Hyperion Research Special Report Overview
Cloud-based AI Activity for HPC: Widespread but Primarily Exploratory

October 2024

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

Highlights from the study include:

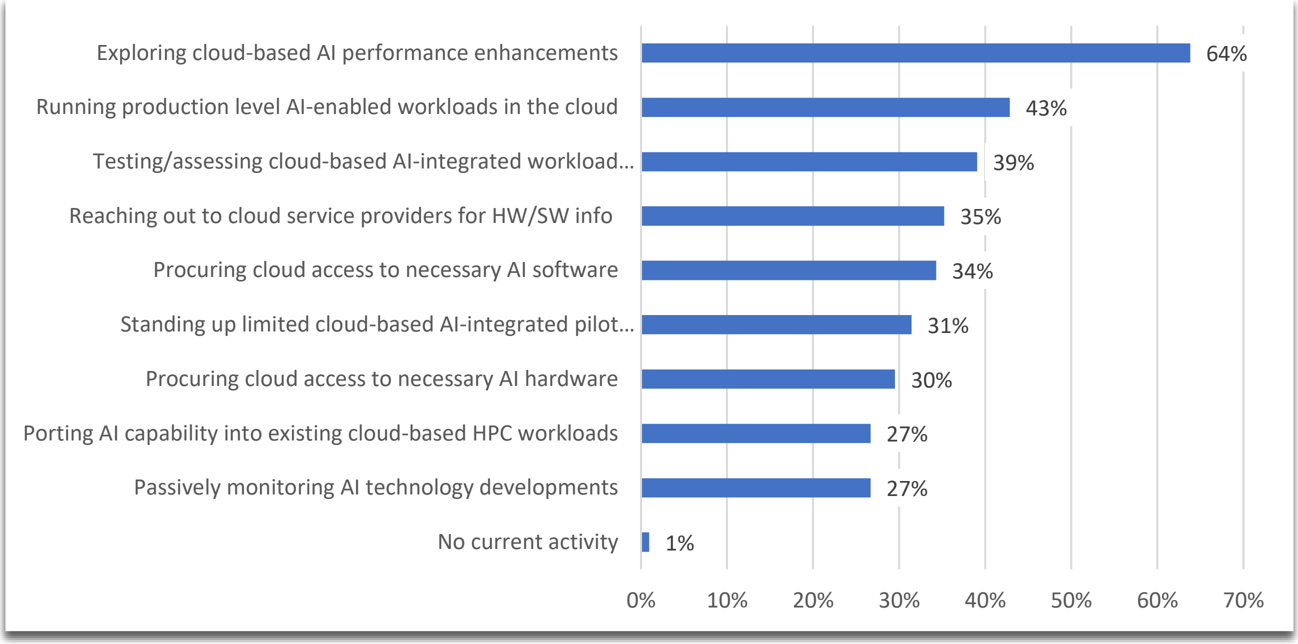
- Public cloud resources are considered a valuable asset in exploration and integration of AI into HPC or compute-intensive environments.
- Respondent organizations are leveraging a wide range of public cloud offerings.
- Numerous architectures and device types are currently used to meet inferencing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

Note: The survey, conducted in July 2024, collected input from 105 survey respondents who indicated current or planned use within the next 12-18 months of AI on public cloud-based resources to support HPC or compute-intensive activities. Respondents came from a mix of major sectors: commercial (76%), academic (2%), and government (14%), representing verticals led by computers and related electronics but also including the financial sector, bioscience, advanced manufacturing, and geosciences.

“Advanced computing users and organizations conducting compute-intensive workloads are currently and increasingly leveraging public cloud resources for AI endeavors,” said Tom Sorensen, lead analyst on the report. “With the speed of growing interest, adoption, and development within the advanced AI market, the cloud has become an agile, reactive environment that users can turn to for the latest offerings and support. While hardware development is already in a stage of increased pace that has not been seen before, CSPs have the advantage of circumventing long on-premises buying cycles, more streamlined installation into existing workloads, and a more composable way of designing compute infrastructure compared with on-premises counterparts. This agility within the cloud allows for CSPs to offer users more varied and up-to-date solutions while on-premises resources must follow a different, often lengthier path to utilization.”

As seen in the figure below, survey respondent organizations are in various stages of integration lifecycle, commonly running production-level workloads while simultaneously testing for improvements, exploring new options, and provisioning new or more application appropriate resources.

Current HPC-related AI Activities in the Cloud



Source: Hyperion Research, September 2024

For more information about the AI in the Cloud Special Study and purchase options, please contact jeansorensen@hyperionres.com.

For more about Hyperion Research's new AI Beacon CIS program and the AI Advisory Committee, go to https://hyperionresearch.com/ai-beacon/.