Special Report

Inferencing in an HPC/Advanced Computing Environment, Application Needs Must Be Well Understood

Tom Sorensen and Bob Sorensen
March 2025

EXECUTIVE SUMMARY

The purpose of this study was to gain a better understanding of the activities and use behaviors of AI users actively engaging in inference processing. Key objectives included creating a picture of user goals for AI inference integration, their current and planned methodologies, budget allocations, expectations, and preferred hardware, software, and cloud resources for their HPC-centric inference endeavors.

Highlights from the study results include:

- Considerable budget allocations are directed towards the procurement and integration of resources specifically to meet AI inferencing needs.
- Spending for on-premises vs cloud resources is a nearly even split.
- There are numerous architectures and device types currently used to meet inferencing needs.
- Complexity, cost, and unexpected technical issues are among the top concerns for HPC users adopting inferencing technology.
- Installation of inference-specific servers and systems is quickly on the rise.

The survey, which was conducted in January 2025, collected input from 100 survey respondents who indicated current or planned use within the next 12-18 months of production or mission-related AI inferencing processes.

HPC users and organizations engaging in compute-intensive workloads, specifically those with scientific and engineering missions, are actively exploring and integrating inference technology into their workflows. Buy-in and confidence in the ameliorative benefits of AI technology for HPC are still high, and this sentiment is reflected most prominently in budget allocations. Training, the other main activity critical to a functioning AI system, is largely front-loaded in its requirements. Oftentimes, one or several large training sessions are then followed by smaller, iterative fine-tuning efforts. Inference, however, presents a greater demand on sustained functionality of hardware devices, software support, and user expertise. While training is almost never a 'one-and-done' activity, inferencing by its very nature presents nearly continuous compute demands and, in some application spaces, has an extremely high ceiling for scaling up.

Key Findings

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

When asked about their AI activities with inferencing, respondents most often indicated that they are currently engaged in exploratory and experimental stages of AI inference processing, even those who have already integrated these technologies into some portion of their production environments. Across the board, these respondents demonstrated preference towards on-premises over cloud options for AI inferencing workloads. However, just 16% of respondents reported their inference needs being met only on-premises, the rest using some level of cloud support to meet their needs. Exploratory, experimental, and testing activities are often associated with proportionally high cloud usage.

Key Finding #2: Respondents' greatest concerns centered on integration complexity, hardware/software cost, and technical issues.

When asked about the major barriers to their adoption and effective use of inferencing technology, respondents indicated that complexity of integration into existing HPC workflows is their highest concern. The cost of inference-specific hardware and software was also high among their listed barriers, as well as concerns regarding technical issues, specifically managing expandability and hallucinations.

Key Finding #3: General purpose servers and devices are more commonly used to meet inference needs than AI-specific hardware.

Despite a wide range of AI-specific processors and GPUs being available on the market, respondents indicated that their AI inferencing processes were mostly being carried out on general purpose GPUs and on non-inference specific servers. However, when asked about the nature of the cloud resources being leveraged to support their inferencing, respondents primarily reported using AI-specific resources. While not the majority, dedicated inference servers are on the rise, carrying out 30% of operations.

General purpose servers and devices are more commonly used to meet inference needs than AI-specific hardware.

Key Finding #4: AWS is being used nearly twice as much as its leading alternative to meet inferencing needs.

The main hyperscalers used to meet AI inferencing needs among HPC respondents are AWS, Microsoft Azure, and Google. 34% of users utilize AWS to carry out inferencing, about as much as Microsoft Azure and Google combined.

Key Finding #5: HPC users report a nearly even split between on-premises and cloud budgeting.

When asked to divide their AI/HPC computing budget by on-premises vs cloud spending, users reported a nearly 50/50 split between the two. Within that cloud spending, it is also a near split between private and public cloud. It is worth noting that only about 33% of respondents indicated that the cloud comprised 50% or more of their compute resources.

HPC users report a nearly even split between on-premises and cloud budgeting.

Key Finding #6: A plurality of the software resources being used to support AI inferencing is open source.

Open-source software remains a powerful force in HPC/AI usage with roughly 25%, a plurality, of HPC users indicating their software is open-source. GPU supplier software is the second most common response, with in-house software third. A robust open-source ecosystem represents healthy innovation in the technology space and is also an effective method of reducing costly procurement of proprietary tools.

Key Finding #7: A significant majority of respondents are already using liquid cooling or plan to use some kind of liquid cooling soon.

Liquid cooling, for a staggering 84% of respondents, is already in use or is a quickly approaching eventuality within their organization. Cooling and power are a looming issue for those continuing to scale up their AI efforts, and the integration of AI technology within on-premises systems hastens the adoption of these infrastructural upgrades.

Key Finding #8: Nvidia has become a major supplier of inferencing systems.

Nvidia maintains a confident leadership position in supplying the hardware for AI inference, training, and other accelerator uses. Notably, Nvidia software offerings remain a substantial draw for users when considering procurement options. This advantageous market position is expected to continue for the foreseeable future.

Summary and Next Steps

The AI inference 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 fast-paced and diverse range of AI and inference-specific hardware offerings (both in the cloud and on-premises), a healthy open-source software ecosystem, and the advancement of infrastructural developments like liquid cooling.

While AI inferencing is part of the larger process of AI/HPC integration along with processes like training and fine tuning, it presents a unique set of needs in terms of compute resources; as with other similar integrative processes, users often turn to the cloud for exploration, testing, and piloting. While there is still high trust in the future efficacy and return for these efforts, concerns surrounding cost, difficulty, and looming technical issues are still prominently represented in respondent sentiment.

Currently, inferencing does not necessitate an AI-specified piece of hardware, or even appliance type, and there is no indication that all inference jobs will ever require inference-specific hardware. Outputs can take on virtually any form of data, from a handful of characters to 3D mesh models and beyond. As such, application-specific knowledge will go a long way in realizing cost-effective and investment-returning inferencing architecture. This knowledge requires a wealth of expertise both in any organization's field of practice and AI inferencing.

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 user behaviors surrounding the inference portion of AI/HPC adoption and integration. For the purpose of this study, AI inference activity 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 and indicated that they are currently or planning to run inference workloads. This effort took a broad-based approach to AI that included the major categories of machine leaning, deep learning, and generative AI, particularly large language models.

Research Approach

This study is based on an independent survey of organizations currently involved in AI inferencing for HPC/AI workload integration or production use, or those 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 inference integration.
- Assessing the current devices, processor types, and server configurations used to meet inference needs.
- Understanding the computing landscape used in AI inference integration and usage within organizations that commonly use HPC.
- Characterizing the interest and expectations of AI inference among users currently or soon planning to integrate AI technology into existing HPC workflows.
- Exploring the details of cloud offerings, on-premises devices, and software sources used to support inference technology.
- Identifying budgetary details to support AI inference and HPC support.

Respondents came from a mix of major sectors: commercial (63%), academic (23%), and government (14%), representing verticals led by computers and related electronics but also including the financial sector, bioscience, advanced manufacturing, and geosciences.

