



HYPERION RESEARCH

Perspectives on HPC-AI in the Cloud

ISC25 Market Update Briefing
June 2025

www.HyperionResearch.com
www.hpcuserforum.com

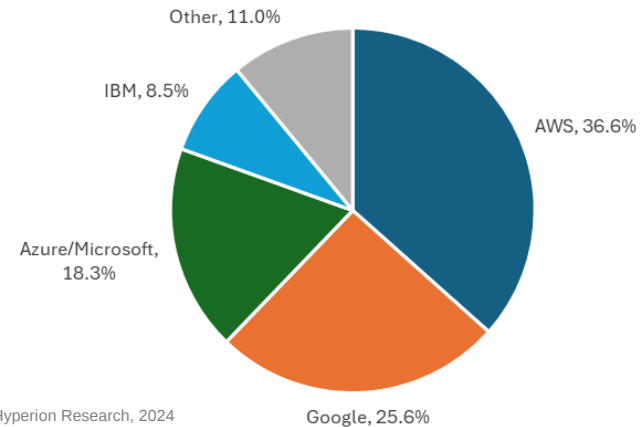
Mark Nossokoff

CSP Preferences – Primary vs. All

Who is your PRIMARY cloud provider / ALL cloud providers for your HPC/AI/HPDA workloads TODAY?

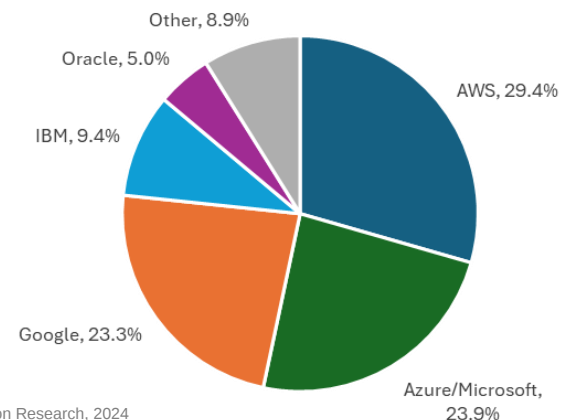
- **AWS the preferred primary CSP among respondents**
- **Google the 2nd most preferred primary CSP**
- **Microsoft the 3rd most preferred primary CSP, but rises to 2nd when considering all CSPs**
 - 180 total responses for CSPs utilized
 - ~2 CSPs per site

Site Preference - **Primary** CSP



N=82
Source: Hyperion Research, 2024

Site Preference - **All** CSPs, Including Primary



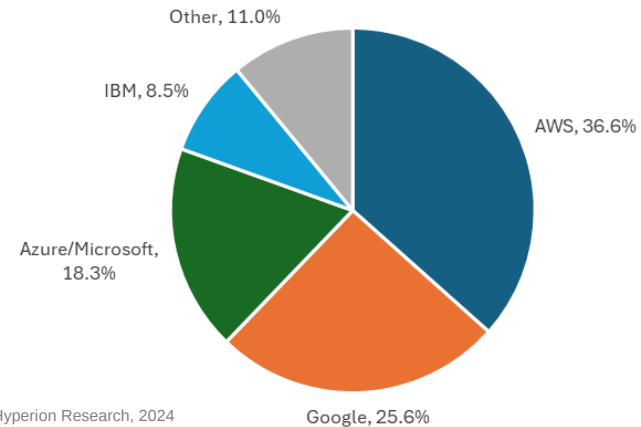
N=82
Source: Hyperion Research, 2024

CSP Preferences – AI Workload Crosscut

Who is your PRIMARY cloud provider / ALL cloud providers for your HPC/AI/HPDA workloads TODAY?

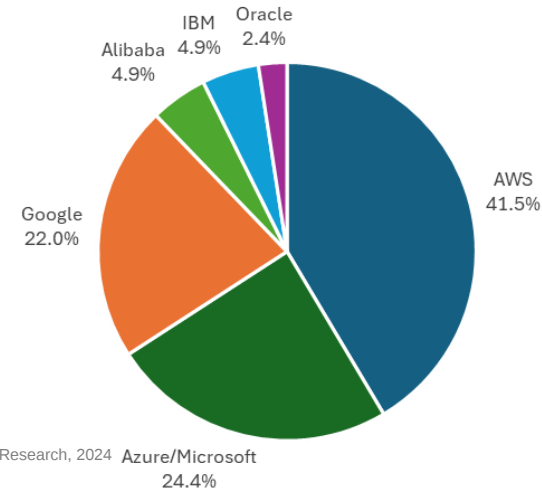
- **AWS the preferred primary CSP among respondents**
- **AWS as the primary CSP preference increases for sites who run >50% of their AI workloads in the cloud**
- **Microsoft moves to 2nd preferred primary preference for sites who run >50% of their AI workloads in the cloud**

