HPC-AI User Forum

HPC-AI Cloud Musings and Highlights from the 88th HPC User Forum

Mark Nossokoff and Earl Joseph
April 2025

OVERVIEW

The 88th HPC-AI User Forum (UF) was recently held from April 8-9, 2025, in Sante Fe, New Mexico. With a mission of promoting the health of the global HPC industry and address issues of common concern to users, the UF brings together global HPC-AI leaders from government, academia, and industry to share best practices and lessons learned from across the entire HPC-AI ecosystem.

HPC-AI in the cloud, inclusive of hyperscalers, full-service cloud service providers (CSP), and GPU/AI as a service (aaS), was one ecosystem area discussed at the 88th UF. Various perspectives were shared from the following individuals and organizations:

- Mark Nossokoff (Research Director, Hyperion Research): A Random Walk in the Clouds
- Sam Skillman (Staff Software Engineer, Google Cloud Platform): Progress in Developing HPC/AI Infrastructure
- Chris Maestas (CTO, Data and AI Storage Solutions, IBM): IBM Storage Scale for HPC and AI Solutions in the Cloud

A RANDOM WALK IN THE CLOUDS

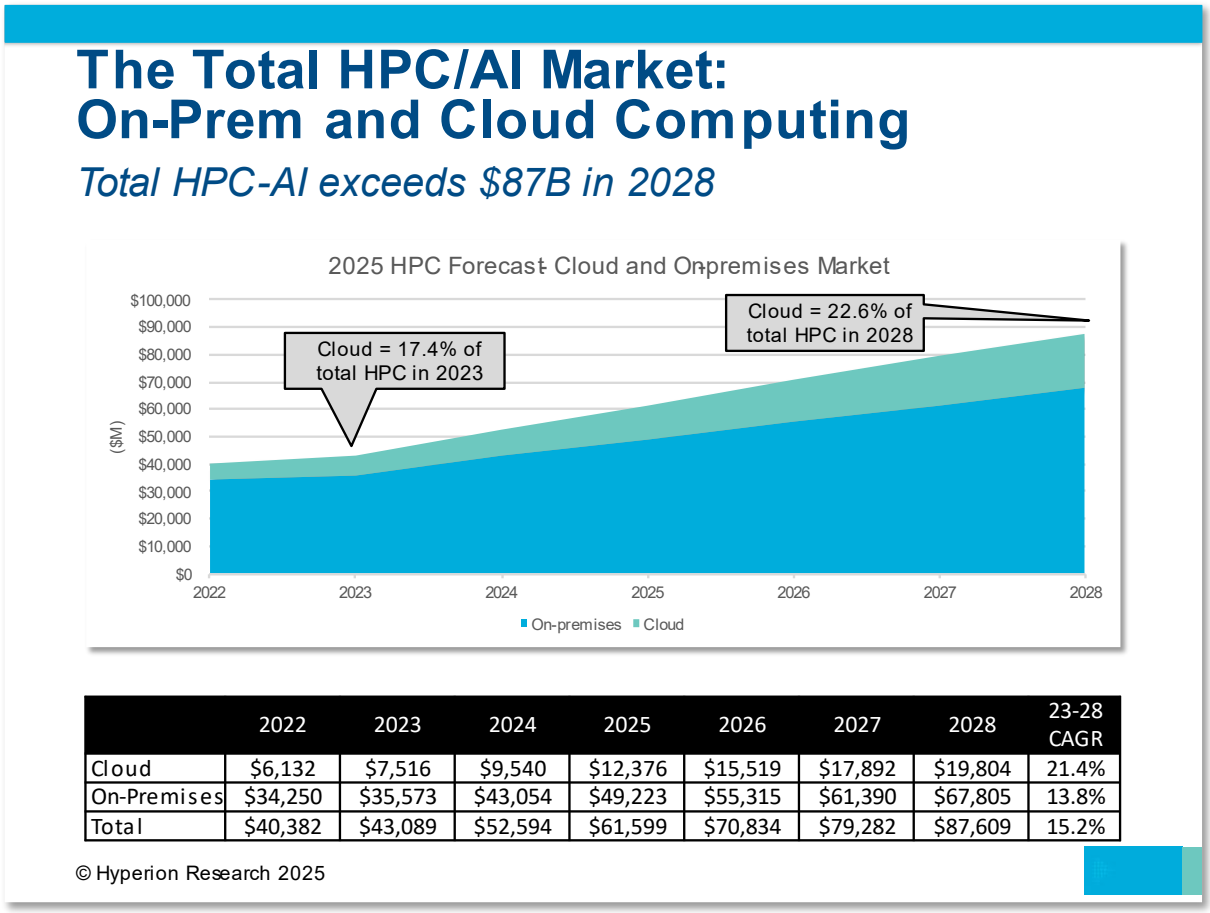
Adoption and Growth Continues for HPC-AI Cloud Resource Spending

User adoption of cloud-based resources to run their HPC-AI workloads continues to grow. No longer is the conversation "on-premises or cloud"; it has evolved into the notion of "continuum computing" where users are more intentionally evaluating the most appropriate infrastructure environment based on their respective priorities of budget, schedule, resources, and other items. Hyperion Research's most recent global site survey shows 77% of participants either currently use or intend to use cloud-based resources in the next 12-18 months.