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 progress, practices, and expectations of HPC users regarding their ongoing AI inferencing activities. The survey design was based on the following goals:

- Understanding the scope, expectations, and goals of organizations managing inferencing processes to support existing HPC workloads.
- 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 inferencing activities 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 AI Inference Activity in Organization's Existing HPC/AI Workloads

Activity	% Selected
Exploring the range of potential performance enhancements by integrating inferencing technology into existing HPC-based scientific and engineering workloads	57.0%
Exploring in-house requirements for integrating inferencing into HPC-based scientific and engineering workloads	52.0%
Testing/assessing inferencing-integrated workload performance	39.0%
Running production level inferencing-enabled workloads	37.0%
Procuring access to necessary inferencing software	30.0%
Procuring access to necessary inferencing hardware	27.0%
Passively monitoring inferencing technology developments	26.0%
Porting inferencing capability into existing workloads	25.0%
Standing up limited inferencing-integrated pilot programs	23.0%
Reaching out to inferencing hardware and software suppliers for information	22.0%
Standing up fully funded inferencing research efforts	17.0%
No current activity or Don't know/Not sure	1.0%
Other	3.0%

N = 100

Respondents could select all options that apply. Those who selected "Other" specified classified work

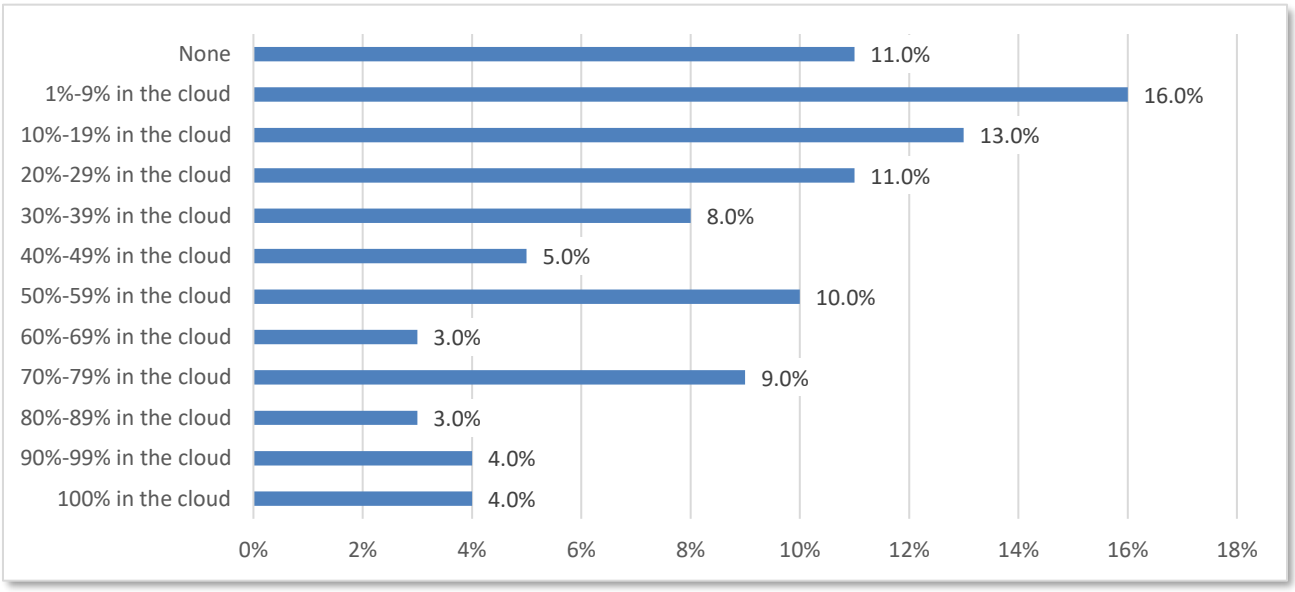
Source: Hyperion Research, 2025

Figure 1, below, focuses on the ratio between cloud and on-premises processing for HPC/AI usage as a whole. The majority of users indicated using the cloud for less than 50% of their total compute resources, with 33% reporting using the cloud for more than half their total compute resources. Leveraging cloud resources for advanced computing needs continues to rise in popularity, especially among those users engaging in AI integration processes. Cloud resources and hyperscalers are often considered a cost-effective, low-committal method of exploring new options and piloting test projects for new technologies.

Previous Hyperion Research studies have collected similar data, and comparisons between previous studies and this data support the trend of a migration towards on-premises/cloud hybrid solutions as a means of furthering application specificity for compute resources, with cloud resources offering diverse, on-demand resources, and on-premises offering long-term, more cost-predictable solutions.

FIGURE 1

Percentage of HPC/AI Workload in the Cloud vs. On-premises



N = 100

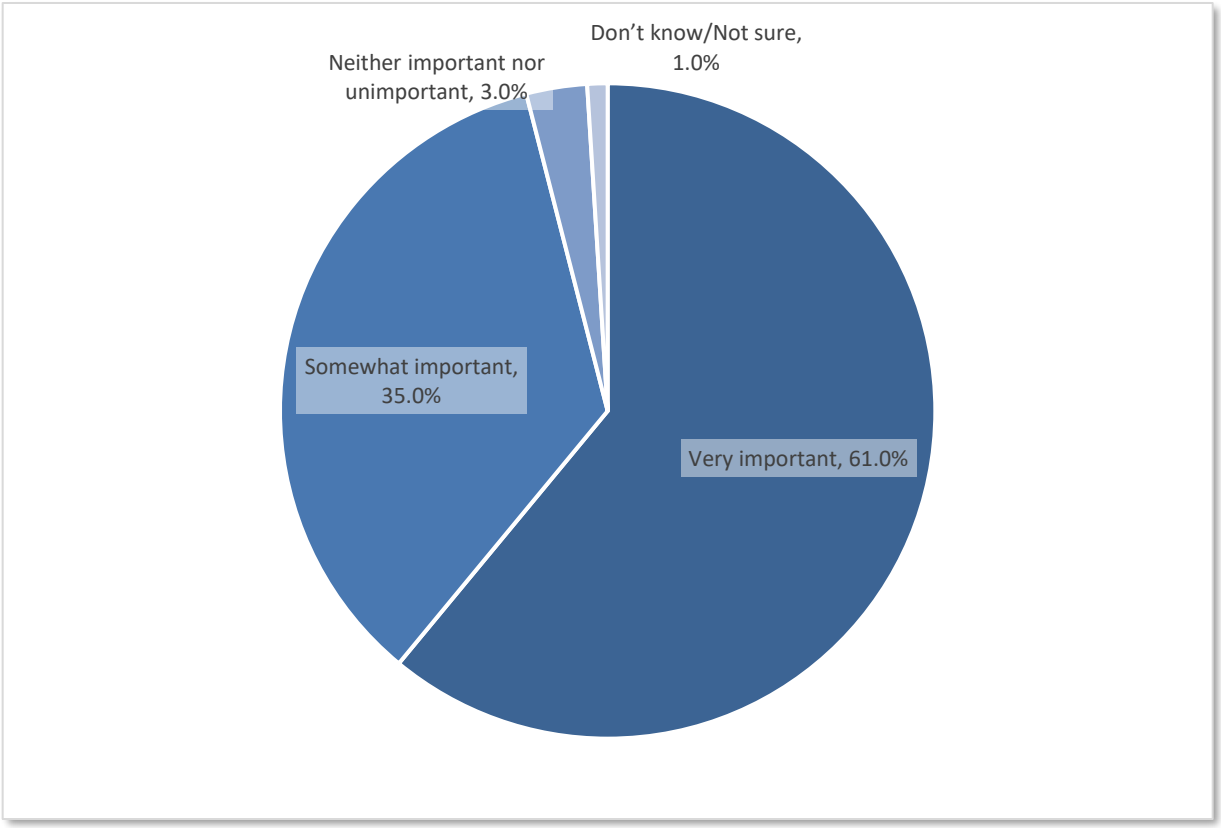
Source: Hyperion Research, 2025

Figure 2, below, reveals the sentiments among HPC users concerning the importance of the role of inference in their current or planned workloads. With 61% reporting high levels of importance and another 35% reporting a level of somewhat important, inference workloads are top-of-mind when it comes to planning the future of AI integrated HPC.

