AI-at-HYPERION: A Special Analysis

AI Investments for HPC and Advanced Computing Continue to Expand and Begin to Show ROI Results

Tom Sorensen, Bob Sorensen, and Earl Joseph
November 2025

EXECUTIVE SUMMARY

The purpose of this study was to gain a better understanding of the return on investments made by AI/HPC across critical science and engineering verticals. Key goals included creating a picture of investment expectations, snapshots of current integration projects in progress, budget allocations, provisioning methods, future plans, and challenges involving achieving desired goals in these areas.

The report reveals a strategic shift among high-performance computing (HPC) and scientific users in their approach to generative AI integration. While continuing a pace of relatively rapid expansion, organizations are now focused on refining existing implementations and responding more deliberately to success metrics, whether by scaling, maintaining, or redirecting their efforts. AI is primarily used today in supportive roles such as data preprocessing and workflow management, reflecting the sector's need for precision and reproducibility. Despite costs often exceeding expectations and monetary returns projected to be years away, users remain committed to AI as a long-term innovation tool, prioritizing research advancement over immediate financial gain.

Some highlights from the study:

- 75.8% of the sites felt that their AI projects met or exceeded expectations.
- 74.9% of respondents indicated plans to moderately or significantly expand generative AI to support HPC workloads, roughly 28% of those characterize this expansion as significant.
- Less than 3% expect to contract their use of AI, none of which would characterize that contraction as significant.
- Roughly 40% of respondents are already using agentic AI models.

However, technical challenges continue to bring hesitation when it comes to broad adoption. Hallucinations, lack of explainability, and integration complexity are persistent concerns, especially in scientific and engineering contexts where accuracy is paramount. These issues, combined with the high computing demands of generative models, have led to a projected slowdown in the high rate of growth. This signals a transition from reactive adoption to more measured, application-specific onboarding. To sustain momentum, future progress will depend on innovations that reduce technical friction and elevate AI from peripheral support to more central roles in HPC workflows.

KEY FINDINGS

The survey was conducted in July 2025 and collected input from 103 survey respondents who indicated current or planned use within the next 12-18 months of production or mission-related AI/HPC integrated processes. Respondents came from a mix of major sectors: commercial (59%), academic (20%), and government (21%), representing verticals led by computers and related electronics but also including the financial, bioscience, advanced manufacturing, and geosciences sectors.

Key Finding #1: Over half of respondents intend on continuing to expand AI budgets by over 20% within the next 12-18 months; over a quarter of respondents intend to increase budgets by over 50% in the same time period.

The large investments of the last 12-18 months show signs of continuing among most users

The large investments of the last 12-18 months show signs of continuing among most users. Most intend to stay on pace or increase AI budget commitments, while an outlying few anticipate decreasing. This data demonstrates a continued trust and reliance on the new integrated technology, and confidence in its ability to provide the desired benefits. This reliance will be reflected in a continued rise in the hardware, software, infrastructure, and expertise procurements to support AI efforts. Survey data reflects the substantial investment seen over the last 12-18 months in which over half of respondents reported over 20% increase in investment during the last 12-18 months and over a quarter of respondents noted an increased investment of over 50% in the same period.

Key Finding #2: No respondents indicated decreasing their AI budgets over the last 18 months; only one outlier reported intent to decrease over the next 18 months.

When asked about the past and future intent to invest in AI capabilities, users unanimously reported having expanded their investment over the last 18 months and all but one respondent indicated no intention to decrease their spending growth within the next 18 months.

- Despite this large increase, over 21% of respondents intend to maintain their current level of large language model (LLM) development or reduce the growth rate in their spending.

With priority being placed on exploring and understanding the potential benefits and in-house requirements of LLM integration, organizations are more swiftly understanding the scope of their own needs as they relate to the capabilities offered by current AI technology. As such, these organizations are more often beginning to slow investment expansion rates and transition to a fine-tuning stage of development as opposed to rapid growth. Those organizations indicating a slowdown in their AI investment growth rates are likely the least satisfied with performance or most disadvantageously positioned to continue expending resources to achieve certain goals.

Key Finding #3: Nearly 75% of respondents plan to moderately or significantly expand generative AI use to support HPC workloads.

The majority of respondents anticipate expanding Gen-AI development either moderately (46.6%) or significantly (28.2%) for their HPC scientific and engineering workloads. A clear demonstration of continued interest and reliance on this new technology, this is and will continue to be reflected in budgetary and resource investment. Note than over 20% of respondents indicated continuing growing at the same level or contracting their expansion of generative AI use.

This expansion comes along with allocation of advanced computing budgets that would traditionally be committed to non-AI procurement. 26% of respondents anticipate relegating 30% or more of advanced computing budget to HPC-based computational requirements to support Gen-AI capabilities over the next 5 years.

This survey data indicates that *generative AI comprises a considerable portion of the advanced computing market*.

Considering its relative novelty when compared to traditional systems, this reflects an explosive investment pattern among a usually slow-moving ecosystem of buyers.

The majority of respondents anticipate expanding Gen-AI development either moderately (46.6%) or significantly (28.2%) for their HPC scientific and engineering workloads.

Key Finding #4: Roughly half of respondents expect a measurable monetary return on investment within 2 years with 30% expecting it in less than one year.

Roughly 18% of respondents expect a measurable monetary return within 0-2 years with ~35% anticipating a return in 2-4 years. 21% of respondents do not know when a measurable monetary return could be expected.

Roughly 18% of respondents expect some measurable monetary return within 0-2 years with ~35% anticipating a return in 2-4 years.

However, approximately 22% of respondents indicated breakeven, negative, or no demonstrable return on investment over the next 3 years. This group was not limited to government or academic respondents. The plurality of respondents fell into the categories of a return on investment of 50%-99% (19.4%) and 25-49% (14.6%). For many, increased revenue was not high on their list of goals compared to providing new insights or faster time to solution.

Key Finding #5: Most users reported LLM integration into HPC/advanced computing workloads met or exceeded expectation.

44% of users reported the technology met their expectations over the last 12-18 months with over 31% reported that it exceeded expectations over the same period.

44% of users reported the technology met their expectations over the last 12-18 months.

With roughly 17% reporting failure to meet expectations, these sentiments appear to be linked to use case activities, with data analysis end use cases being most closely linked to satisfaction.

