



HYPERION RESEARCH

A Random Walk in the Clouds

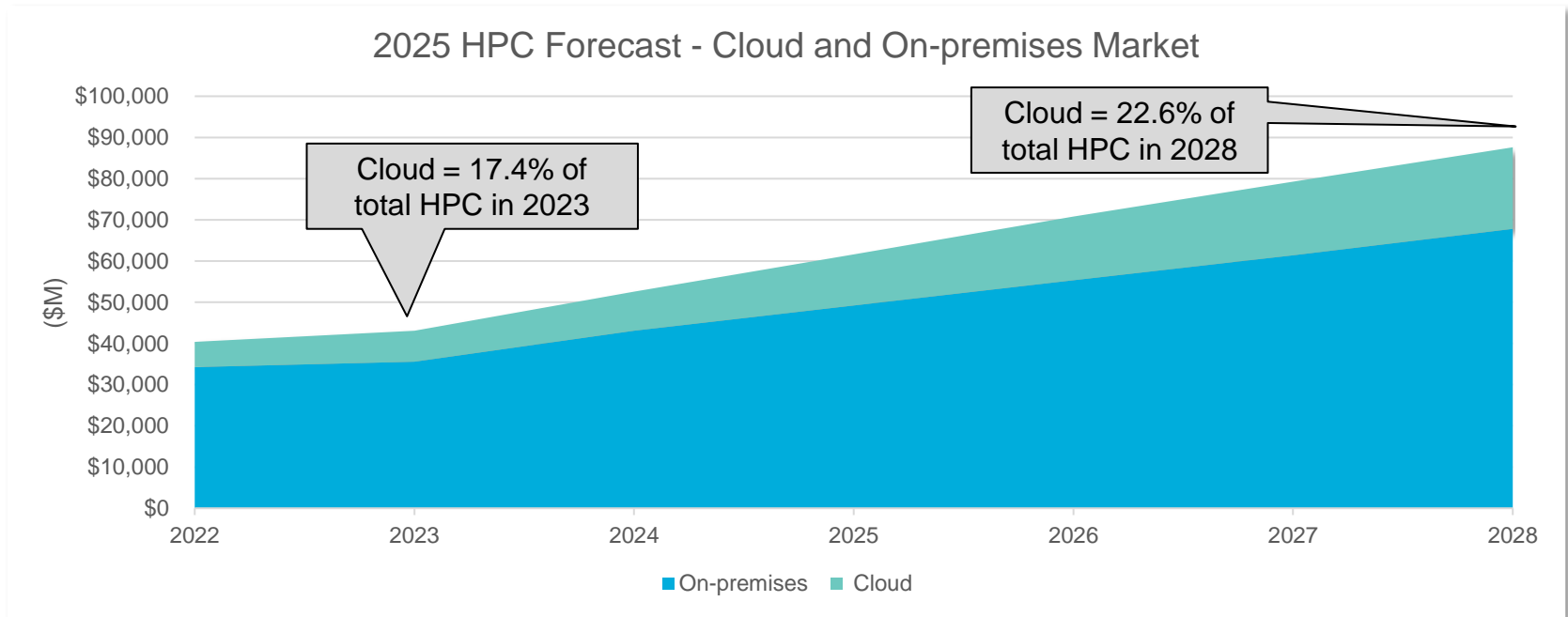
HPC User Forum
Santa Fe, NM
April 2025

www.HyperionResearch.com
www.hpcuserforum.com

Mark Nossokoff

The Total HPC/AI Market: On-Prem and Cloud Computing

Total HPC-AI exceeds \$87B in 2028

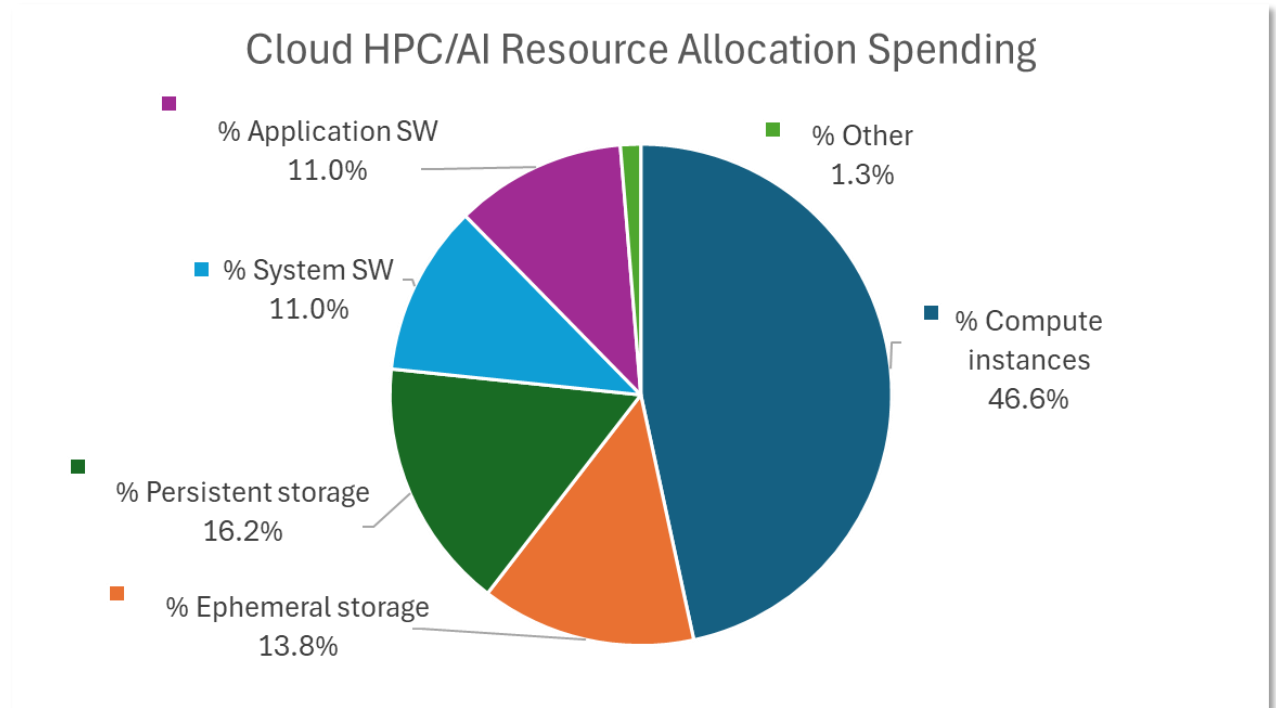


	2022	2023	2024	2025	2026	2027	2028	23-28 CAGR
Cloud	\$6,132	\$7,516	\$9,540	\$12,376	\$15,519	\$17,892	\$19,804	21.4%
On-Premises	\$34,250	\$35,573	\$43,054	\$49,223	\$55,315	\$61,390	\$67,805	13.8%
Total	\$40,382	\$43,089	\$52,594	\$61,599	\$70,834	\$79,282	\$87,609	15.2%

HPC-AI Cloud Resource Allocation Spending

11 sites spent 80% or more of their cloud budget on compute

- End user spending in the cloud for HPC-AI resources
- EXCLUDES what resource providers are spending to provision HPC-AI resources in the cloud

