Special Analysis

Assessment of Cloud-based AI Activity and Anticipated HPC User Benefits

Tom Sorensen and Bob Sorensen
August 2024

CURRENT STATUS

In a soon to be available Hyperion Research study focusing on HPC users leveraging cloud resources for AI tools and capabilities, respondents most often indicated that they are currently engaged in exploratory and experimental stages of AI adoption, even those who have already integrated AI capabilities and technologies into some portion of their production environments. In this study, respondents demonstrated a preference for cloud over on-premises options for a range of AI-centric activities. Figure 1 compares anticipated cloud with on-premise AI workload allocation for the next 12–18 months.

FIGURE 1

AI Approaches Between Cloud and On-premises Solutions for Next 12-18 Months

Exploring the range of potential on-premises AI performance enhancements	50%
Exploring the range of potential cloud-based AI performance enhancements	64%
Reaching out to AI hardware and software suppliers for information	30%
Reaching out to cloud service providers for hardware and software information	35%
On-premises hardware procurement for AI activities	25%
Cloud-based hardware procurement for AI activities	30%
On-premises software procurement for AI activities	20%
Cloud-based software procurement for AI activities	34%
Standing up limited on-premises AI-integrated pilot programs	22%
Standing up limited cloud-based AI-integrated pilot programs	31%
Testing/assessing on-premises AI-integrated workload performance	25%
Testing/assessing cloud-based AI-integrated workload performance	39%
Running production level AI-enabled workloads on-premises	30%
Running production level AI-enabled workloads in the cloud	43%

Note: n = 103, Respondents could select multiple options

Source: Hyperion Research, 2024

Table 1 summarizes the anticipated benefits of integrated AI capabilities supported by cloud computing. Attracting, developing, and/or maintaining a more effective workforce was the most often selected at 58%, with automated coding and application modernization narrowly behind at 55%. General automation was selected at only a rate of 25%, next to workforce scheduling enhancement at 26%. Using expert systems to develop new insights in R&D was selected by only 1% of respondents.

Table 1

Main Benefits of Cloud-Based HPC AI Capabilities

	Currently
Attract, develop, and/or maintain a more effective workforce	58%
Automate coding or applications modernization	55%
Boost existing compute intensive application performance	34%
Create new content with generative AI	31%
Drive business process optimization	30%
Drive customer-related recommendation strategies	27%
Enhance computer-vision capabilities	26%
Enhance workforce scheduling optimization.	26%
Improve robotics/automation capabilities	25%
Improve customer service capabilities	22%
Personalize customer experiences	17%
Use expert systems to develop new insights in R&D	1%

N = 105

Respondents could select multiple options.

Source: Hyperion Research, 2024

ANALYST COMMENT

The study, which also explores budgetary considerations, algorithm types leveraged, and other details of HPC users leveraging cloud resources for AI, suggests that there are a number of benefits of using cloud resources for fast moving, new technologies such as advanced AI. While many accept the powerful enhancement capabilities of AI for HPC, users and organizations are still seeking the best, most efficient way to integrate AI into their work. Cloud resources offer a low-committal, high-diversity environment in which to explore, experiment, and pilot new methods of compute efficiency, and have an agile relationship with advancements in AI software and hardware that is not currently met in all on-

premises offerings. With procurement of advanced AI hardware for on-premises being protracted, and organizations still discovering and settling on exactly how to most efficiently and effectively leverage these tools, the cloud is seen as invaluable test bed for novel techniques and developments.

It is expected that as AI integration in HPC matures, on-premises resources will be more frequently leveraged. But as advanced AI hardware and software are currently progressing swiftly, cloud service providers are the quickest to offer the newest, most diverse architectures and configurations, despite the potential cost-savings of well-profiled and more application-specific on-premises devices.