Site Preference - **Primary** CSP



N=84
Source: Hyperion Research, 2024

Primary CSP - > 50% AI in the Cloud

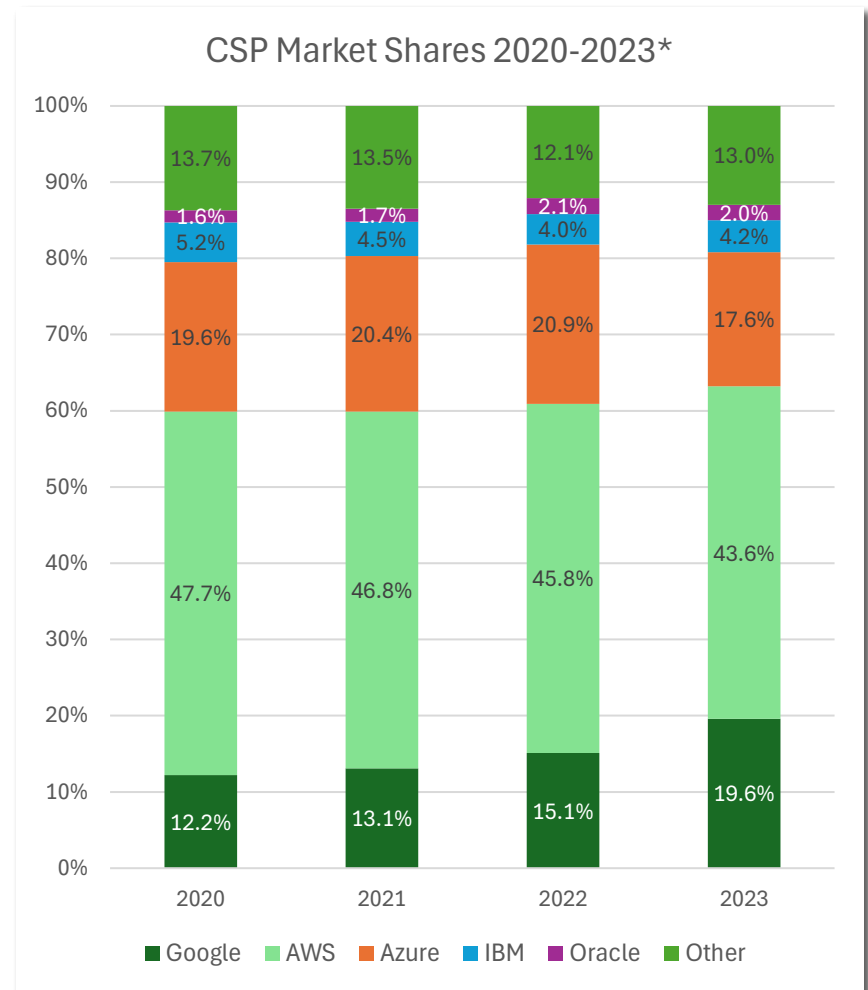


N=41
Source: Hyperion Research, 2024

Estimated CSP HPC-AI Market Shares

AWS maintains highest share

- **Google is gaining share**
- **“Other” is also gaining share**
 - European clouds
 - China clouds
 - Neo-clouds (AlaaS, GPUaaS)



*2024 year-end results not available at the time of this recording

Source: Hyperion Research, 2025

What's Happened in HPC-AI in the Cloud Since SC24?

