI or LLM training, sector or company specific fine tuning, inference, and validation/verification).
- Among these respondents, Gen AI/LLM training consumes the plurality of their workload commitment in terms of AI-based compute intensive activity.

User responses indicate an interest in ground level activity in the whole cycle of AI integration. These AI workload stages, in some cases, have discreet compute requirements and resource needs, creating incentives for a diverse architectural landscape. This diverse set of needs can more easily be met, at least quickly, by the aid of cloud service providers and their variety of hardware and software configurations.

FIGURE 5

Percentage of Cloud-based AI Workloads and their Stages



N = 105

Source: Hyperion Research, 2024

Focus On Inferencing

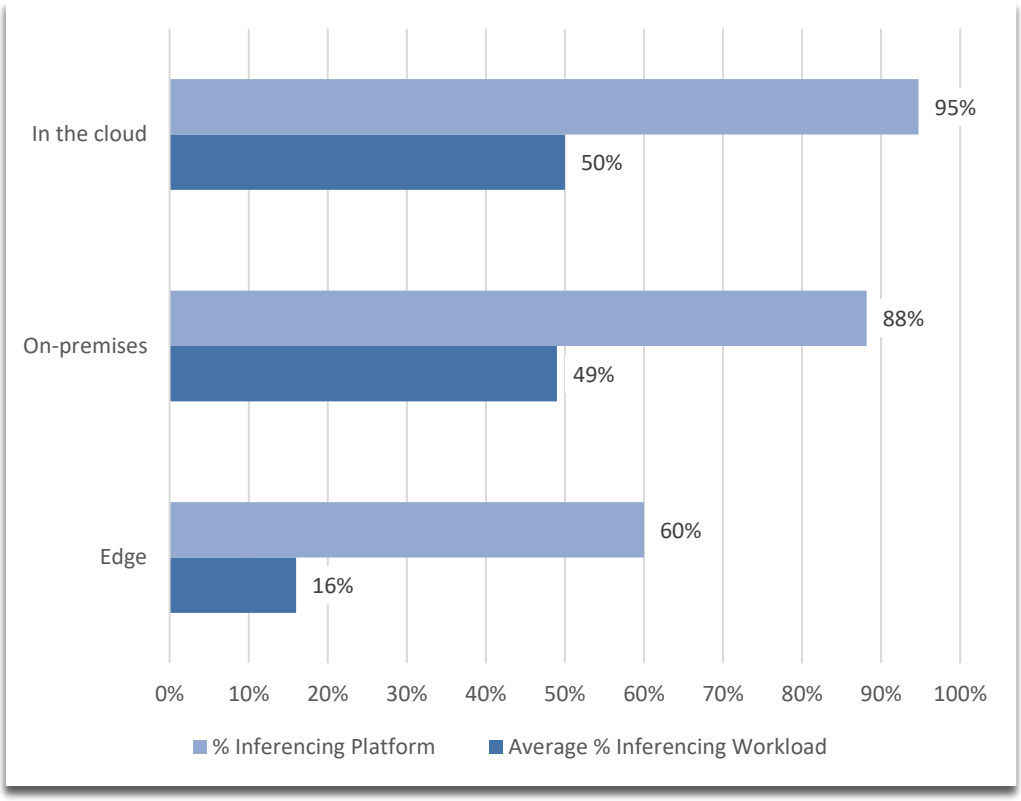
Figure 6 indicates that the majority of respondents are running inference workloads, and those who are utilize a generally even split of cloud computing environments and on-premises resources, with 16% of the general inference workload also being carried out on edge devices. Specifically:

- 95% of respondents conduct inferencing in the cloud to address an average of 50% of their inferencing workload.
- 88% conduct inferencing on-premises to address an average of 49% of their workload.
- 60% conduct inferencing at the edge to address an average of 16% of their workloads.