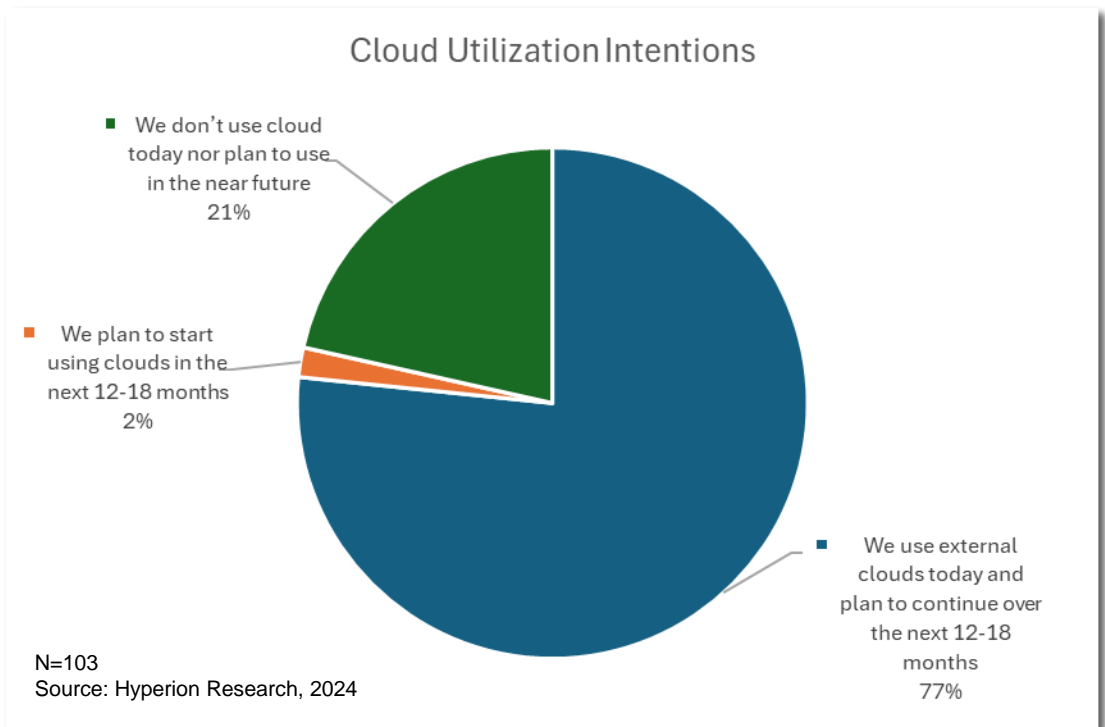


N=78
Source: Hyperion Research, 2024

Cloud Utilization for HPC-AI Workloads - Intentions

Are you using or planning to use external cloud resources for any of your HPC, AI, big data, or quantum workloads?

- **77% of respondents currently use or intend to use cloud within the next 12-18 months**
- **21% don't use the cloud or intend to use the cloud within the next 12-18 months**