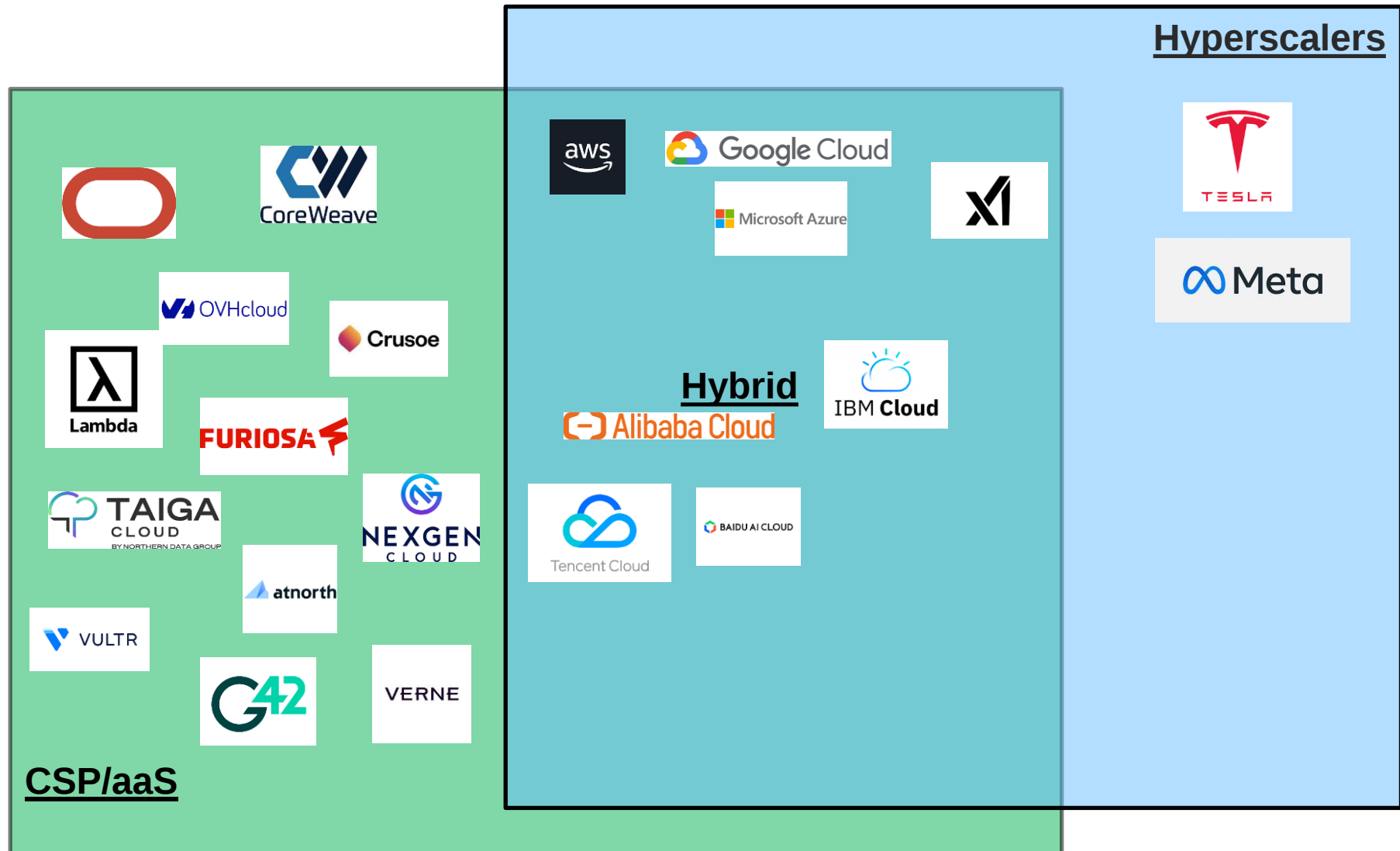
- **Google Cloud Platform**
 - New TPU7
 - New H4D CPU VMs
 - NVIDIA Blackwell support
 - Cluster Toolkit and Cluster Director
 - Google Cloud Managed Lustre
 - Agent Engine in Agent Space
- **AWS**
 - Trainium (GA for T2; preview announcement for T3; EC2 instances)
 - EC2 P6-B200 NVIDIA Blackwell instances
 - FSx for Lustre support for Elastic Fabric Adapter (EFA) and NVIDIA GPUDirect Storage (GDS)
- **Microsoft Azure**
 - NVIDIA Blackwell support
 - Azure HPv5 VMs
 - New in-house custom silicon beyond Maia and Cobalt (Hardware Security Module [HSM], Boost DPU)
- **Coreweave IPO**
- **UK Met shifts operations to Azure**

The Neo-Cloud Rises

Multiple factors will accelerate users to use CSP resources, including AlaaS and GPUaaS providers, to meet their compute needs

- **Acceleration of Cloud Adoption for AI Workloads**
 - AlaaS and GPUaaS providers ("neo-clouds") offer instant access to state-of-the-art hardware
 - Supply chain delays and frequent hardware refresh cycles drive demand for cloud-based solutions
- **Faster Access to Cutting-Edge Technology**
 - Expensive GPUs with yearly iterations encourage low-commitment cloud adoption
 - Rapid compute access accelerates AI/ML/DL integration/time-to-market
 - Supply chain uncertainty hinders smaller on-premises build-outs
- **Diversification of Application-Specific Hardware**
 - CSPs appeal to organizations in pilot, testing, and pre-production phases
 - Specialized AI data centers focus on refined service models over traditional CSPs (e.g., AWS, Google, Microsoft)
- **Sustainability as a Catalyst for Change**
 - Organizations avoid costly upgrades (e.g., liquid cooling) while reducing their carbon footprint
 - CSPs innovate energy management practices, promoting renewable energy and green architectures

Hyperscaler/CSP/aaS – Taxonomy



Hyperscaler/CSP/aaS Taxonomy

Focus	Characteristic	CSP/aaS	Hybrid	Hyperscaler
External Technology & service provider	Provisions instances for external consumption	X	X	
	Concentrated service offerings (e.g., AI-focused)	X		
	Full array of services and support		X	
Internal Technology consumer	Consumes latest technology at scale	X	X	X
	Develops custom silicon		X	X
	Utilizes infrastructure resources for internal consumption; does not provision instances based on custom silicon		X	X

Upcoming Studies

Several cloud-based studies in process

- **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**

Questions?



mnossokoff@hyperionres.com