- While there are concerns for difficulty of integration, operational costs, efficacy, and budgetary allocations, detailed in a later table, this reflects the overall high confidence in AI inference endeavors.
- The amount of 'cloud native' or cloud-only users is on the rise. This class of HPC user is set apart from those using on-premises or hybrid compute resources for many reasons, notably their lack of exposure to the influences of the hardware market, infrastructure demands like power and cooling, and server/device maintenance and renewal.

Figure 2

Importance of the Use of AI Inferencing to Support Current or Planned HPC/AI Workloads



N = 100

Source: Hyperion Research, 2025

Exploring Barriers and Goals for AI Inferencing

Table 2 arrays the most common barriers or challenges for AI inferencing efforts among surveyed organizations. Complexity of integration is a standout among these challenges with a selection rate of nearly half of respondents (47%). Integration of new technology always presents critical challenges, even more so for the mission critical and possibly legacy science and technology workloads these users carry out.

Not far behind are cost concerns, specifically that of inference-specific hardware or software. AI-specific hardware can be costly, slow to acquire, and possibly outside of or peripheral to the knowledge base of in-house expertise. These forces can incentivize cloud migration for some users, at least temporarily. Similarly, the actual or perceived high cost of AI inferencing-specific software is likely

driving users to leverage open-source inference software products, a behavior detailed more fully later in this report.

Concerns with technical issues surrounding inferencing such as expandability and hallucinations is a close third among the top concerns. This concern is of particular importance for users engaging in high-precision, low-tolerance science and technology projects in which inferencing outputs are less easily validated than, for instance, code rewriting or chatbot models.

Table 2

Top Three Key Challenges with Using AI inference for HPC/AI Workloads

Challenge	Currently
Complexity with integrating inferencing into existing HPC-based scientific and engineering workloads	47.0%
Concerns with cost of inferencing-specific hardware or software	31.0%
Concerns with technical issues surrounding inferencing such as expandability and hallucinations	29.0%
High/uncertain operational costs	24.0%
Uncertainty about the right application or hardware or software to use	24.0%
High/uncertain development costs	23.0%
Too computationally intensive	22.0%
Lack of in-house expertise in inferencing	16.0%
The technology is moving too fast for credible assessment of value	16.0%
Long/uncertain implementation times	15.0%
Lack of demonstrated return on investment	12.0%
Lack of reproducibility	11.0%
Lack of precision	10.0%
Confusion/uncertainty with inference vendor selection	8.0%
Uncertainty of demonstrated computational performance improvements	7.0%
Other	5.0%

N = 100

Respondents could select multiple options.

Source: Hyperion Research, 2025

Table 3 characterizes those operational goals considered most important by integrating AI inferencing capability into an organization's base of existing HPC workloads. The goal selected most often among HPC users is for AI inferencing to usher in capabilities (insights or knowledge) beyond the purview of traditional systems by themselves (47%). It is important to note that this goal does not necessarily imply that an AI model or AI inferencing will subsume or replace the processes carried out by traditional compute methods.

Another standout response is the second most selected goal envisioned by integrating AI inferencing into existing HPC/AI workloads: a faster time to solution for existing HPC-based science and engineering solutions (44%). Most HPC users consider AI integration as a means of refining and improving existing workflows, not creating entirely new ones. Data identification and processing, efficiency monitoring and insights, usage reports, and post-processing are some areas considered most potentially fruitful for AI inferencing.

Table 3

Most Important Goals Envisioned by Integrating AI inference into Existing HPC/AI Workloads

Goals	% Selected
Providing new insights or knowledge not possible on traditional systems by themselves	47.0%
Faster time to solution for key HPC-based scientific or engineering solutions	44.0%
Opening new lines of HPC-based scientific and engineering research	23.0%
Greater efficiency on existing HPC-based scientific and engineering workloads	20.0%
Greater computational fidelity on key HPC-based scientific and engineering workloads	14.0%
Increasing revenue	12.0%
Lowering overall HPC software costs	10.0%
Lowering overall HPC hardware costs	9.0%
Opening new lines of multiple vertical end uses	8.0%
Opening new lines of vertical-specific end uses	6.0%
Reducing HPC staffing requirements	4.0%

Table 3

Most Important Goals Envisioned by Integrating AI inference into Existing HPC/AI Workloads

Goals	% Selected
Lowering overall HPC power requirements	2.0%
Other	1.0%

N = 100

Respondents could select multiple options

Source: Hyperion Research, 2025

According to Table 4, over a quarter of respondents reported using open-source software for their AI inferencing needs. With hardware/software cost concerns being prominently represented among key barriers, open-source software provisioning is an effective way of maximizing budget allocations, especially among those users still in an exploratory or experimental stage.

The prevalence and apparent effectiveness of open-source software solutions speaks further to the confidence of users for future AI integration payoff. A healthy, robust open-source ecosystem can contribute critically to the advancement of a growing technology.

GPU suppliers also comprise a considerable share of inference software usage. With AI effectiveness and GPU technology being inextricably linked, GPU and other processor suppliers are expected to continue to offer not just state-of-the-art devices but cost-effective proprietary support software. In this way, the effectiveness of GPU hardware is not solely linked to the server device itself but also to the software and its continued maintenance.

- It is expected that as AI expertise within HPC organizations continues to grow, in-house development of inference support software and other AI related tools will grow.

Table 4

Source of AI Inferencing Software

Software Source	% Selected
Open source %	25.4%
GPU supplier %	19.5%
In-house developers %	14.6%
General-purpose computer systems supplier %	14.4%

Table 4

Source of AI Inferencing Software

Software Source	% Selected
General purpose AI-centric software firms %	8.2%
AI-centric accelerator supplier %	5.8%
Vertical or sector specific inferencing software companies %	3.1%
Hyperscale/CSP software offerings %	2.1%
Don't know/ Not sure %	5.4%
Other %	1.5%

N = 100

Source: Hyperion Research, 2025

Sourcing AI Inference Technology and Resources

Table 5 arrays a simple breakdown of inference workloads by the source of their compute. Users most often selected a primarily on-premises compute environment supported by cloud resources (35%), but also noticeably represented were an even mix of on-premises and cloud resources (20%) and cloud-heavy resources (19%).

