



HYPERION RESEARCH

Understanding the Evolving Use of AI in HPC

June 2025

www.HyperionResearch.com
www.hpcuserforum.com

Tom Sorensen

Maturing AI Use Raises New Questions

As the technology continues to be further integrated into HPC environments, challenges and opportunities expand

- **Continued integration progress of AI among HPC users prompting longer-term perspectives:**
 - How to efficiently procure resources
 - How extensively cloud resources should be used
 - Comprehensiveness of regulatory guidelines
- **Despite realized advantages, users are more realistically assessing challenges:**
 - High cost of upkeep including power & infrastructure
 - Continual education of in-house expertise
 - Management of shifting regulatory demands

Forecasting in a Shifting Environment

Hyperion Research AI forecasts are still being fine tuned

- **Forecasts for server and other hardware procurements is evolving due to major changes in the market**
 - Increased yet often exploratory use of cloud resources
 - Continued assessment of appropriate hardware/software for application
 - Hastened accelerator/GPU release cycles
 - Diversification of language models in domains

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

Source: Hyperion Research, 2025

Ongoing Hyperion Research Studies

Hyperion Research continues long-term series of HPC/AI studies

- **Continued series of studies tracking HPC/AI user behaviors, expectations, and challenges**
 - Began with LLM study, continued to integration, inference, and moving on to ROI in June 2025
 - Inference provisioning and management has become an area of heightened focus
 - As “hype” fades, ROI will receive greater attention
- **End User Inferencing: Completed Last Month**
 - Targeted towards the inferencing side of production and near-production integration of advanced AI/LLM
 - The survey dove into the hardware and software requirements of user groups and organizations managing high inferencing demands, as well as related budgetary and infrastructure requirements
 - Survey respondents provided insights on their specific inference types, level of integration and experimentation, and other details of their advanced AI usage including plans and methods of scaling

Inference Study Select Key Findings

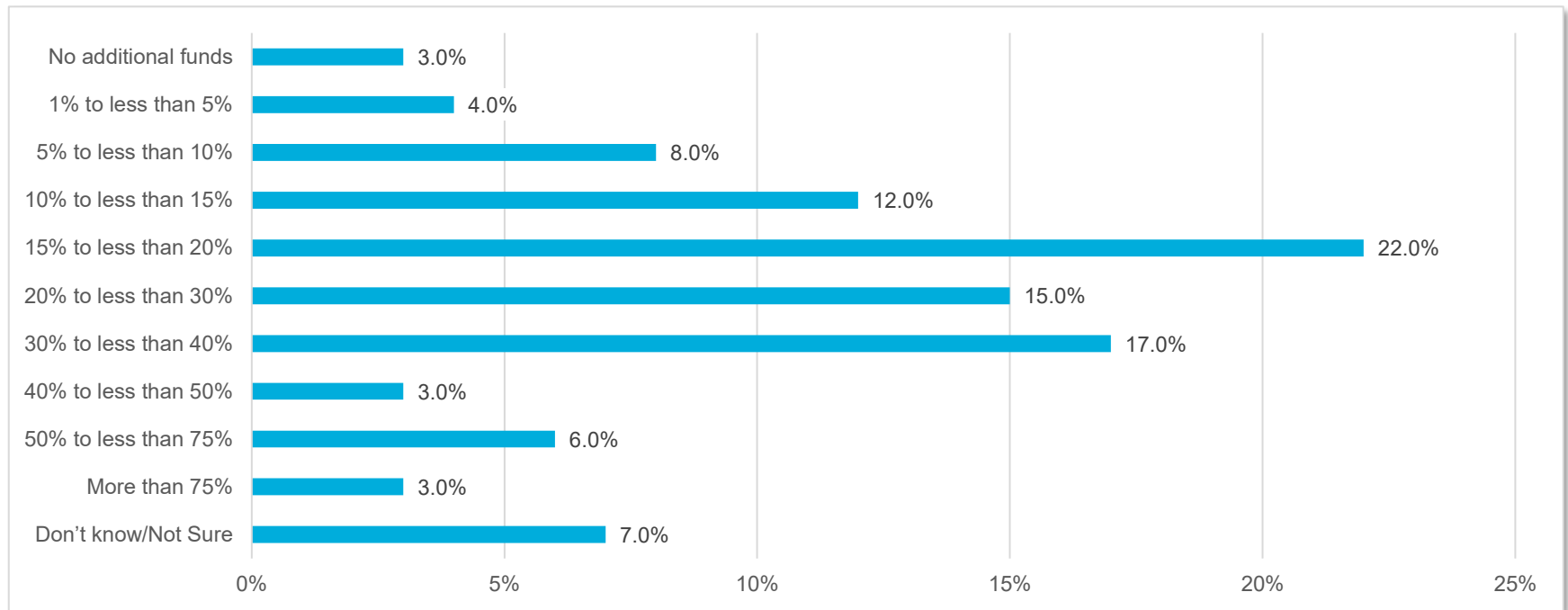
Inference is of high importance to HPC users, experimentation continues

- Users most frequently indicated that they are still exploring/experimenting with AI-centric options both in the cloud and on-premises
- Concerns centered on integration complexity, hardware/software cost, and technical issues
- HPC users report a nearly even split between on-premises and cloud budgeting
- A plurality of the software resources being used to support AI inferencing is open source

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%

5-Year Anticipated % of Overall Advanced Computing Budget for AI Inferencing

Confidence in considerable efficiency and productivity gains remains very high



N=100

Source: Hyperion Research, 2025

- **The inference spending portion is expected to rise, with some outliers expecting a decrease**
- **Additional outliers expecting to reach the >75% threshold**

ROI Study Highlighted Survey Questions

Next AI study focuses on return on investment, management of challenges, and shifting allocation of resources

- **To what extent did integrating generative AI models into your HPC workload environment meet performance and cost expectations?**
- **How have budgetary plans to support gen-AI change over the last 12-18 months?**
- **If there have been measurable monetary gains from HPC/AI integration, how long will it be to recover from initial investment?**
- **To what degree will your organization expand or contract gen-AI development moving forward?**

Top of Mind: AI Maturity Brings New Questions

As efforts to adopt and integrate AI gain traction among industry leaders, new use cases, optimization, regulatory developments, and ROI will become a new focus for users

- **HPC/AI integrators have come to expect:**
 - Robust return on investment
 - New levels of efficiency
 - Effective regulatory guidelines
- **As AI integrated systems become the norm, the effectiveness and limitations of the technology will become better understood**
- **Aspirant goals will be realized for many users, but some may face costly challenges of unexpected severity such as:**
 - High cost of upkeep
 - Continual education of in-house expertise
 - Rising emphasis on effective oversight
 - Management of regulatory demands

Top of Mind: LLM Training Needs a Reboot

The rapid rise of compute requirements for large language model training runs will begin to slow with a shift in emphasis on smaller and more efficient models using more focused training data sets

- **Current LLM training requirements 10^{26} total training operations**
 - Projections call for an increase of two to three order of magnitude in the next few years (10^{28} to 10^{29})
 - This is out of reach for all but the most aggressive, well-funded organizations: e.g., Anthropic, OpenAI, Tesla, Meta, Google
- **The mainstream HPC world will instead focus on less demanding LLMs or small language model training**
 - Requires less total compute, perhaps three to four orders of magnitude less
 - Based on training data sets that are smaller, more disciplined or subject focused, appropriately curated, and perhaps even proprietary to a targeted end use or end users

Questions, Comments And Suggestions Are Welcome



Please contact me at:
tsorensen@hyperionres.com