Inference workloads can vary greatly in their time to completion, required compute power, and queries per day based on their output type, stage in the model's overall development or integration, and overall application efficiency. Mixed resources offer users the ability for application specificity, smarter scaling, and experimentation to meet these needs. Edge devices, existing on-premises resources, and the plethora of cloud resource options will likely help to meet these diverse needs as they grow and change.

FIGURE 6

Where Inference Workloads are Conducted



N = 76, Cloud, On-Prem, Edge = 67,72,46 respectively. Workloads based on Runtime

Source: Hyperion Research, 2024

Figures 7 demonstrates current preferred platforms for carrying out inference needs. Again, there is a relatively even distribution suggesting a diverse set of requirements based on idiosyncratic elements of application type, model development stage, and user needs.

It also suggests that there are not well-settled methods for carrying out inference workloads across the board. As the need of inference scaling comes to light both within organizations and across industries, it is expected that device types and other supporting technologies for these workloads will become more narrowly focused on what is most optimized.

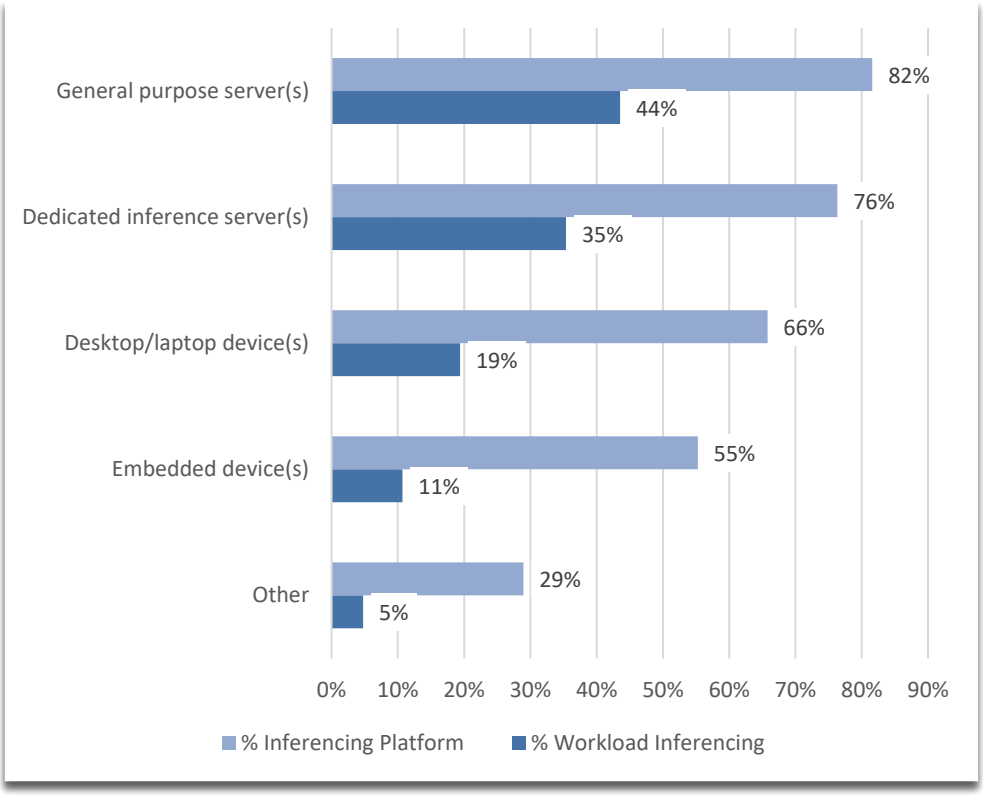
Currently, in addition to exploring and experimenting with options, organizations are also making use of existing hardware on-premises. It is expected that more hardware technologies will be developed in the future that are specifically designed for meeting inferencing needs, and that application profiles will become more well understood surrounding this workload type.

- These preferred device types, broken out, do not show large changes over the next 12-18 months, but there is a trend toward dedicated inference servers, spurred by increased

budgetary consideration for AI endeavors, developments in workload-specific hardware, and a general scaling up of inference runs.