When asked about the breakdown of end uses for generative AI supported in their AI integrated HPC scientific and engineering workloads, the vast majority indicated the use of scientific data analysis applications. In line with the analytic and language generation strengths of current generative AI capabilities, data analysis and code/text generation were among the top 4 selected end uses. Key use cases included:

- Scientific Data Analysis: 80.6%
- Time Series Data Analysis: 61.2%
- Text Generation: 60.2%
- Code Generation: 56.3%

- Synthetic Data Generation: 54.4%
- Image Creation: 43.7%

Key Finding #6: Most users reported their AI integration exceeded cost expectations at least moderately, but they intended to continue investing in the technology.

Roughly 52% of respondents indicated that their integrated generative models exceeded their cost expectations with only 6% reporting a lower-than-expected cost. These outcomes did not appear linked to the source of software solutions, use case, or overall budgetary profile of the respondent organization. Despite this sentiment, most organizations reported the intention to continue growing their AI capabilities.

In contrast to previous studies, uncertain/high development costs were not among the highest concerns.

Key Finding #7: Respondents indicated integration complexity, technical issues, and high computational demands as their top challenges.

When asked about the major barriers to their adoption and effective use of AI technology, respondents indicated that complexity of integration into existing HPC workflows was their highest concern. Technical concerns such as explainability and hallucinations along with high computational demands were also high on the list. In contrast to previous studies, uncertain/high development costs were not among the highest concerns.

Key Finding #8: New insights/knowledge not possible on traditional systems and faster time to solution stand out as top Gen-AI goals for HPC users.

Innovations in scientific and engineering computing take center stage with new insights and knowledge not possible on traditional systems and faster time-to-solution for science and engineering problems stand out as frontliners among respondents when asked to describe their most important goals envisioned by integrated Gen-AI. Notably, increasing revenue was often considered a background goal and was selected only at a rate of 11.7%. Top goals included:

- Providing new insights/knowledge not possible on traditional systems: 47.6%
- Faster time to solution for key HPC-based scientific or engineering solutions: 43.7%
- Opening new lines of HPC-based scientific or engineering research: 24.3%
- Greater compute efficiency on existing HPC-based workloads: 22.3%

Key Finding #9: 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 generative LLMs for HPC, respondents most often indicated that they are currently engaged in exploration and experimental activities of generative LLM integration, 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 workloads.

Key Finding #10: Transformer Models and Generative Adversarial Networks are used by the majority of respondents.

Notably, Agentic AI, a relatively new model type to the landscape of AI models, was selected at a rate of 40.8%. Please note that respondents could select multiple options:

- Transformer Models: 67%
- Generative Adversarial Networks (GANs): 51.5%
- Reinforced learning for generative tasks: 48.5%
- Recurrent Neural Networks (RNNs): 44.7%
- Agentic AI: 40.8%

Key Finding #11: OpenAI and Google stand out as top commercial developers of Gen-AI used in the last 12-18 months with Anthropic at third; the open-source marketplace for models maintains a healthy diversity.

When questioned on the commercial developers of AI tools used to support their efforts, respondents reported OpenAI as a standout with a 75.7% use rate of their products. Google was a relatively close second with 65.1% and Anthropic was in third place at 34%. This represents a continued maturation of the software service industry, with OpenAI and Google continuing to emerge as the major players.

SUMMARY

The user sentiment, behaviors, and expectations reflected in this survey demonstrate a continued intent to fine-tune existing integrated generative AI capabilities for HPC/advanced computing workloads as well as continuing exploring potential benefits and requirements for expanded integration. Demonstrated monetary return on investment, for the most part, is anticipated to be years away. While cost concerns are not of primary importance to HPC users, as evidenced by both the prioritizing of their own challenges and the continued expansion of AI capabilities despite admitted excessive costs beyond expectations, technical issues like hallucinations and explainability as well as the complexity of integration remain top-of-mind among science and engineering users.

The precision, reproducibility, and high stakes of scientific and engineering HPC/advanced computing workloads currently relegates AI integration to mostly data pre/post processing, workflow management and other ameliorative, but not mission critical roles. While organizations are currently willing to onboard technology that exceeds their cost expectations and that will not yield monetary results for upwards of 5 years, continued innovation in the area will be required to soften the roadblocks of considerable computing requirements, technical concerns, and ease-of-integration for this growth to maintain its strength.

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 sentiment and intent regarding the use of generative AI models to support science and engineering workloads in an HPC/advanced computing environment. Spending levels and the ROI from AI were key focus areas for the study. Respondents represented organizations that currently used AI capabilities including but not limited to large language models (LLM) to support current or planned HPC-based or advanced computing workloads. Respondents represented organizations operating on-premises and/or cloud-based HPC both for research and production environments.

The organizations represented in this data carry out scientific, engineering, and finance applications and excludes AI used for social media and other non-technical computing areas.

Research Approach

This study is based on a survey of organizations currently involved in using generative LLMs for HPC/Technical Computing/AI workload 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 returns on investment and performance for generative AI models' integration into HPC and advanced computing workloads.
- Assessing the current sentiment, anticipated usefulness, challenges, and goals integrated generative AI models in HPC environments.
- Understanding the computing landscape and resource types used in AI integration and usage within organizations that commonly use HPC.
- Exploring the details of budgetary allocations both present and anticipated, as well as the elements that influence expected purchasing changes.

Notes:

- *A more detailed description of respondent demographics can be found in the Appendix.*
- *Some numbers in this document may not be exact due to rounding.*
- *All monetary values shown in US dollars unless specified otherwise*.

SURVEY RESULTS

Characterizing Goals, Expectations, and Activities

The survey results demonstrate a diverse range of current and planned activities to support HPC endeavors: many users reported continued engagement in exploration of AI 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 variety of HPC-centric AI activity that spans AI technology development, exploration, hardware and software options, and workload production activities. As the fusion of these technologies matures, respondents consistently report engaging in exploring potential performance enhancements and in-house requirements even while their integration efforts enter more advanced stages.