- Notably, when correlated with Table 1, some users who use a mix of on-prem and cloud for their overall HPC/Technical computing workloads continue to carry out inferencing entirely in the cloud.

Table 5

Method of Supporting Existing/Planned AI Inferencing for Compute Needs

Method	% Selected
Mostly on-premises hardware, but also supported by cloud resources	35.0%
An even mix of on-premises and cloud resources	20.0%
Mostly cloud resources, but also supported by on-premises hardware	19.0%
Entirely on-premises hardware	16.0%
Entirely cloud resources	8.0%
Don't know/Not sure	2.0%

N = 100

Source: Hyperion Research, 2025

Table 6 arrays the supplier market landscape used for standing up inference related systems. The plurality of responses went to NVIDIA at just over a quarter of total responses (25.8%), with Dell (16.6%) and HPE (12.3%) in second and third.

- While market favorability of NVIDIA GPUs and accelerators is currently undeniable, their software support offerings play a considerable role in their popularity as well.

Table 6

Server Suppliers Used to Meet On-premises AI Inferencing Needs

Server suppliers	% Selected
NVIDIA %	25.8%
Dell %	16.6%
HPE %	12.3%
IBM %	9.2%
Fujitsu %	3.4%
Cisco %	3.3%
Supermicro %	3.2%
Lenovo %	2.8%
ASUS %	2.0%
Eviden/ATOS %	2.0%
Gigabyte %	1.3%
NEC %	1.3%
Lambda %	1.1%
ASA %	0.5%
Quanta %	0.5%
Liqid %	0.2%
Inspur %	0.2%

Table 6

Server Suppliers Used to Meet On-premises AI Inferencing Needs

Server suppliers	% Selected
Penguin %	0.2%
Exxact %	0.2%
Don’t know/Not sure %	11.0%
Other %	2.8%

N = 92

Respondents could select multiple answers

Source: Hyperion Research, 2025

Table 7 shows the prevalence among users of liquid cooling in current and planned servers used to support AI inferencing. Liquid cooling is quickly growing in adoption and usage, and the increase of GPUs and AI-specific processors to meet project needs is a driving factor.

Table 7

Percentage of Liquid Cooled AI Inferencing Servers Current/Planned

Percentage	% Selected
None	16.3%
1%-9%	7.6%
10%-19%	4.4%
20%-29%	12.0%
30%-39%	5.4%
40%-49%	7.6%
50%-59%	5.4%
60%-69%	6.5%
70%-79%	5.4%
80%-89%	3.3%

Table 7

Percentage of Liquid Cooled AI Inferencing Servers Current/Planned

Percentage	% Selected
90%-99%	3.3%
All - 100%	6.5%
Don't know/Not sure	16.3%

N = 92

Source: Hyperion Research, 2025

Hardware Allocation, Sources, and Suppliers

Table 8 contains respondent data on primary GPU or accelerators used to meet HPC/AI related inference needs. Most respondents reported use of state-of-the-art high-performance GPUs (37%) such as NVIDIA Blackwell, AMD MI300X, or NVIDIA H200. Many respondents, however, are effectively using mainstream general-purpose GPUs (27.2%).

Relatively few respondents indicated the use of AI-specific accelerators from commercial suppliers such as Cerebras, Graphcore, and SambaNova. These numbers could rise as user groups expand exploratory and experimental stages and begin provisioning longer term solutions for on-premises inferencing.

- Respondents are more likely to use AI-specific cloud resources than AI-specific on-premises.

Table 8

Primary GPU or AI-Centric Accelerator Used for HPC/AI Inference Needs

Primary GPU or Accelerator	% Selected
High-performance GPUs (e.g. Blackwell, MI300X, H200)	37.0%
Mainstream GPUs (H100, MI100)	27.2%
Hyperscale AI-centric accelerators (e.g. AWS, Google, Microsoft)	17.4%
AI-centric accelerators from commercial suppliers (e.g. Cerebras, Graphcore, SambaNova)	7.6%
Don't know/Not sure	6.5%
Other	4.4%

Table 8

Primary GPU or AI-Centric Accelerator Used for HPC/AI Inference Needs

Primary GPU or Accelerator	% Selected

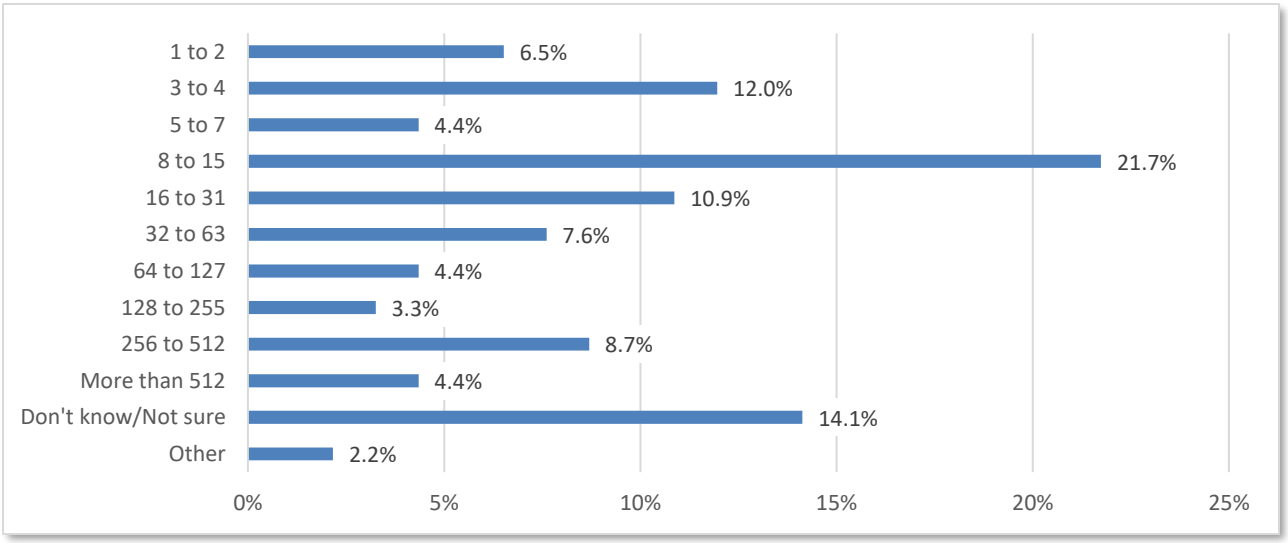
N = 92

Source: Hyperion Research, 2025

As Figure 3 demonstrates, the average number of GPUs or AI-specific accelerator use for inferencing servers was widespread, with the most popular option being between 8-15. This indicates a wide range of complexity in an applications output type or required performance.

FIGURE 3

Average Number GPUs or AI-specific Accelerators in AI Inferencing Servers



N = 92

Source: Hyperion Research, 2025

As seen in Table 9, inference needs are most often being met by GPUs alone (47.5%), either by general purpose GPUs (26.7%) or AI-specific GPUs (20.8%). A mix of general-purpose GPUs and CPUs is the second most common form of hardware allocation (21.1%).