FIGURE 7

On-Premises Inferencing Divided by Device Type by Runtime



N = 76

Source: Hyperion Research, 2024

Similar to the device types, Figure 8 identifies current cloud inferencing platforms by their categorized reported optimization configuration by cloud service providers. Not all CSPs identify their environments or instances in the same manner, but there are a set of common generic instances widely used. Specifically:

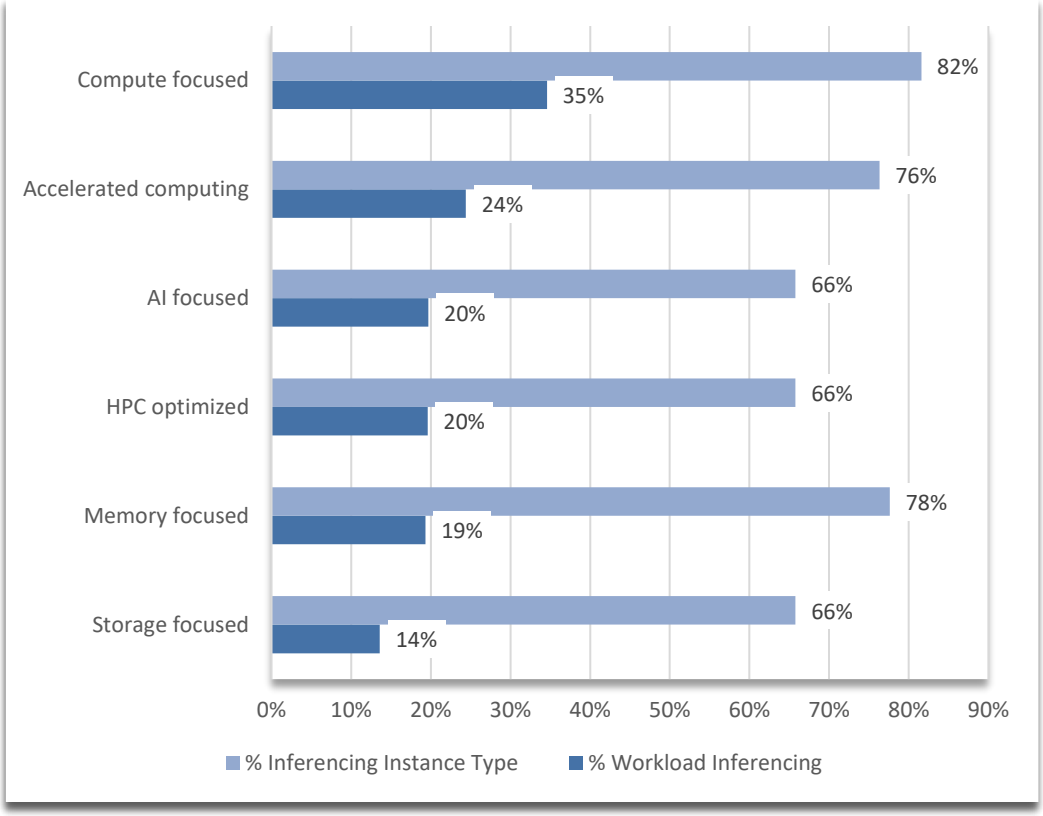
- 82% do compute focused cloud-based inferencing for an average of 35% of their cloud-based inferencing workload.
- 76% do accelerated computing focused cloud-based inferencing for an average of 24% of their cloud-based inferencing workload.

Similar to device types in on-premises, there is a wide range of uses for all cloud configuration types, even those well out of the range of typical AI needs in terms of their stated optimization.

- Notably, it is compute focused and accelerated computing focused cloud configurations which are most currently running and expected to run the bulk of inference workloads.
- Memory focused cloud offerings are the second most leveraged option but are next to the lowest in terms of overall workload commitment.
- While not shifting the position above compute and accelerated categories, respondents indicated expecting that HPC optimized and AI focused cloud offerings will become more commonly used in the next 12-18 months. However, this change is occurring slowly.