Table 1

Current Level of AI Inference Activity in Organization's Existing HPC/Technical Computing/AI Compute Workloads

Activity	% Selected
Exploring the range of potential performance enhancements of integrating generative AI into existing HPC-based scientific and engineering workloads	48.5%
Exploring in-house requirements for integrating generative AI into HPC-based scientific and engineering workloads	43.7%
Testing/assessing generative AI-integrated workload performance	42.7%
Porting generative AI capability into existing workloads	35.9%
Running production level generative AI-enabled workloads	33.0%
Reaching out to generative AI hardware and software suppliers for information	27.2%
Standing up fully funded generative AI research efforts	27.2%
Procuring access to necessary generative AI software	26.2%
Procuring access to necessary generative AI hardware	25.2%
Standing up limited generative AI-integrated pilot programs	21.4%
Passively monitoring generative AI technology developments	12.6%
Don't know/ Not sure	0.0%

Table 1

Current Level of AI Inference Activity in Organization's Existing HPC/Technical Computing/AI Compute Workloads

Activity	% Selected

Notes: N = 103. Respondents could select all options that apply.

Source: Hyperion Research, 2025

Table 2, below, breaks down the generative model types most often used in HPC/advanced computing AI integration efforts. Transformer Models (67.0%) and Generative Adversarial Networks (GANs) (51.5%) were selected in more than half of respondents' answers.

Many respondents reported leveraging multiple model types, and there was a general spread among these types. It is expected that agentic AI systems designed to autonomously plan, execute, and adapt multi-step tasks will experience increased activity in the coming months and years. Notably, roughly 40% of respondents are already using agentic AI models, a relative newcomer to the landscape of generative AI.

Roughly 40% of respondents are already using agentic AI models.

Table 2

Breakdown of Generative AI Model Types Used Among Respondents

Option	% Selected
Transformer Models	67.0%
Generative Adversarial Networks (GANs)	51.5%
Reinforced learning for generative tasks	48.5%
Recurrent Neural Networks (RNN)	44.7%
Agentic AI	40.8%
Flow-based models	32.0%
Diffusion Models	31.1%
Autoregressive Models	27.2%
Variational Autoencoders (VAEs)	23.3%

Table 2

Breakdown of Generative AI Model Types Used Among Respondents

Option	% Selected
Don't know/Not Sure	1.9%
Other (please specify)	1.0%

Notes: N = 103. Respondents could select multiple options.

Source: Hyperion Research, 2025

Table 3 reveals specific end use cases among HPC/advanced computing users for their generative AI models. Analysis of scientific data (80.6%) and time series data (61.2%) were the most representative, with text generation (60.2%) and code generation (56.3%) not far behind.

- These responses show that HPC/advanced computing scientific and engineering applications generative AI models are most commonly leveraged to support the front and back end of existing traditional HPC methods such as modeling and simulation today.
- Most respondents reported engaging in several activities linked to their generative AI models, demonstrating a high confidence and willingness to explore new options for their AI technology.

Table 3

Specific End Uses of Generative AI

End Use	% Selected
Scientific data analysis	80.6%
Time series data analysis	61.2%
Text generation	60.2%
Code generation	56.3%
Synthetic data generation	54.4%
Image creation	43.7%
Audio or music generation	14.6%
Other (please specify)	3.9%

Table 3

Specific End Uses of Generative AI

End Use	% Selected
Don't know/Not Sure	0.0%

Notes: N = 103. Respondents could select multiple options.

Source: Hyperion Research, 2025

Complexity with integrating gen-AI models into existing HPC-based scientific and engineering workloads is the standout challenge among users at 51.5% of users selecting this option, as shown in table 4. Integration complexity issues can be alleviated with AI expertise, vendor guidance, and the continued establishment of industry best practices, which all continue to grow more robust as the fusion of traditional HPC and generative AI matures.

- The second largest challenge at 31.1% concerns technical issues surrounding generative AI-models such as explainability or hallucinations.
- The third largest concern is high generative AI computational demands at 29.1% of sites.

Few respondents (6.8%) are concerned with lack of software support on-site, and even less so in the cloud (2.9%).

- This and other Hyperion Research studies suggest that HPC/advanced computing users engaging in science and engineering workloads face particular challenges with technical concerns such as explainability or hallucinations when compared to their enterprise counterparts.

Table 4

Key Challenges Faced in Integrated AI-Based Models into HPC/Advanced Computing Environment

Challenges	% Selected
Complexity with integrating generative AI-based models into existing HPC-based scientific and engineering workloads	51.5%
Concerns with technical issues surrounding generative AI-models such as explainability or hallucinations	31.1%
High generative AI computational demands	29.1%
Lack of generative data precision	23.3%
Lack of in-house generative AI expertise	22.3%

Table 4

Key Challenges Faced in Integrated AI-Based Models into HPC/Advanced Computing Environment

Challenges	% Selected
Technology is moving too fast for credible assessment of value	17.5%
Lack of required hardware onsite	16.5%
Lack of return on generative AI investment	16.5%
High/uncertain generative AI development costs	14.6%
Confusion/uncertainty with generative AI hardware vendor selection	11.7%
Confusion/uncertainty with generative AI software vendor selection	11.7%
High/uncertain generative AI operational costs	10.7%
Uncertainty of demonstrated computational performance improvements	9.7%
Lack of required hardware in the cloud	6.8%
Lack of required software onsite	6.8%
Long/uncertain generative AI implementation times	5.8%
Lack of required software in the cloud	2.9%
We have not faced any challenges	2.9%
Don't know/not sure	2.9%
Other (please specify)	1.9%

Notes: N = 103. Respondents could select multiple options.