While it is anticipated that inference-specific processors will grow in popularity, they are underrepresented among HPC users (2.2%) and could represent a viable area for innovation as inference needs become more clearly understood.

Table 9

Specific Processor/Accelerator Type Used to Meet On-premises AI Inferencing Needs

Processor type	% Selected
General purpose GPUs %	26.7%
AI-specific GPUs %	20.8%
A mix of general-purpose GPUs and CPUs %	21.1%
A mix of AI-specific GPUs and CPUs %	13.7%
CPUs only %	9.3%
Inference-specific processors %	2.2%
No on-premises AI inferencing %	0.8%
Don’t know/Not sure %	5.3%
Other %	0.2%

N = 92

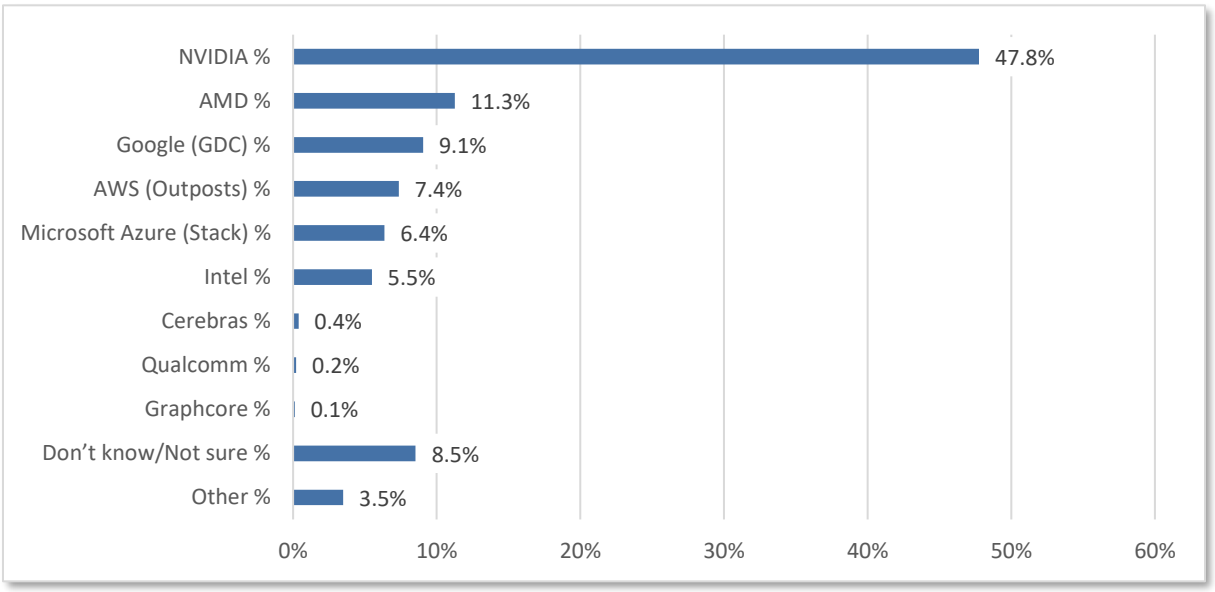
Source: Hyperion Research, 2025

It should come as no surprise that NVIDIA remains the top supplier for GPUs or related AI-centric accelerators in the next three years, with 47.8% of respondents expecting to use NVIDIA hardware for their AI inferencing needs. With their reputation for providing and, in some cases, leading the way for AI-related processors combined with industry favored software support, NVIDIA is expected to remain at the top of this heap for the foreseeable future.

- NVIDIA's dominance in this area does not imply a lack of innovative and market influential moves being made by competitors like AMD but instead points to its long-established presence as a leading-edge supplier of relevant hardware and software.

FIGURE 4

Anticipated Supplier of GPUs or Related AI-centric Accelerators in the Next Three Years to Support On-premises AI inferencing



N = 92

Source: Hyperion Research, 2025

Table 10 shows that the plurality of respondents is still using general purpose servers to meet their inference needs (47.1%). The number of respondents using dedicated inference servers (30.7%) is expected to rise as production level AI inference continues to enter production workflows.

- With so many organizations admittedly still exploring/experimenting with AI inference and its contribution to their existing HPC workloads, these numbers are expected to change as inferencing needs become more clearly understood and are better designed for more effective scaling.
- Such scaling efforts will likely lead to a higher prevalence of dedicated inference servers.

Table 10

System-level Devices to Support On-premises AI Inferencing

Device	% Selected
General purpose server(s) %	47.1%
Dedicated inference server(s) %	30.7%
Desktop/laptop devices(s) %	10.1%

Table 10

System-level Devices to Support On-premises AI Inferencing

Device	% Selected
Edge device(s) %	5.2%
No on-premises inferencing %	2.6%
Don't know/Not sure %	4.4%
Other %	0.0%

N = 92

Source: Hyperion Research, 2025

Nearly half of all HPC related inferencing jobs carried out by respondents (42%) are handled by 1-4 GPU/accelerators. That said, the reported accelerator count for a given inference job has a wide variation, another indicator of diversified output types in terms of complexity and time to completion.

- Users are most likely to fall between using less than 8 or more than 32 GPUs/accelerators.

Table 11

GPU/AI Accelerator Count Typically Used For Each HPC/AI Inferencing Job

GPU/Accelerator count	% Selected
1 to 2	17.0%
3 to 4	25.0%
5 to 7	15.0%
8 to 11	18.0%
12 to 15	8.0%
16 to 32	6.0%
Greater than 32	11.0%