User spending in 2023 on HPC-AI cloud resources was $7.5B, representing 17.4% of overall global spending on HPC-AI resources. Spending is forecast to grow at a 21.4% 5-year CAGR to $19.9B in 2028, representing 22.6% of the global HPC-AI market. Figure 1 shows Global 2023-2028 HPC-AI Forecast.

FIGURE 1

2024-2028 Global HPC-AI Forecast



	2022	2023	2024	2025	2026	2027	2028	23-28 CAGR
Cloud	$6,132	$7,516	$9,540	$12,376	$15,519	$17,892	$19,804	21.4%
On-Premises	$34,250	$35,573	$43,054	$49,223	$55,315	$61,390	$67,805	13.8%
Total	$40,382	$43,089	$52,594	$61,599	$70,834	$79,282	$87,609	15.2%

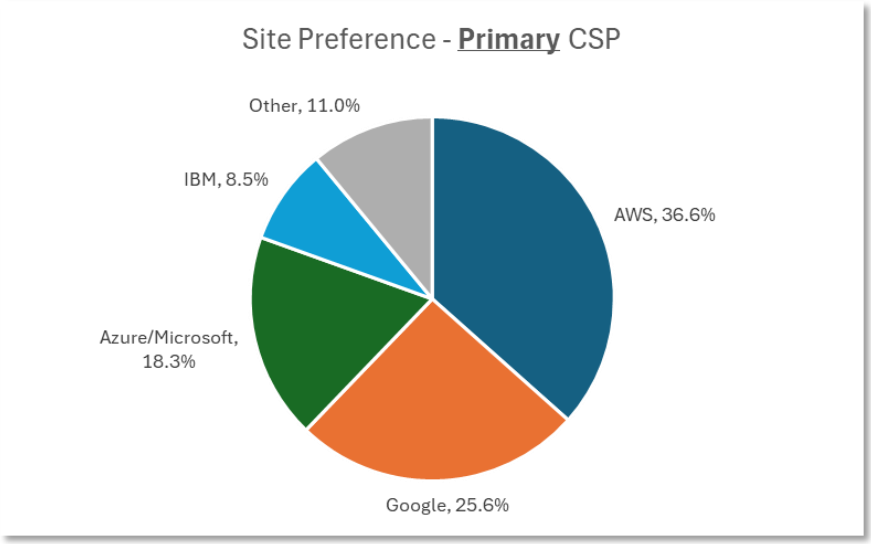
Source: Hyperion Research, 2025

AWS Identified as the Preferred HPC-AI Cloud Service Provider

Hyperion Research's most recent global site survey identified AWS as the preferred primary HPC-AI cloud service provider, identified by 36.6% of the respondents who indicated they utilized cloud resources for their HPC-AI workloads. Google Cloud was second, followed by Microsoft Azure, identified by 25.6% and 18.3% of the respondents, respectively. Figure 2 summarizes the preferred primary cloud resources provides among the 72 respondents.

FIGURE 2

Preferred Primary HPC-AI Cloud Service Providers



Source: Hyperion Research, 2025

Hyperscalers, Cloud Service Providers, and GPU/AIaaS Providers – An Emerging Taxonomy