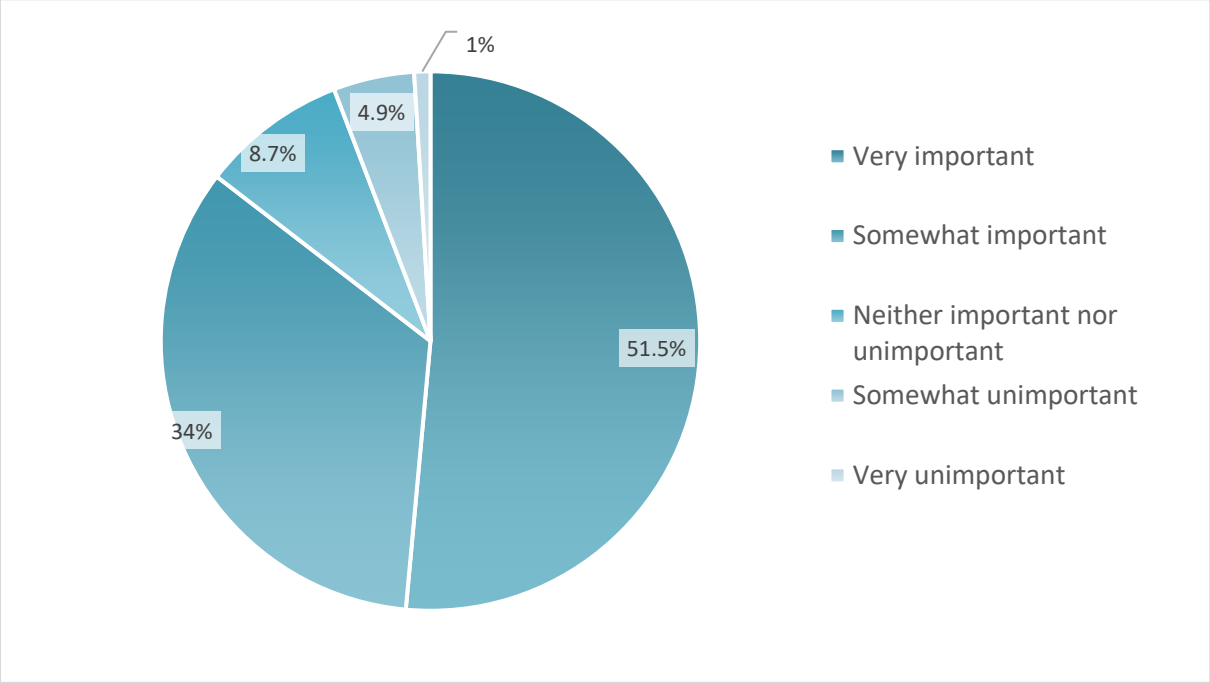
Source: Hyperion Research, 2025

Figure 1 characterizes the sentiments of HPC/advanced computing users regarding the importance of their AI integration into traditional workloads. With over 85% reporting that their generative AI models are either very important (51.5%) or somewhat important (34.0%) in the last 12-18 months, confidence in an already successfully integrated process is high.

- When considered with the continued involvement in exploratory efforts to bring more potential benefits and provision AI-focused computing resources, the importance of AI for HPC/advanced computing users is expected to continue its rise.

FIGURE 1

Importance of Generative Models to Support HPC-Based/Advanced Computing Workloads in the last 12-18 months



N = 103

Source: Hyperion Research, 2025

Table 5, below, arrays the most important goals envisioned among users in their integrated generative AI workloads. The two breakout responses, providing new insights or knowledge (47.6%) and faster time to solution (43.7%), are in line with previous Hyperion Research studies and demonstrate the focus on novel elements to benefit traditional workflows.

- With options like increased revenue (11.7%), lowering HPC hardware and software costs (8.7% and 2.9% respectively) and lowering power requirements (1.9%) on the low end of importance for users, the urgency for return on investment appears low.

Table 5

Most Important Goals Envisioned by Integrated Gen-AI Models Into Existing Workloads

Goal	% Selected
Providing new insights or knowledge not possible on traditional systems by themselves	47.6%

Table 5

Most Important Goals Envisioned by Integrated Gen-AI Models Into Existing Workloads

Goal	% Selected
Faster time to solution for key HPC-based scientific or engineering solutions	43.7%
Opening new lines of HPC-based scientific and engineering research	24.3%
Greater compute efficiency on existing HPC-based scientific and engineering workloads	22.3%
Greater computational fidelity on key HPC-based scientific and engineering workloads	18.4%
Increasing revenue	11.7%
Lowering overall HPC hardware costs	8.7%
Opening new lines of vertical-specific end uses	7.8%
Opening new lines of multiple vertical end uses	7.8%
Lowering overall HPC software costs	2.9%
Reducing HPC staffing requirements	2.9%
Lowering overall HPC power requirements	1.9%
Don't know/ Not sure	0.0%
Other (please specify)	0.0%

Notes: N = 103. Respondents could select multiple options.

Source: Hyperion Research, 2025

Table 6 gauges the sentiment among HPC/advanced computing users as to the level at which their integrated generative models met their expectations over the last 12-18 months. 75.8% felt that their AI projects met or exceeded their expectations.

75.8% felt that their AI projects met or exceeded their expectations.

With over 30% of HPC/advanced computing users indicating that their AI projects exceeded expectations and over 17% reporting failure to meet expectations, it is important to consider use cases, software selection and other idiosyncratic circumstances that may contribute to this discrepancy.

- Of the 18 respondents who reported their generative models somewhat or significantly failed to meet expectations, 78% of them were engaged in code generation activities.

- Among the respondents whose expectations were greatly exceeded, synthetic data generation and scientific data analysis were the most common generative model activities.

Table 6

Extent to Which Integrated Generative AI Models Met Performance Expectations Over the Last 12-18 Months

Expectations	% Selected
Greatly exceeded expectations	10.7%
Moderately exceeded expectations	21.4%
Met expectations	43.7%
Somewhat failed to meet expectations	15.5%
Significantly failed to meet expectations	1.9%
Have not yet integrated generative AI into our HPC workloads	4.9%
Don't know/ Not sure	1.9%

Notes: N = 103

Source: Hyperion Research, 2025

Table 7 measures user sentiment based on cost expectations and reality over the last 12-18 months.

- Over half (52.4%) of respondents indicated integration being more costly than expected.
- 34.0% felt that the costs met expectations.
- 6.8% found that costs were lower than expected.

With AI integration efforts demonstrating beneficial use cases in many HPC/advanced computing applications and AI integration being encouraged by many organization leaders as a matter of principle, higher-than-expected costs had no remarkable impact to their willingness to continue and ramp up future investment in the technology. However, users who indicated a failure to meet performance expectations were more likely to indicate a lower anticipated increase in budgetary allocation to AI integration.

Table 7

Extent to Which Integrated Gen AI Models Met Cost Expectations Over the Last 12-18 Months

Cost Expectation	% Selected
Significantly more cost than expected	11.7%
Moderately more cost than expected	40.8%
Met expectations	34.0%
Somewhat less costly than expected	4.9%
Significantly less costly than expected	1.9%
Have not yet integrated generative AI into our HPC workloads	3.9%
Don't know/ Not sure	2.9%

Notes: N = 103

Source: Hyperion Research, 2025

Budgeting, Provisioning, and Supporting Continued AI/HPC Integration

Table 8 contains a breakdown of budget allocations across a range of AI development and provisioning needs. Most users indicated engaging in all or most of these activities, with the most budget being allocated to development of in-house generative AI models at a rate of 21.1%. Most users are leveraging a hybrid compute environment for their activities, while 29.0% reported on-premises only operations.