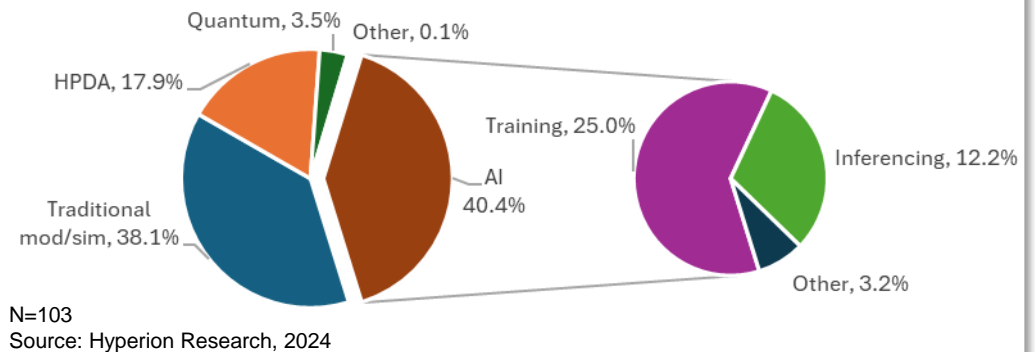


HPC-AI Workload Distribution by Environment - % Runtime

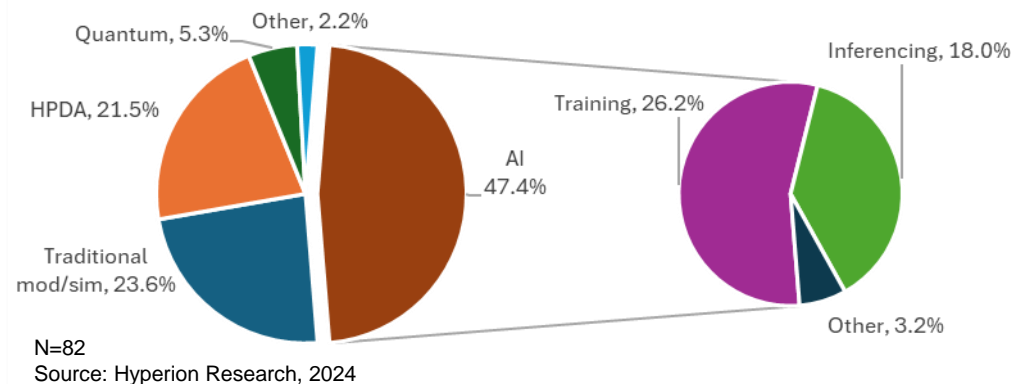
Of all your workloads in your HPC/AI/HPDA on-premises/cloud environments, please distribute your utilization time by the following:

- **AI identified as the primary workload based on runtime**
- **AI approaching 50% of the workload runtime in the cloud**
- **Traditional mod/sim runtime is 61% greater on-prem than in the cloud**

HPC-AI On-premises Workload Distribution - % Runtime



HPC-AI Cloud Workload Distribution - % Runtime

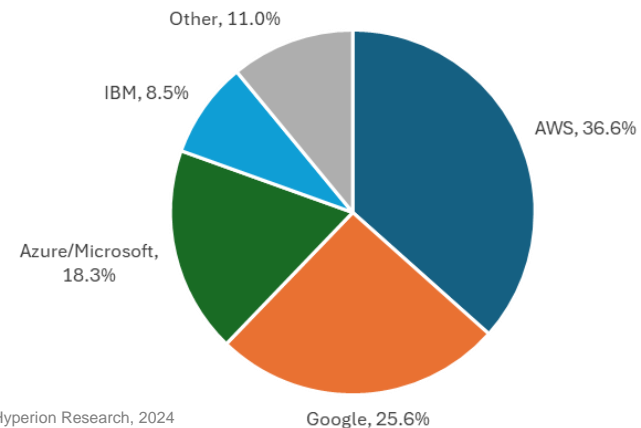


CSP Preferences – Primary vs. All

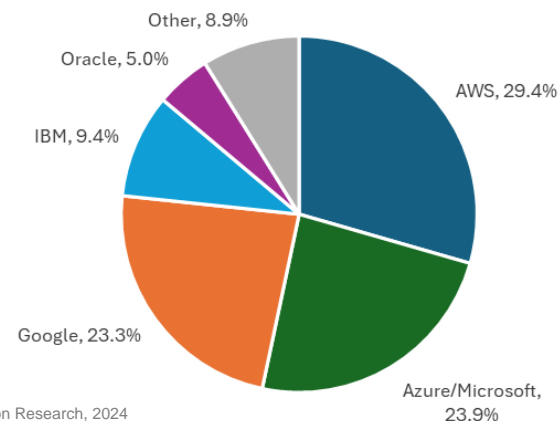
Who is your *PRIMARY* cloud provider / *ALL* cloud providers for your HPC/AI/HPDA workloads TODAY?