FIGURE 8

Current Cloud Inferencing Divided by Device Type by Runtime



N = 76

Source: Hyperion Research, 2024

On-premises HPC-centric AI Spending

When asked whether inferencing would require the purchase of more computational power on-premises, 48% of respondents answered positively. Table 12 quantifies the timespan in which this expectation is confined:

- 20% have reported it is already needed.

- 82% are expecting to need it within the next 12 months.
- 96% are expecting to need it within the next 2 years.

Table 12

Time Period in Which More On-premises Computational Power Will Be Needed for Inferencing

	% Selected
It is already needed	20%
Within the next 0-6 months	24%
Within the next 6-12 months	38%
Within the next 1-2 years	14%
Within the next 2-5 years	4%
5 years or more	0%

N = 50

Source: Hyperion Research, 2024

In order to support overall AI-based processes, Table 13 reveals that 54% of those respondents are expecting an increase in their annual budget of over 7% in the next 12-18 months, with 12% expecting an increase greater than 20% to meet these new needs.

Table 13

Expected Additional Annual Budget to Support On-Premises AI-based Processes in 12-18 months

	% Selected
Less than 1% of existing IT budget	4%
1% to 3% of existing IT budget	12%
More than 3% to 5% of the existing IT budget	24%
More than 5% to 7% of existing IT budget	6%
More than 7% to 10% of existing IT budget	34%
More than 10% to 15% of existing IT budget	8%

Table 13

Expected Additional Annual Budget to Support On-Premises AI-based Processes in 12-18 months

	% Selected
More than 15% to 20% of existing IT budget	0%
More than 20% of existing IT budget	12%

N = 50

Source: Hyperion Research, 2024

SUMMARY AND NEXT STEPS

HPC users and organizations carrying out compute-intensive workloads are currently and increasingly leveraging public cloud resources for AI endeavors. While these resources are often used in conjunction with small devices, on-premises servers, and larger compute clusters, the benefits offered by cloud services are unique and varied in a manner that offers specific benefits to users, especially when exploring, testing, integrating, and running AI-based applications to support compute-intensive workloads.

The study data suggests that this increased usage of cloud resources for AI workloads is pragmatic:

- Overall, users feel more comfortable experimenting with and exploring new AI methodologies and solutions in a low-commitment cloud environment.
- Users feel free to access computing provisions from multiple sources and of various architectural configurations to seek optimization.
- Users are still in different stages of determining application-specific compute needs, and the wide availability of different compute configurations in clouds allow for optimization to be discovered before an on-premises buy-in.

This willingness to explore options is also reflected in the range of ML/DL/LLM algorithm types currently being explored by respondent organizations. Furthermore, budgetary commitments to meet new compute needs, especially related to inferencing, are reported to increase dramatically, demonstrating hopefulness and trust in new and integrated technologies. This confidence is also reflected in a willingness to carry out new AI workloads on existing on-premises hardware.

AI production integration across HPC and compute-intensive focused organizations has been a continuing effort for many months, and data suggests that, while many users are still engaged in exploring options and experimenting with optimization, those with the intention of adopting this new technology are already on their path. Although it is too early to suggest a saturation of the market, there is a dwindling number of advanced technical groups that are not in some way connected or cognizant of AI technology and its potentially beneficial relationship with traditional HPC workloads.

With the speed of excitement, adoption, and development within the advanced AI market, the cloud has become an agile, reactive environment to which users can turn for the latest offerings and support. While hardware development is already in a stage of increased pace that has not been seen before, CSPs have the advantage of circumventing long buying cycles, more streamlined installation into existing workloads, and a more composable way of designing compute infrastructure. This agility within the cloud allows for CSPs to offer users more varied and up-to-date solutions while on-premises resources must follow a different, often lengthier path to utilization.

- On-premises resources are still of great benefit to users, and an on-premises compute environment would still ultimately be more cost-effective than cloud solutions, provided the workloads are well understood and profiled.
- Cloud resources, while agile and numerous, are not well suited for every organization leveraging HPC for workloads, especially in cases involving specific data management laws, privacy, and sovereignty.