- Previous Hyperion Research studies have indicated the popularity of using cloud resources as a platform for piloting projects, experimenting with new or unfamiliar workflows, or assessing the hardware needs of their AI activities.
- It is expected that on-premises solutions will rise as processes evolve from experimental to more full scale workloads.

Furthermore, while fine-tuning is an ongoing process that will continue at the current or increased rate, it is expected that budgetary allocations to meet inferencing needs will rise sharply, especially as HPC/advanced computing organizations hone their models and more coherently scope the needs and boundaries of their applications.

Table 8

Breakdown of Current Gen-AI Hardware Budget

Budget Allocation Area	% Selected
Development of in-house generative AI models	21.1%
Procurement of externally trained generative AI models	17.0%
On-premises generative AI model training efforts	17.0%
Cloud-based generative AI model training efforts	13.2%
On-premises generative AI fine tuning	9.3%
Cloud-based generative AI fine tuning	7.6%
On-premises inferencing operations	6.5%
Cloud-based inferencing operations	6.2%
No costs to date	1.6%
Don't know/ Not sure	8.3%

Notes: N = 103

Source: Hyperion Research, 2025

As Table 9 reveals, no respondent indicated a decrease in their AI spending over the last 12-18 months. Indeed, over the last 12 to 18 months:

- 58.2% of the sites raised their budgets by over 20%
- 47.5% increased their budgets by over 30%,
- 34.9% increased by over 50%
- 13.6% more than doubled their spending

The overall average growth rate for the sites in this study was 42.6%.

This outcome was not unexpected as HPC users and organizations have generally shown a high level of interest in the benefits of AI demonstrated across varied application spaces, especially at enterprise sites, over the past 3 years.

Table 9

Budget Changes to Support Gen-AI Models for Overall HPC/Advanced Computing Workloads Over Last 12-18 Months

Budget Change	% Selected
Increased over 100%	13.6%
Increased 75%-99%	1.9%
Increased 50%-74%	19.4%
Increased 30-49%	12.6%
Increased 20-29%	10.7%
Increased 15-19%	4.9%
Increased 10-14%	9.7%
Increased 5%-9%	7.8%
Increased 1-4%	3.9%
No change	11.7%
Decreased 1-9%	0.0%
Decreased 10-19%	0.0%
Decreased 20-29%	0.0%
Decreased 30-49%	0.0%
Decreased 50-74%	0.0%
Decreased 75%-99%	0.0%
Decreased 100%	0.0%
Don't know/ Not sure	3.9%

Notes: N = 103

Source: Hyperion Research, 2025

Table 10 contrasts with Table 9 as to the anticipated change over the next 12-18 months to support generative AI models for HPC/advanced computing workloads. Over the next 12 to 18 months 57.4% of the sites expect to increase their budgets by over 20%, 39.9% plan to grow by over 30%, 25.3% to

grow by over 50% and 9.7% plan to more than double their spending. Although slightly lower than the last 12 to 18 months, this represents continued major growth.

Only 18.4% of respondents indicated a higher rate of investment over the next 12-18 months compared to the previous time slot. However, the anticipated investment amount still represents a significant increase over the traditional pace of advanced computing investment, suggesting continued strong growth of markets supporting advanced AI technology.

- The overall average expected growth rate was 36.6%, compared to 42.6% for the previous 12 to 18 months. This represents a slightly lower but still high rate of growth.
- As noted above, performance expectations rather than cost expectations were a greater indicator of continued spending increases among respondents.

The overall average expected growth rate was 36.6%, compared to 42.6% for the previous 12 to 18 months. This represents a slightly lower but still high rate of growth.

Table 10

Anticipated Budget to Support Gen-AI Models for Overall HPC/Advanced Computing Workloads Over Next 12-18 Months

Percentage Change	% Selected
Increase by over 100%	9.7%
Increase 75%-99%	3.9%
Increase 50%-74%	11.7%
Increase 30-49%	14.6%
Increase 20-29%	17.5%
Increase 15-19%	5.8%
Increase 10-14%	9.7%
Increase 5%-9%	7.8%
Increase 1-4%	1.9%
No change	12.6%
Decrease 1-9%	0.0%
Decrease 10-19%	0.0%
Decrease 20-29%	0.0%
Decrease 30-49%	1.0%

Table 10

Anticipated Budget to Support Gen-AI Models for Overall HPC/Advanced Computing Workloads Over Next 12-18 Months

Percentage Change	% Selected
Decrease 50-74%	0.0%
Decrease 75%-99%	0.0%
Decrease 100%	0.0%
Don't know/ Not sure	3.9%