N = 100

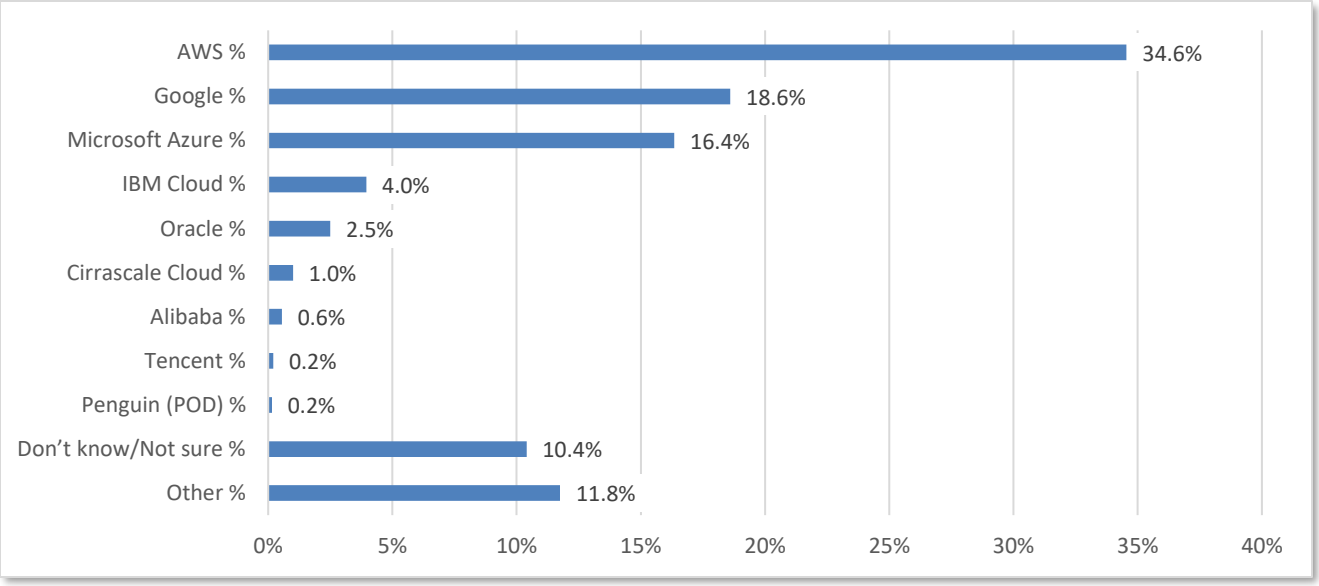
Source: Hyperion Research, 2025

Figure 5 details the hyperscalers or other CSPs used to source compute capacity for inferencing needs. AWS, in line with other previous Hyperion Research studies, remains on top (34.6%) with roughly the sum of its two leading competitors, Google (18.6%) and Microsoft Azure (16.4%). AWS was quick to the punch in terms of market visibility and cloud development to address inferencing needs with inference specific hardware options and architectures.

- With no majority representation among CSPs/hyperscalers, the landscape of compute provisioning for inference needs is more diverse and evenly spread than on-premises device suppliers.

FIGURE 5

Hyperscalers or Other CSPs Currently Used to Support AI Inferencing Needs



N = 100

Source: Hyperion Research, 2025

Table 12 details the supplier-identified compute resource type or instance offered by application specificity and its use among HPC organizations.

As expected, AI-focused instances are the most common, but not by a wide majority (29.1%). Compute-focused (21.4%) and HPC-optimized (20.8%) are not far behind with a sharp drop-off afterwards.

- Workload and application specificity is a major advantage of cloud usage, especially during early stages of technology adoption.
- This data may indicate a disconnect between the expectations of application needs between CSPs/hyperscalers and HPC users.

Table 12

Current or Planned Cloud-based Inferencing by Hyperscaler-defined Instance Type

GPU/Accelerator count	% Selected
AI focused %	29.1%
Compute focused %	21.4%
HPC optimized %	20.8%
Accelerated computing focused %	7.6%
Storage focused %	4.2%
Memory focused %	4.1%
Don’t know/ Not sure %	7.8%
Other %	5.0%

N = 100

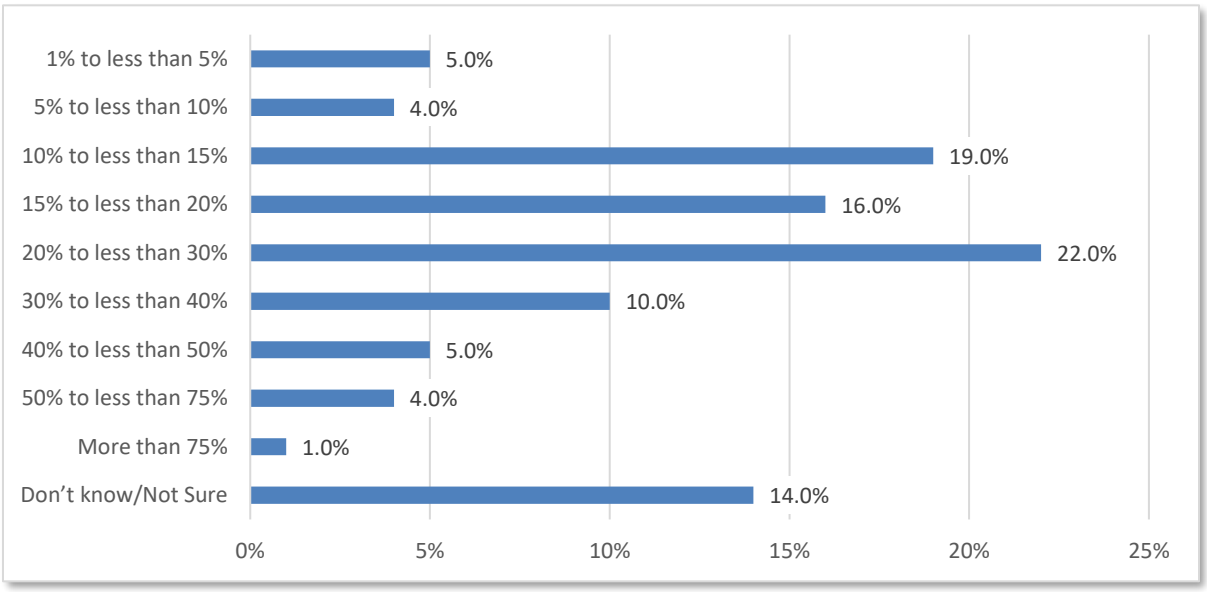
Source: Hyperion Research, 2025

Figure 6 is the first of multiple data sets detailing budget information regarding HPC/AI/Technical computing and inference allocation both in the cloud and on-premises. With 42% of respondents indicating that over 20% of their entire advanced computing budget is dedicated to procuring and maintaining AI inference capabilities, the confidence and financial commitment to the technology is high.

- This data represents only the budget dedicated to inferencing capabilities. The budget allocation for AI capabilities as a whole including training, fine tuning, and other supportive technology would be considerably higher.

FIGURE 6

Percentage of Overall Advanced Computing Annual Budget for Procuring and Maintaining AI inferencing Capabilities



N = 100

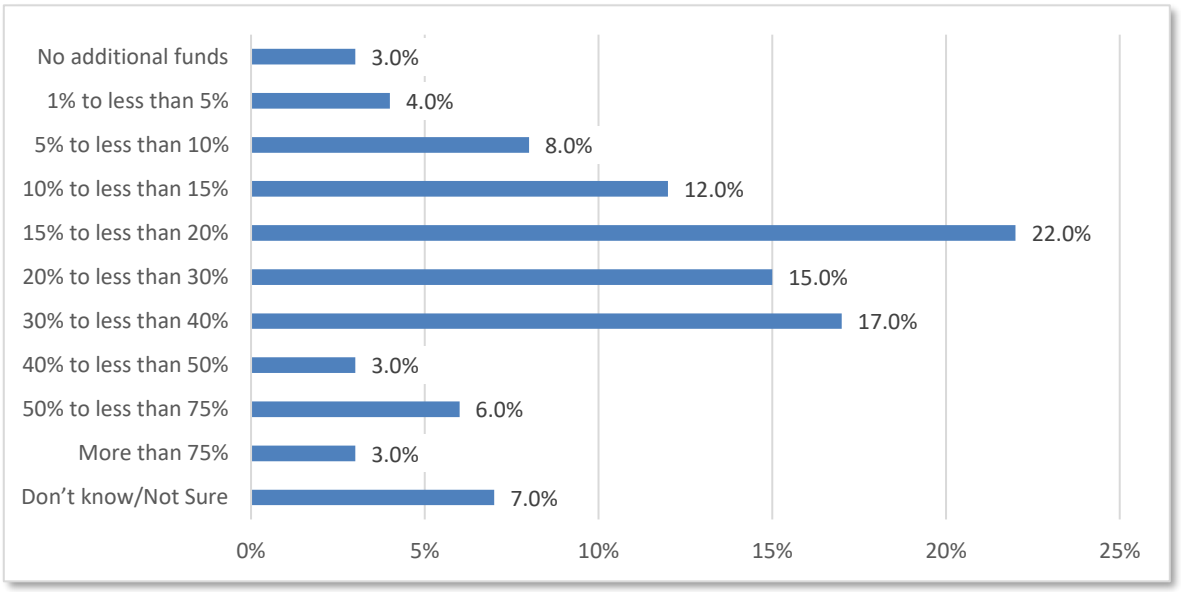
Source: Hyperion Research, 2025

As demonstrated in Figure 7, wherein respondents were asked to envision the same budget allocation in a five-year timeframe, the expectation for inference budget is increasing. This indicates a high confidence in the technology, its ability to return investments, and scale up.