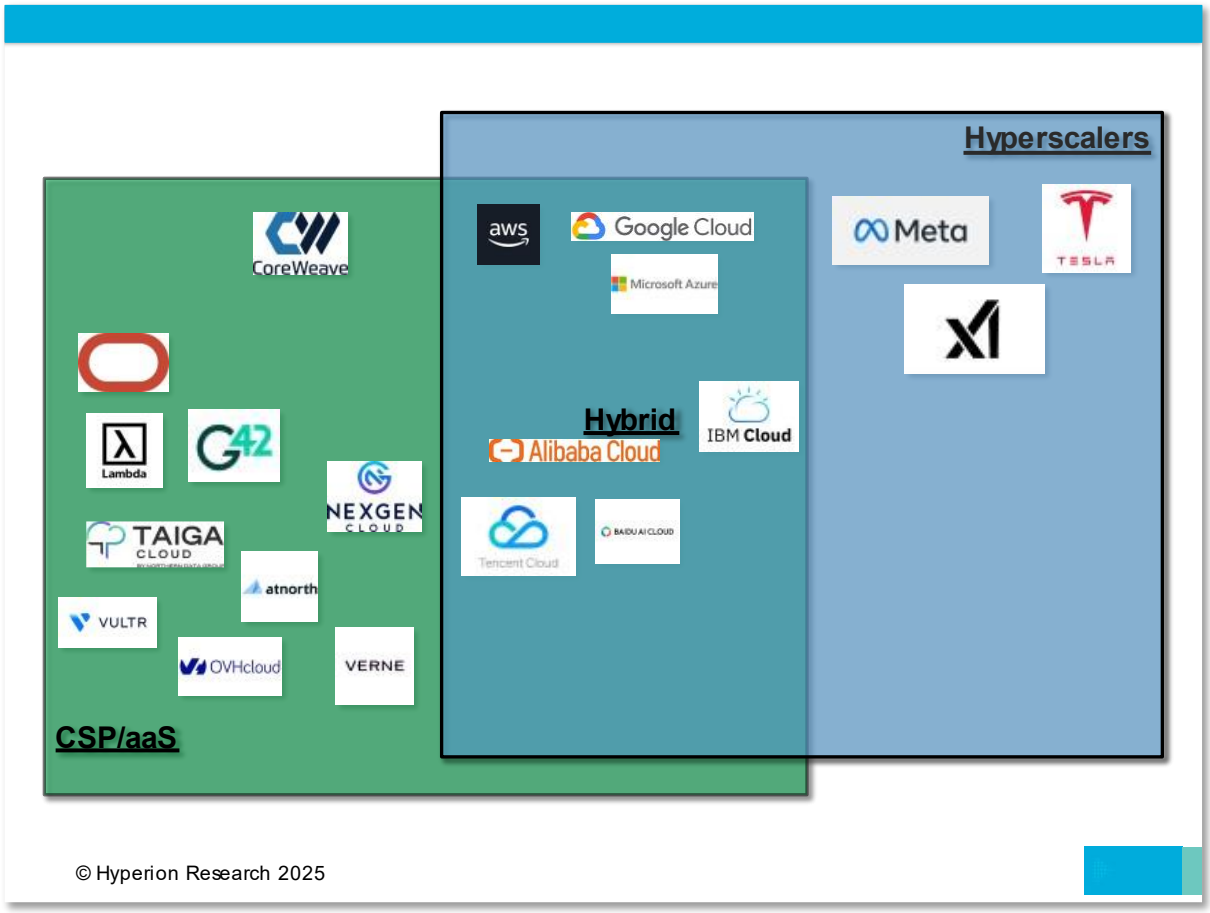
While similar in many ways, hyperscalers cloud service providers (including full service providers and GPU/AIaaS providers) and hybrid service providers also bear significant differences. First, some definitions:

- Hyperscalers: These organizations consume and deploy HPC-AI resources at very large scale. They both consume merchant technology as well as develop their own silicon. They do not provision instances of their technology for public consumption
- CSPs: CSPs are full service providers of compute and storage instances and broader HPC-AI services and support. They deploy HPC-AI resources at a moderate or a very large scale. They consume merchant technology and provision instances as a resource for external, third-party consumption. They may also develop their own CPU and/or accelerator technology for internal consumption. If the latter, they are considered "hybrid" in this taxonomy.
- GPU/AIaaS: These CSPs deploy merchant silicon at moderate to large scales but focus on accelerator and AI-focused infrastructure with limited additional services or support.

Figure 3 provides a mapping of this taxonomy and identifies example providers in each category. Table 1 summarizes the characteristics of each category.

FIGURE 3

Hyperscaler, Cloud Service Provider, and GPU/AIaaS Provider Taxonomy



Source: Hyperion Research, 2025

TABLE 1

Hyperscaler, Cloud Service Provider, and GPU/AIaaS Taxonomy Characteristics

Focus	Characteristic	CSP	GPU/AIaaS	Hybrid	Hyperscaler
External	Provisions instances for external consumption	X	X	X	
Technology & service provider	Reduced service offerings (e.g., AI-focused)		X		

TABLE 1

Hyperscaler, Cloud Service Provider, and GPU/AIaaS Taxonomy Characteristics

Focus	Characteristic	CSP	GPU/AIaaS	Hybrid	Hyperscaler
	Full array of services and support	X		X	
Internal	Consumes latest technology at scale	X	X	X	X
Technology developer and consumer	Develops custom silicon			X	X
	Utilizes resources for internal consumption			X	X