Notes: N = 103

Source: Hyperion Research, 2025

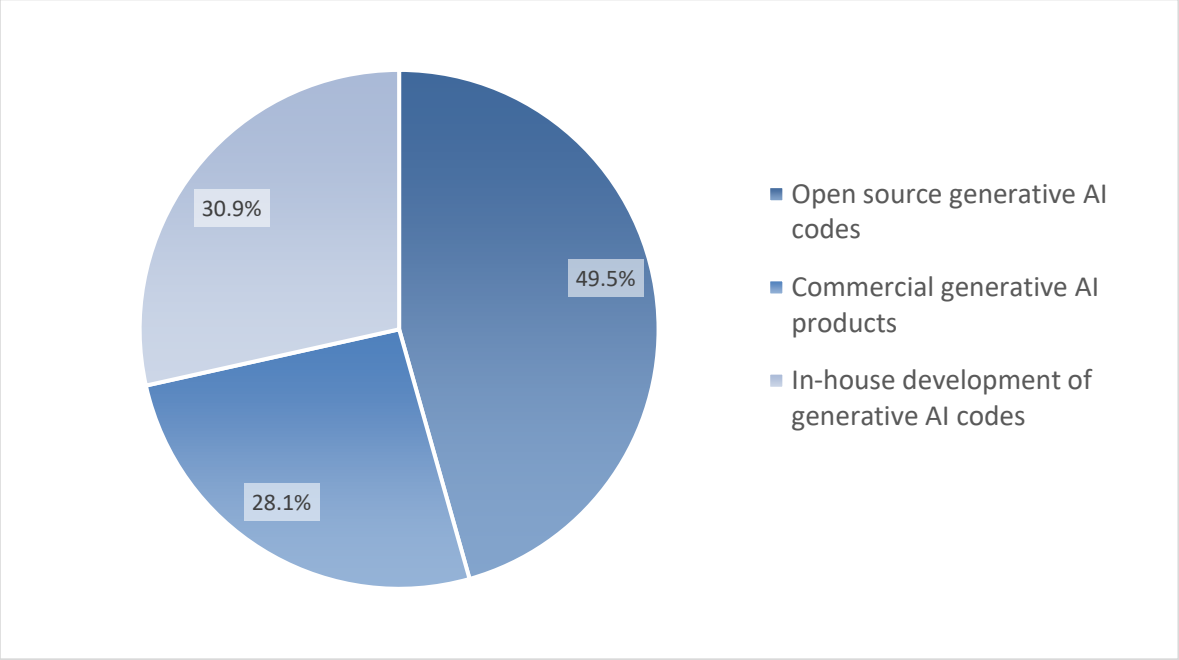
Current Tools and Resources Leveraged

Figure 2 displays a breakdown of generative AI-related software sources, with open-source representing the plurality of responses (49.5%). This is not unexpected, as the open-source software landscape to support generative AI models is well established as a robust and reliable source of production-ready solutions.

- Most respondents indicated that they are using more than one source for generative AI-related software.
- The source of software to support AI did not have a significant impact on the cost expectations realized among respondents, suggesting that the rationale of an integration effort in meeting, exceeding, or failing to meet cost expectations lies outside of software provisioning.

FIGURE 2

Breakdown of Sources for Gen-AI Related Software



N=103

Source: Hyperion Research, 2025

Table 11 is a breakdown of commercial developers whose models were used by respondents over the last 12-18 months. As with the previous survey data, most respondents indicated the use of multiple sources for their models.

- The top three developers/models were OpenAI, Google, and Anthropic.
- 10.7% of respondents indicated using only one source for generative AI models.
- Of those, only one respondent reported using a model from a source other than Anthropic, Google, or OpenAI.

Table 11

Commercial Developers of Gen-AI Models Used in Last 12-18 Months

Developer	% Selected
OpenAI	76.7%
Google	65.1%
Anthropic	34.0%

Table 11

Commercial Developers of Gen-AI Models Used in Last 12-18 Months

Developer	% Selected
Cohere	7.8%
Inflection AI	4.9%
AI21	1.9%
Don't know/ Not sure	7.8%
Other (please specify)	11.7%

Notes: N = 103. Respondents could select multiple options.

Source: Hyperion Research, 2025

Table 12 arrays the open-source providers of generative AI models and their frequency used in the last 12-18 months. While the breakouts of DeepSeek, Hugging Face, Meta Llama 3, TensorFlow, and PyTorch are familiar on the top, most respondents indicated using three or more sources. Moreover, the lack of significant outliers in this group further demonstrates the sufficiency and usability of open-source solutions in the generative AI for industry space.

Table 12

Open-Source Providers of Gen-AI Models Used in Last 12-18 Months

Provider	% Selected
PyTorch	67.0%
TensorFlow	60.2%
Meta Llama 3	47.6%
DeepSeek	42.7%
Hugging Face	35.9%
Scikit-leanr	29.1%
Keras	23.3%
OpenCV	20.4%

Table 12

Open-Source Providers of Gen-AI Models Used in Last 12-18 Months

Provider	% Selected
LangChain	19.4%
Mistral AI	16.5%
Grok.AI	15.5%
Cohere	11.7%
AlphaCode	8.7%
Falcon	6.8%
Stable Diffusion	6.8%
Bloom	5.8%
LeanML	5.8%
Don't know/ Not sure	5.8%
Other (please specify)	0.0%

Notes: N = 103. Respondents could select multiple options.

Source: Hyperion Research, 2025

Plans for Expansion/Contraction and ROI Outlook

Table 13 reports responses when survey participants were asked about their future plans to expand, contract, or maintain course on their AI investment strategies.

74.9% of respondents indicate plans to moderately or significantly expand generative AI use to support HPC/advanced computing workloads, with roughly 28% of those characterizing this expansion as significant.

Less than 3% expect to contract their support at all, none of which would characterize that contraction as significant.

74.9% of respondents indicate plans to moderately or significantly expand generative AI use to support HPC/advanced computing workloads.

Table 13

Plans to Move Forward or Expand Gen-AI Development

Plan	% Selected
Significantly expand generative AI use to support HPC/advanced computing workloads	28.2%
Moderately expand generative AI use to support HPC/advanced computing workloads	46.6%
Continue at the same level of generative AI use	18.5%
Moderately contract or decrease use of generative AI for HPC/advanced computing workloads	2.9%
Significantly contract or decrease use of generative AI for HPC/advanced computing workloads	0.0%
Entirely discontinue use of generative AI for HPC/advanced computing workloads	0.0%
Don't know/Not sure	3.9%

Notes: N = 103

Source: Hyperion Research, 2025

Table 14 arrays the timespan in which respondents anticipated a measurable monetary return on their generative AI integration efforts. Over 70% of respondents expect a return within the next 3 years, with the plurality (29.1%) in less than one year. A notable portion (7.8%) expect no return on investment.

This further demonstrates the confidence placed in AI, especially considering the high level of ongoing budget investment. Furthermore, the belief in the inherent value of generative AI integration is demonstrated with only 11.7% of the same respondents cited increasing revenue as a main goal of their AI/HPC activity.

Over 70% of respondents expect a return within the next 3 years.

Table 14

Anticipated Demonstrable Measurable Monetary Return On Investment by Year

Time	% Selected
Less than one year	29.1%
One year to less than two years	21.4%
Two years to less than three years	20.8%

Table 14

Anticipated Demonstrable Measurable Monetary Return On Investment by Year

Time	% Selected
Three years to less than four years	7.6%
Four years to less than five years	4.2%
More than five years	4.1%
Never	7.8%
Don't know/ Not sure	5.0%

Notes: N = 103

Source: Hyperion Research, 2025

Table 15 shows the expected amount of return on investment over the next 3 years, which was the longest offered window of monetary return on investment. Between 50%-99% ROI was the most common response selected by 19.4% of the sites surveyed. 45.7% of the sites expect to see an ROI of at least 25% within 3 years and 31.1% expect to see a return of over 50%.