- These allocations include both cloud and on-premises budget, which will be broken out in a later question.
- The data also indicates the expectation among users that inferencing technology will demand continued maintenance and upcycle costs likely in the form of hardware upgrades, increased power consumption, and continual expertise acquisition.

FIGURE 7

Envisioned Annual Percentage of Budget for AI Inferencing in Five Years



N = 100

Source: Hyperion Research, 2025

The budgets of organizations represented in this study are widely varied. Table 13 contains data on the annual budgets from representative organizations for both on-premises and cloud-based HPC support. It shows a generally even distribution, with the plurality of 39% between $1M and $20M.

- Respondents reported similar concerns regarding cost, technical issues, and complexity regardless of budget.
- Issues such as expertise availability most severely affect those organizations with the lowest budgets.

Table 13

Total Annual Budget for On-premises and Cloud-based HPC/AI Workloads

Budget range	% Selected
Under $100,000	1.0%
$100,000 to less than $250,000	5.0%
$250,000 to less than $500,000	7.0%
$500,000 to less than $1 million	7.0%

Table 13

Total Annual Budget for On-premises and Cloud-based HPC/AI Workloads

Budget range	% Selected
$1 million to less than $5 million	17.0%
$5 million to less than $10 million	12.0%
$10 million to less than $20 million	10.0%
$20 million to less than $50 million	5.0%
$50 million to less than $100 million	5.0%
$100 million to less than $500 million	11.0%
More than $500 million	2.0%
Don't know/Not sure	18.0%

N = 100

Source: Hyperion Research, 2025

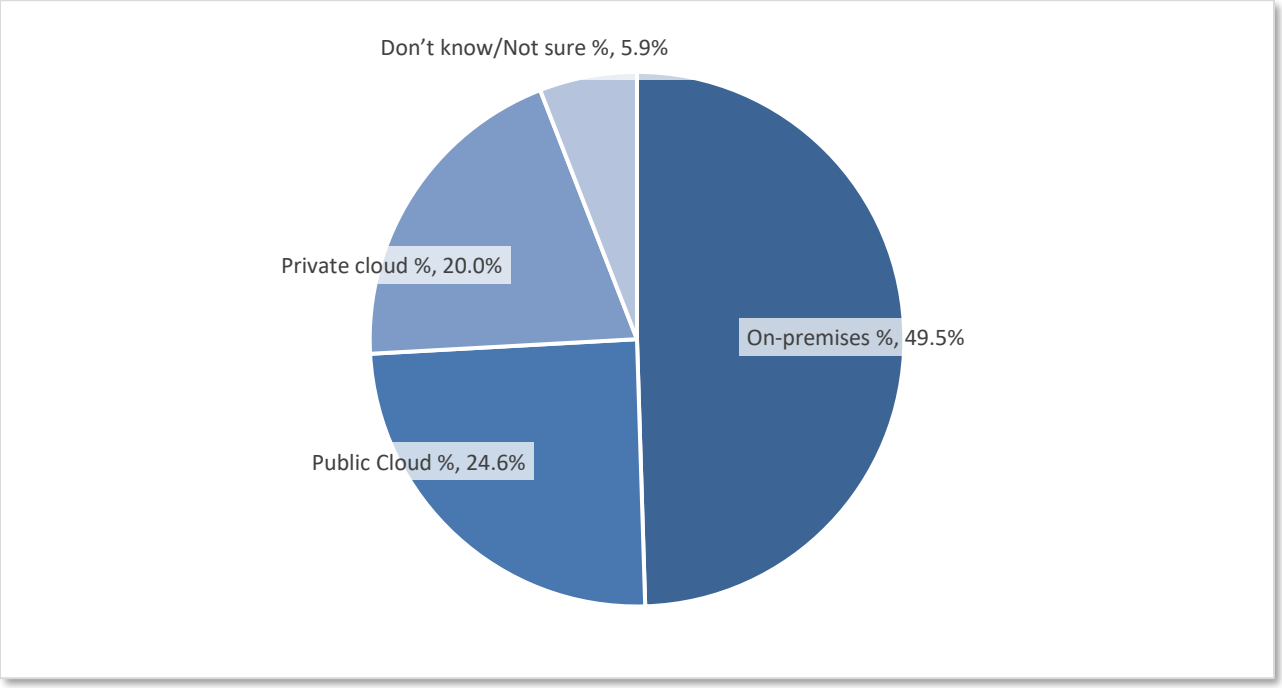
When asked how their advanced computing budgets are divided between on-premises and cloud (both private and public), there was a near even split between cloud resources and on-premises (see Figure 8). This data is of particular interest considering the apparent preference among this user group for on-premises computing. This could be an indicator of a higher cost of compute sourced from the cloud.

This ratio will be closely watched in proceeding studies, as it is expected that the maturation of AI integration processes will lead to a migration towards on-premises resources to meet inference needs. Scaling up in an on-premises environment is traditionally considered a lower cost in the long term but only if the applications are well understood and well-suited hardware is procured.

- High levels of cloud investment are common in experimental stages of technology adoption.
- More established HPC organizations may seek to avoid using cloud resources for production tier scale-ups, opting for the control and ownership offered by on-premises procurement.

Figure 8

Current Annual Budget Allocations to Support On-premises, Public Cloud, and Private Cloud



N = 100

Source: Hyperion Research, 2025

SUMMARY AND NEXT STEPS

HPC users and organizations engaging in compute-intensive workloads, specifically those with scientific and engineering missions, are actively exploring and integrating inference technology into their workflows. Buy-in and confidence in the ameliorative benefits of AI technology for HPC are still high, and this sentiment is reflected most prominently in budget allocations. Training, the other main activity critical to a functioning AI system, is largely front-loaded in its requirements. Oftentimes, one or several large training sessions are then followed by smaller, iterative fine-tuning efforts. Inference, however, presents a greater demand on sustained functionality of hardware devices, software support, and user expertise. While training is almost never a 'one-and-done' activity, inferencing by its very nature presents nearly continuous compute demands and, in some application spaces, has an extremely high ceiling for scaling up.

As revealed in this study, HPC users currently engaged in or looking to leverage inference technology are concerned not only with how to make it work properly and efficiently but how to keep the inevitably costly maintenance and continual renewal costs as low as possible.