There is a dwindling number of advanced technical groups that are not in some way connected or cognizant of AI technology.

Data collected in this survey is reflective of HPC users' management of new technologies and how they are explored and integrated into a production environment. Cloud environments can serve as a test bed for groups seeking to adopt AI technology without a sizeable buy-in of state-of-the-art AI specific hardware, which can often be more expensive or have a protracted acquisition process. Users are more likely to run experiments, test jobs, and even carry out production level runs in the cloud despite having on-premises resources, even if it is more costly on the front end if it achieves certain operational goals.

As AI technology matures within the advanced computing world, the cloud will continue to play a pivotal role. With that considered, there will not always be a wave of curiosity and experimentation with this new technology. As workload parameters eventually become more well understood within organizations and their industries, optimized solutions will become easier to come by and will ‘float to the top.’ As the standard means of meeting inference and training needs become clearer, the benefits of cloud as a test bed for new methodologies will need to be paired with a cost-effective, reliable means of providing organizations with efficient solutions.

APPENDIX: SURVEY DEMOGRAPHICS: RESPONDENTS AND ORGANIZATIONS

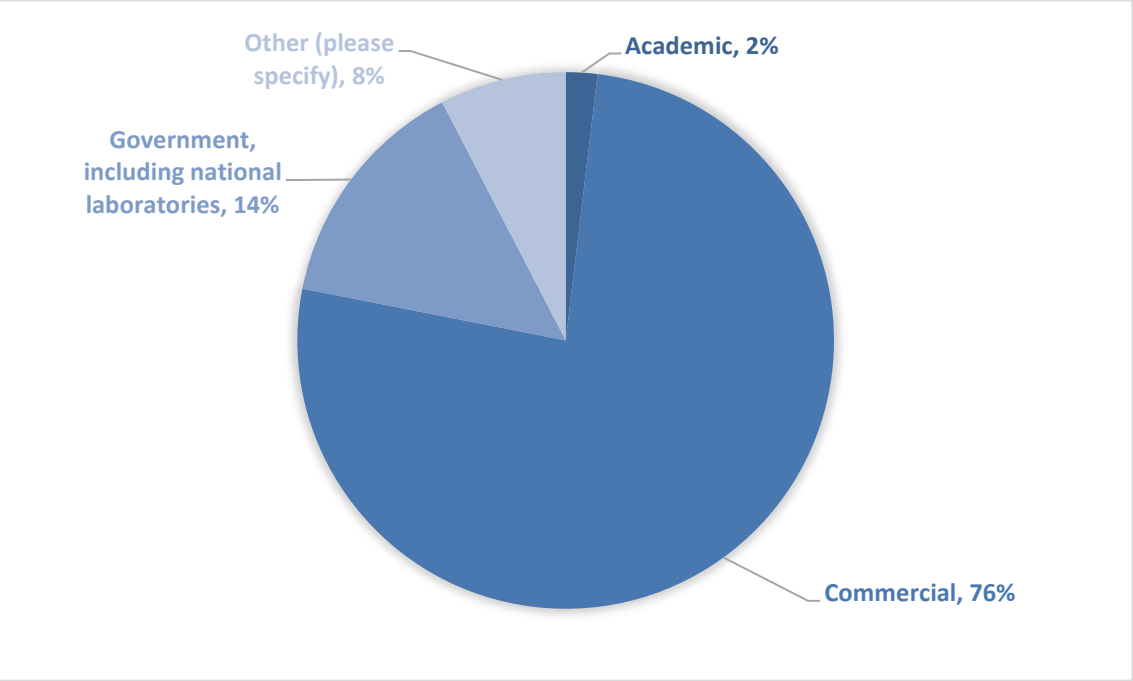
This appendix provides additional details on the demographics of the survey respondents and the organizations they represented. It consists of three main segments: individual survey respondents' background information, general demographics about the organization they represent, and select data on the budgetary levels of those organizations.