45.7% of the sites expect to see an ROI of at least 25% within 3 years.

Considering the high level of investment and confidence, the spread of expectations is surprisingly wide, with a relatively large portion of respondents indicating either negative ROI or no demonstrable monetary return on investment (19.4% combined), which is common with R&D investments. This, combined with those who expect a breakeven, comprises one fifth of sites expecting no monetary return on their AI integration efforts.

- Users who anticipated negative ROI were nearly evenly split across commercial, government, and academia when adjusted for demographic representation. However, government respondents were disproportionately more likely to anticipate no demonstrated monetary return.
- This combined with the 11.7% of respondents who expect over 100% return on investment make for a reasonably scoped expectation among users.

Table 15

Anticipated Total Return on Investment Over 3 Years

Percent Return	% Selected
100%+ return on initial investment	11.7%
50-99% return on initial investment	19.4%
25-49% return on initial investment	14.6%
10-24% return on initial investment	18.5%
1-9% return on initial investment	5.8%
Breakeven return on investment	2.9%
Negative ROI – the investments are greater than the financial returns (we are doing it for other reasons than financial ROI)	9.7%
No demonstrated monetary return on investment	9.7%
Don't know/ Not sure	7.8%

Notes: N = 103

Source: Hyperion Research, 2025

The data contained in Table 16 helps color the organizational identity of the respondents. With outliers on the spectrum of under $100,000 and over $500M, a wide spread of organizational budgets was achieved. The budget allocations of an organization had no remarkable bearing on whether their AI technology met cost expectations. The plurality of respondents was in the range of $1M to $10M.

Table 16

Total Annual Budget to Support HPC-Based Computational Requirements

Total Budget	% Selected
Under $100,000	3.9%
$100,000 to less than $250,000	7.8%
$250,000 to less than $500,000	6.8%
$500,000 to less than $1 million	5.8%

Table 16

Total Annual Budget to Support HPC-Based Computational Requirements

Total Budget	% Selected
$1 million to less than $5 million	12.6%
$5 million to less than $10 million	17.5%
$10 million to less than $20 million	6.8%
$20 million to less than $50 million	4.9%
$50 million to less than $100 million	8.7%
$100 million to less than $500 million	10.7%
More than $500 million	3.9%
Don't know/ Not sure	10.7%

Notes: N = 103

Source: Hyperion Research, 2025

Most respondents, indicated by the data in Table 17, anticipate the budget percent of HPC-based computational requirements to support generative AI capabilities to stay below 30% over the next 5 years. Respondents were more likely to expect greater allocation based partially on their expectations being met or exceeded by their current activity.

Table 17

Anticipated Budget Percent of HPC-Based Computational Requirements to Support Gen-AI Capabilities Over 5 Years

Budget Percent	% Selected
1% to less than 5%	4.9%
5% to less than 10%	5.8%
10% to less than 15%	14.6%
15% to less than 20%	13.6%
20% to less than 25%	15.5%

Table 17

Anticipated Budget Percent of HPC-Based Computational Requirements to Support Gen-AI Capabilities Over 5 Years

Budget Percent	% Selected
25% to less than 30%	11.7%
30% to less than 40%	10.7%
40% to less than 50%	8.7%
50% to less than 75%	3.9%
More than 75%	2.9%
Don't know/ Not sure	7.8%

Notes: N = 103

Source: Hyperion Research, 2025

SUMMARY AND NEXT STEPS

This report offers a detailed look into how users in high-performance computing (HPC) and advanced scientific domains are approaching the integration of generative AI technologies. The survey responses reflect a shift in mindset: while increasing the expansion of AI capabilities, users are now focused on refining and optimizing existing implementations while more responsibly responding to successes by expanding, keeping course, or changing directions. This pivot suggests a maturation strategy, where organizations are moving beyond the initial excitement and toward more deliberate, scoped applications, particularly in areas like data preprocessing, workflow management, and post-processing. These roles remain supportive and usually not mission-critical, which aligns with the precision and reproducibility demands of scientific and engineering workloads.

Despite the high costs associated with generative AI, often exceeding expectations, users remain willing to increase investments, with the understanding that monetary returns may not materialize for three to five years. Interestingly, cost is not the dominant concern for most HPC users. Instead, they prioritize solving domain-specific challenges and advancing research capabilities. This tolerance for delayed ROI returns underscores the strategic importance of AI as a long-term innovation tool, rather than a short-term financial asset.

However, technical barriers continue to loom large. Concerns around hallucinations, explainability, and complex integration are top-of-mind for many users. These issues are particularly acute in scientific and engineering contexts, where accuracy and transparency are non-negotiable. The difficulty of embedding AI into existing HPC workflows without compromising performance or reliability remains a significant hurdle, and one that must be addressed for broader adoption to continue.

This tolerance for delayed ROI returns underscores the strategic importance of AI as a long-term innovation tool, rather than a short-term financial asset.

This report also highlights a notable trend: continued high growth in AI spending, although at a lower rate of growth compared to the previous 12 to 18 months. This suggests a transition from the rapid, possibly reactive adoption seen in previous years to a more measured, readiness-focused approach. Organizations are increasingly evaluating the suitability of LLMs for specific applications, and this scoping process is helping them develop a clearer, more grounded perspective on AI onboarding. 74.9% of respondents indicate plans to moderately or significantly expand generative AI use to support HPC/advanced computing workloads, with roughly 28% of those characterizing this expansion as significant. Less than 3% expect to contract their support at all, none of which would characterize that contraction as significant.

Looking ahead, the continued integration of generative AI into HPC environments will depend on overcoming several key challenges. Reducing inaccuracy and false results from AI models, improving performance, improving model reliability, and simplifying deployment pathways are all critical to sustaining momentum. While current use cases remain largely peripheral, ongoing innovation in model architecture and tooling will be essential to unlock deeper value and potentially elevate AI to more central roles within scientific and engineering workflows.