Source: Hyperion Research, 2025

Emergence of Continuum Computing

Users are becoming much more sophisticated in determining where to best run their workloads across both on-premises and cloud-based resources. This sophistication is evolving into a balanced on-premises/cloud partnership, with on-premises emerging as more performance-driven and cloud becoming resource driven:

- On-premises – performance driven
  - Multiple (integrated) partitions that meet key mission-critical workload specifics with targeted composition(s) of processors, accelerators, memory, etc.
  - Overall architecture deeply committed to current and planned workload requirements
  - Low latency access with direct connection to (primarily) local storage
  - Direct and long-term resource management/provisioning/staffing
  - Stable budget/resource schedule
  - Stable if not diverse software stack
- Cloud – resource driven
  - Wide range of instance options to address range of (perhaps changing) workloads
  - Quick access to new hardware, software for exploration
  - Cloud-based data supports collaboration, reduces data stovepipes
  - Elastic compute access: surge/step function/programmatic-specific
  - Flexible pricing options
  - Virtual/container-supported software environment

PROGRESS IN DEVELOPING HPC/AI INFRASTRUCTURE AT GOOGLE CLOUD

Identified as the secondary preferred provider in Hyperion Research's most recent global site survey, Google Cloud continues to advance it HPC-AI cloud resource services. Notable HPC-AI innovations aimed to advance scientific discovery highlighted and shared during the UF include announcements made at the concurrently running Google Cloud Next 2025 in Las Vegas:

- Multiple generations of Google's Tensor Processing Unit (TPU) roadmap
- New H4D CPU VMs
- GPU accelerated computing evolution
- Cluster Toolkit and Cluster Director
- Google Cloud Managed Lustre

Evolution of TPUs

Google's TPUs are on the 7th generation with their introduction of Ironwood. Designed to be Google's most performant, scalable, and energy-efficient TPU, Ironwood is intended to satisfy the demanding requirements of users' AI inferencing workloads. The core of Google's AI Hypercomputer, the liquid-cooled Ironwood is architected to scale up to 9,216 chips, leading to a fully configured AI Hypercomputer power consumption approaching 10MW.

H4D CPU VMs

Optimized for scientific HPC applications, the H4D VMs provide configuration options allowing users flexibility relative to core counts, memory size, and local SSD capacity and performance.

Accelerated Computing

Based on NVIDIA's latest lineup of accelerators, the new family of accelerated computing offerings are aimed at large-scale AI workloads and scientific computing applications.

Cluster Toolkit and Cluster Director

Cluster Toolkit provides users complete blueprints for specific scientific workloads to allow fully tested environments for ready-to-use applications. Toolkit items included for tested blueprints include software and applications, scheduler, storage type, machine type, and operating system.

Cluster Director supports the creation and management of purpose-built supercomputer clusters aimed to optimize performance for tightly coupled workload and inter-process messaging, maximize uptime, and provide resiliency at scale.

Google Cloud Managed Lustre

Providing high throughput, low latency, and high capacity storage, Google Cloud Managed Lustre aims to deliver a fully managed parallel file system for high performance AI training and checkpointing – two key items with the AI data pipeline.

IBM STORAGE SCALE FOR HPC AND AI SOLUTIONS IN THE CLOUD

Storage and data platforms are increasingly being recognized as critical elements with HPC and AI architectures. IBM shared investments they've been making within their storage portfolio deployed in their own IBM HPC Cloud and procured by third party cloud providers and leadership on-premises HPC-AI sites.

Initially developed and brought to market as GPFS, the parallel file system has been re-branded several times and is now offered as IBM Storage Scale. IBM has been advancing Storage Scale to address the diverse I/O requirements across the various phases of the AI data pipeline. Taken to market as part of the IBM Storage Scale 6000 system, Storage Scale has been deployed in several notable cloud-based and on-premises environments, including:

- IBM HPC Cloud
- Vela, IBM's cloud-based AI supercomputer
- CoreWeave's storage infrastructure for elements of its GPU/AIaaS offerings
- Scratch storage for JUPITER, Europe's first exascale supercomputer

FUTURE RESEARCH

User's understanding of continuum computing to easily and properly balance competing requirements (e.g., cost, performance, energy consumption, talent, and resources) is nascent, at best. Users across all HPC-AI sectors (industry, government, academia) are looking for guidance for how to best evaluate total cost of ownership (TCO) and develop an informed opinion of how to optimally leverage both on-premises and cloud resources. Hyperion Research is engaged in several current studies to address these questions:

- Value of Open Science Research Computing in the Cloud
- Establishing a Framework for Continuum Computing in Advancing Science
- Creating a Value Model for the Strategic Use of Continuum Computing
- Developing a Strategy for Enabling the Transition to Continuum Computing

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 and vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). 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.hpcuserforum.com and www.HyperionResearch.com