Respondent Demographics

Figure 9 breaks down respondent sector affiliation. This study focused generally on commercial respondents with consideration also made for governmental and academic respondents. 76% reported representing commercial or industrial organizations, 14% from government including national labs, and 2% from academic groups.

FIGURE 9

Respondent's Sector Affiliation



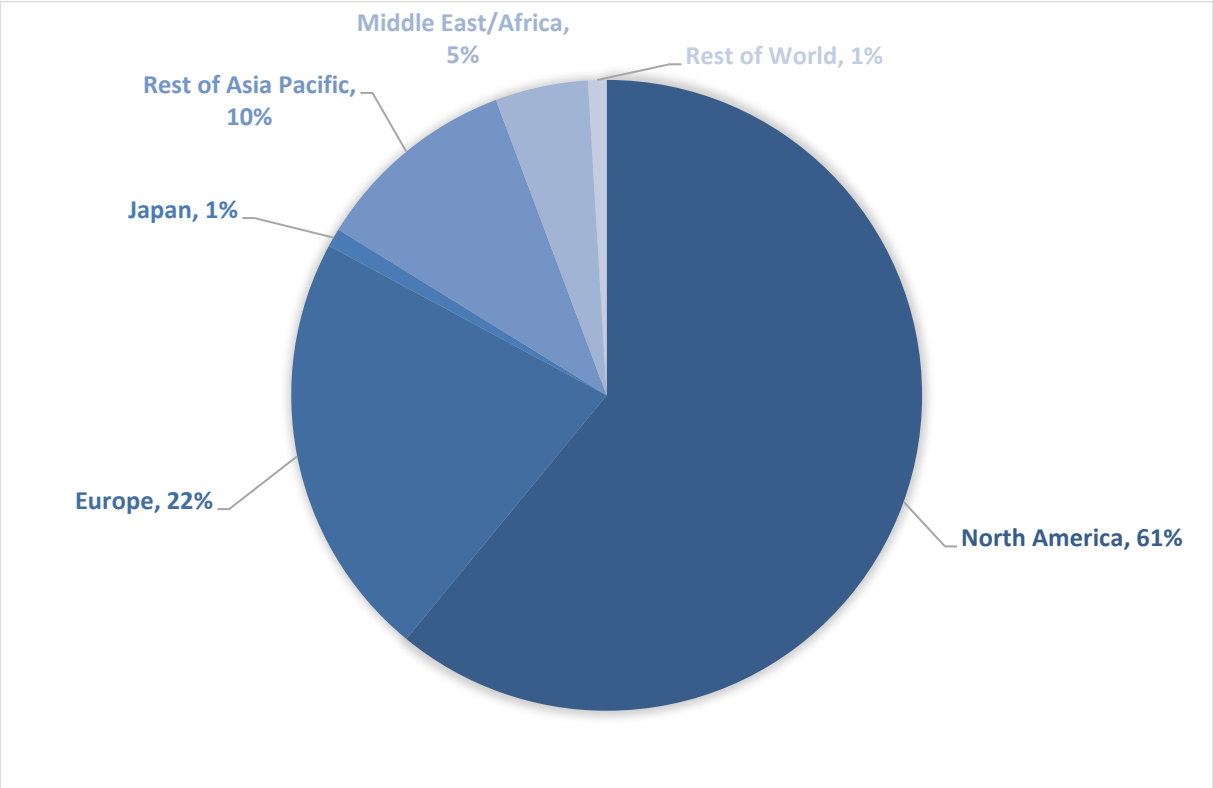
N= 105

Source: Hyperion Research, 2024

Figure 10 outlines the respondents' headquarters location. Over half (61%) were from North America, 22% were from organizations headquartered in Europe, Japan represented around 1%, and the rest of APAC was around 10%. The middle east, Africa, and the rest of the world contributed around 6% of responses.

FIGURE 10

Respondent's Organization Headquarters Location



N=105

Source: Hyperion Research, 2024

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). 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.HyperionResearch.com and www.hpcuserforum.com