APPENDIX: SURVEY DEMOGRAPHICS: RESPONDENTS AND ORGANIZATIONS

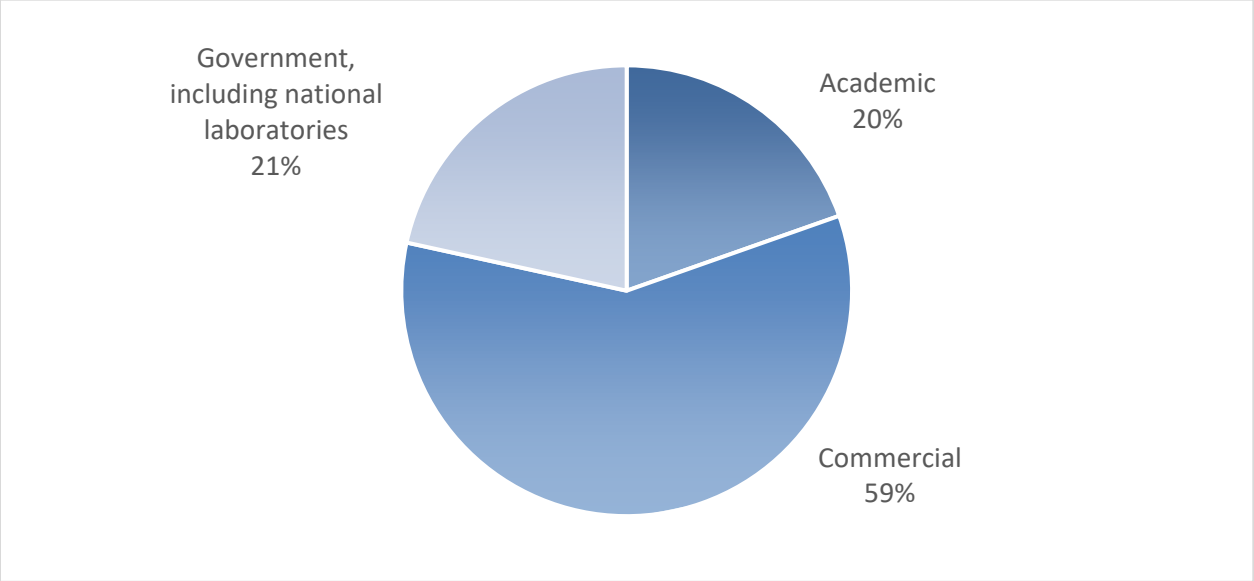
This appendix provides additional details on the demographics of the survey respondents and the organizations they represent. It consists of three main segments: individual survey respondents' background information, general demographics about the organization they represent, and select data on the budgetary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

Respondent Demographics

Figure 3 breaks down respondent sector affiliation. This study focused generally on commercial respondents with consideration also made for governmental and academic respondents. 59% reported representing commercial or industrial organizations, 21% from government including national laboratories, and 20% from academic groups.

FIGURE 3

Respondent's Sector Affiliation



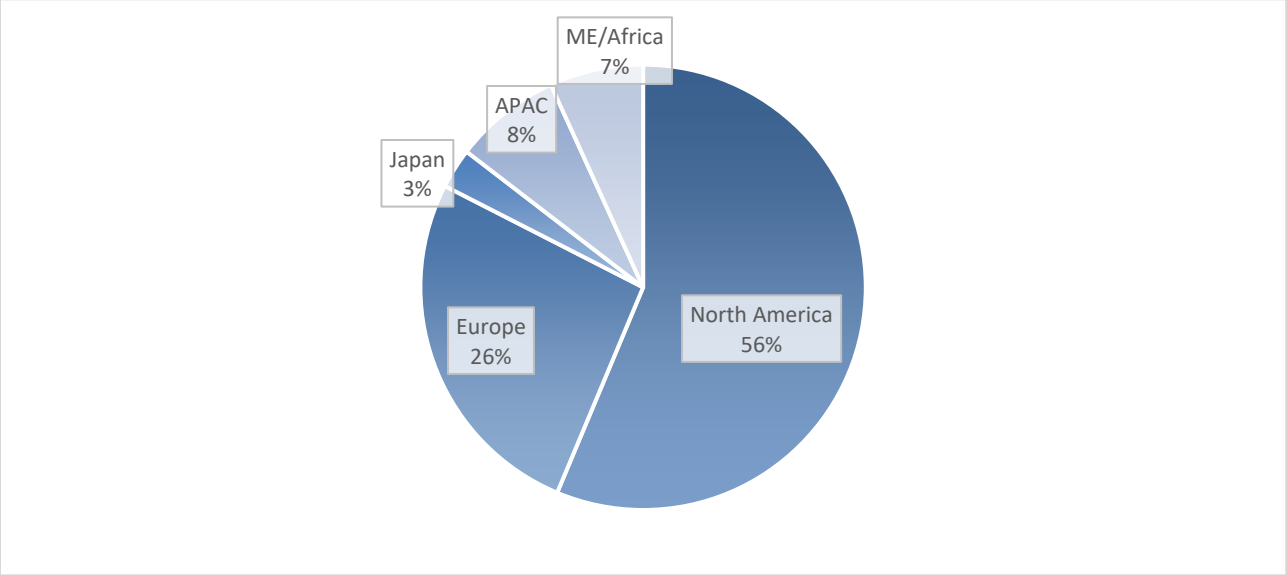
N=103

Source: Hyperion Research, 2025

Figure 4 outlines the respondents' headquarters location. Over half (56%) were from North America, 26% were from organizations headquartered in Europe, 3% represented Japan, and the remainder, approximately 8%, were from APAC. The Middle East, Africa, and the rest of the world contributed approximately 7% of responses.

FIGURE 4

Respondent's Organization Headquarters Location



N=103

Source: Hyperion Research, 2025

About Hyperion Research, LLC

Hyperion Research is the leading global provider of strategic planning information on high-performance computing and related areas for government, industry and academia. By evaluating advanced technologies, user requirements, applications, and adoption dynamics, Hyperion Research clarifies the behavior of the computationally and data-intensive technical computing market, including emerging areas such as cloud computing, AI (including machine and deep learning), big data, precision medicine, automated driving systems, smart cities, IoT, cyber security, and quantum computing. As Hyperion Research, the team continues all the worldwide activities that have made it the world's most respected HPC industry analyst group for more than 30 years, including HPC and AI market sizing and tracking, HPC subscription services, custom studies and papers, and operating the HPC User Forum. For more information, see http://hyperionresearch.com/

.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.hyperionres.com