Many users are admittedly still exploring options and experimenting with how to best deploy integrated inference technology, and their use patterns indicate that this process, across industries, is largely still one in transition.

- Overall, cloud resources are most favored for pilot projects, test phases, and experimentation.
- Despite the high cost of cloud resources, the instant access to new and diverse technology incentivizes cloud use for those not yet committed to a single method of integration.
- Complexity of integration is top of mind for users in terms of barriers, a concern that could serve to elongate the exploration/experimentation stages of integration.

With budgets for cloud and on-premises HPC resources evenly split, yet considerably more work being done in an on-premises environment, the cloud may not be the ideal environment for a stable, highly cost-efficient inferencing workflow. Elements keeping organizations from realizing a fully efficient inference arm of their HPC ecosystem include reconciling application specifics with the diverse hardware offerings available, questions regarding scaling up, and managing the verification and validation processes required to identify hallucinations among outputs. For most, navigating and solving these barriers is best done in the cloud where there is a low barrier of commitment, and high technical support, and accessing and scaling different hardware can be done with the press of a button.

Currently, inferencing does not necessitate an AI-specified piece of hardware, or even appliance type, and there is no indication that all inference jobs will ever require inference-specific hardware. Outputs can take on virtually any form of data, from a handful of characters to 3D mesh models and beyond. As such, application-specific knowledge will go a long way in realizing cost-effective and investment-returning inferencing architecture. This knowledge requires a wealth of expertise both in any organization's field of practice and AI inferencing.

AI inferencing, as one of the critical pillars of a functioning integrated AI framework, presents unique challenges for users and site-runners in terms of scalability, procurements and continued cost, and technical performance. For now, many users have turned to the cloud to explore options and develop usable, efficient systems. A migration towards on-premises hardware solutions is expected as compute needs become better understood and the cost-demands of scaling continue to grow. While the ability for AI to usher in a higher level of investment return and efficiency in HPC environments is largely undebated, the ability to fully realize those boons lies in an organization's willingness and capability to assess their application requirements, identify appropriate hardware/software resources, and scale up in a manner appropriate to their needs. Inference technology and its infrastructure is not one-size-fits-all, and as such, its effective use requires a deep understanding of a project's goals and requirements.

Application specific knowledge will go a long way in realizing cost-effective and investment-returning inferencing architecture.

APPENDIX: SURVEY DEMOGRAPHICS: RESPONDENTS AND ORGANIZATIONS

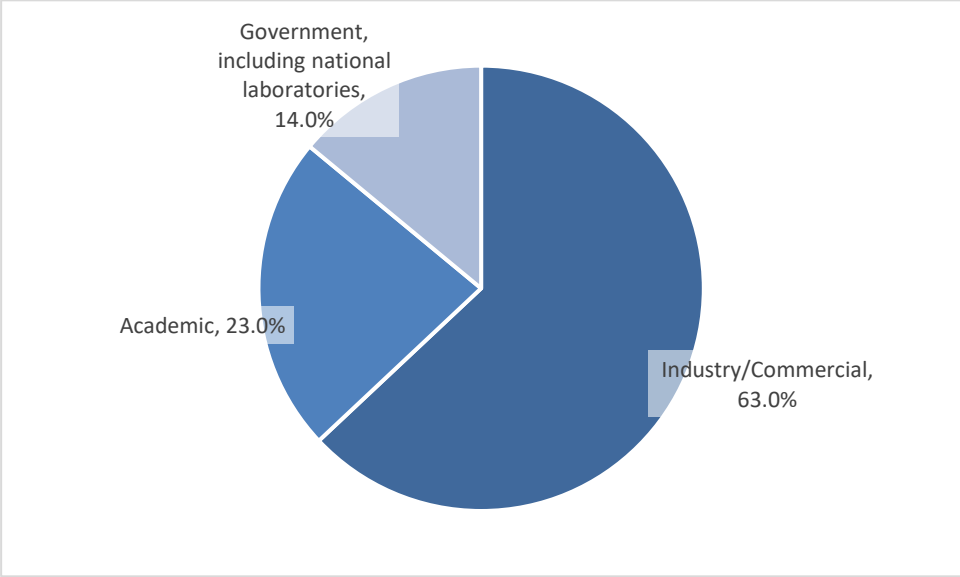
This appendix provides additional details on the demographics of the survey respondents and the organizations they represent. It consists of three main segments: individual survey respondents' background information, general demographics about the organization they represent, and select data on the budgetary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

Respondent Demographics

Figure 9 breaks down respondent sector affiliation. This study focused generally on commercial respondents with some consideration also made for governmental and academic respondents. 63% reported representing commercial or industrial organizations, 14% from government including national labs, and 23% from academic groups.

FIGURE 9

Respondent Sector Affiliation



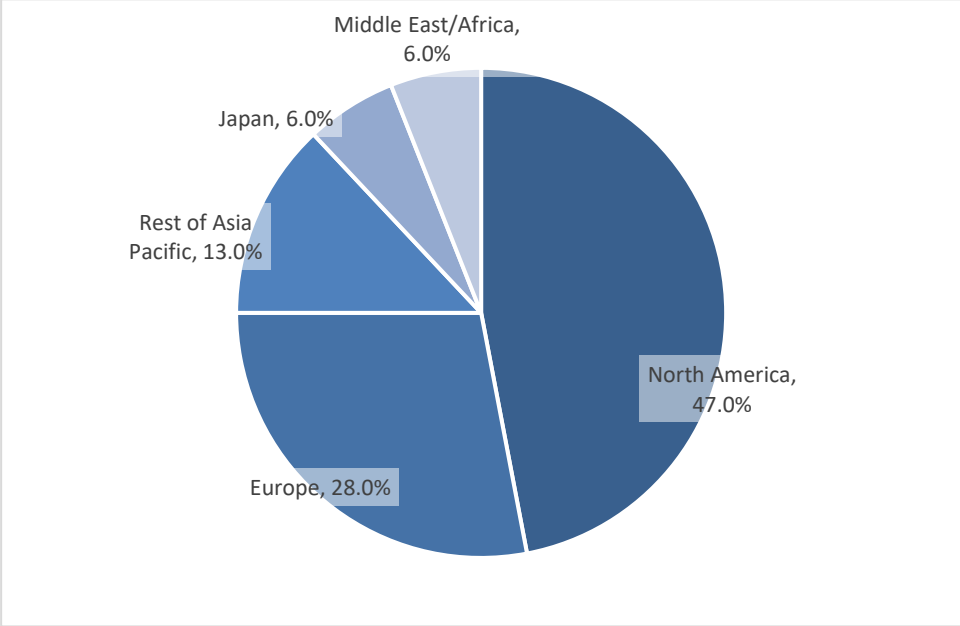
N= 100

Source: Hyperion Research, 2025

Figure 10 outlines the respondents' headquarters location. Nearly half (47.0%) were from North America, 28.0% were from organizations headquartered in Europe, Japan represented 6.0%, and the rest of APAC was around 13.0%. The Middle East, Africa, and the rest of the world contributed approximately 6.0%.

FIGURE 10

Respondent Organization Headquarters Location



N=100

Source: Hyperion Research, 2025

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 AI. 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