- **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



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

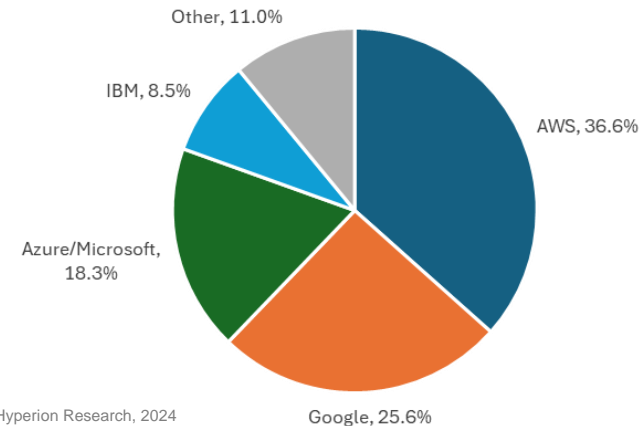


CSP Preferences – AI Workload Crosscut

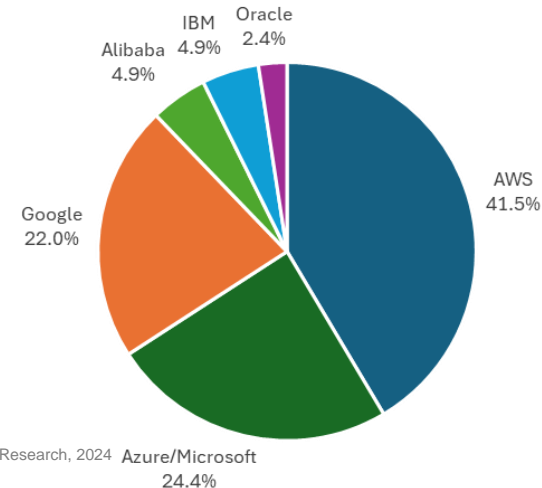
Who is your PRIMARY cloud provider / ALL cloud providers for your HPC/AI/HPDA workloads TODAY?

- **AWS the preferred primary CSP among respondents**
- **AWS as the preferred primary CSP 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



Primary CSP - > 50% AI in the Cloud



Mastering the Cloud-On-Prem Continuum

Users will more fully embrace the idea of “continuum computing”, incorporating the cloud as a viable tool in conjunction with (or instead of) their on-premises infrastructure

- **Optimized Resource Allocation**
 - Align infrastructure with workload-specific demands
 - Enable cost-effective and outcome-driven computing strategies
- **Enhanced Efficiency and Agility**
 - Dynamically shift resources between cloud and on-premises
 - User ability to respond rapidly to changing business needs and priorities
- **The ability to add or access new technologies more quickly**
- **Advancing Orchestration Tools**
 - New tools to simplify transitions across hybrid environments
 - Ensure interoperability and minimizes disruption

Surviving the New Advanced Computing Environment: An On-Prem/Cloud Partnership

Adopting an integrated collection of hardware, software, and architectures that harnesses the best features of both options

- On-prem HPC: 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-based HPC: 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

What Will It Take to Get There?

Both sides need to cooperatively meet in the middle

- Software standardization and interoperability
 - Open standards and common, containerization, virtualization, migration between on-prem and cloud(s)
- Planned annual budgets reconciled with CSP access/instance activity
 - Requires complementary/longer-term/deterministic budget agreements from CSPs
- Integrated hardware/software/architectures
 - CSP for exploration of new technology compliments on-prem longer term commitments
 - Data storage stressing accessibility, portability, security while minimizing overall costs (a balance of \$ and KPI considerations)
- Automated orchestration/scheduling
 - Balanced KPIs: time to solution, queue wait time, storage access scheme, time to science, job priority, fluctuating compute demands, and cost
- Holistic procurement process
 - Seeking an integrated solution across on-prem and CSP suppliers
 - Does this require a revamped budget/procurement process?
- Training for HPC and SME staff to navigate new ecosystem: hiding details through interfaces
 - Recognizing that most new hires will be primarily CSP-centric

Finally: What Steps Can End Users Take To Ensure Computing Capability, Relevancy, and Results?

Workloads drive architecture, and users drive workloads

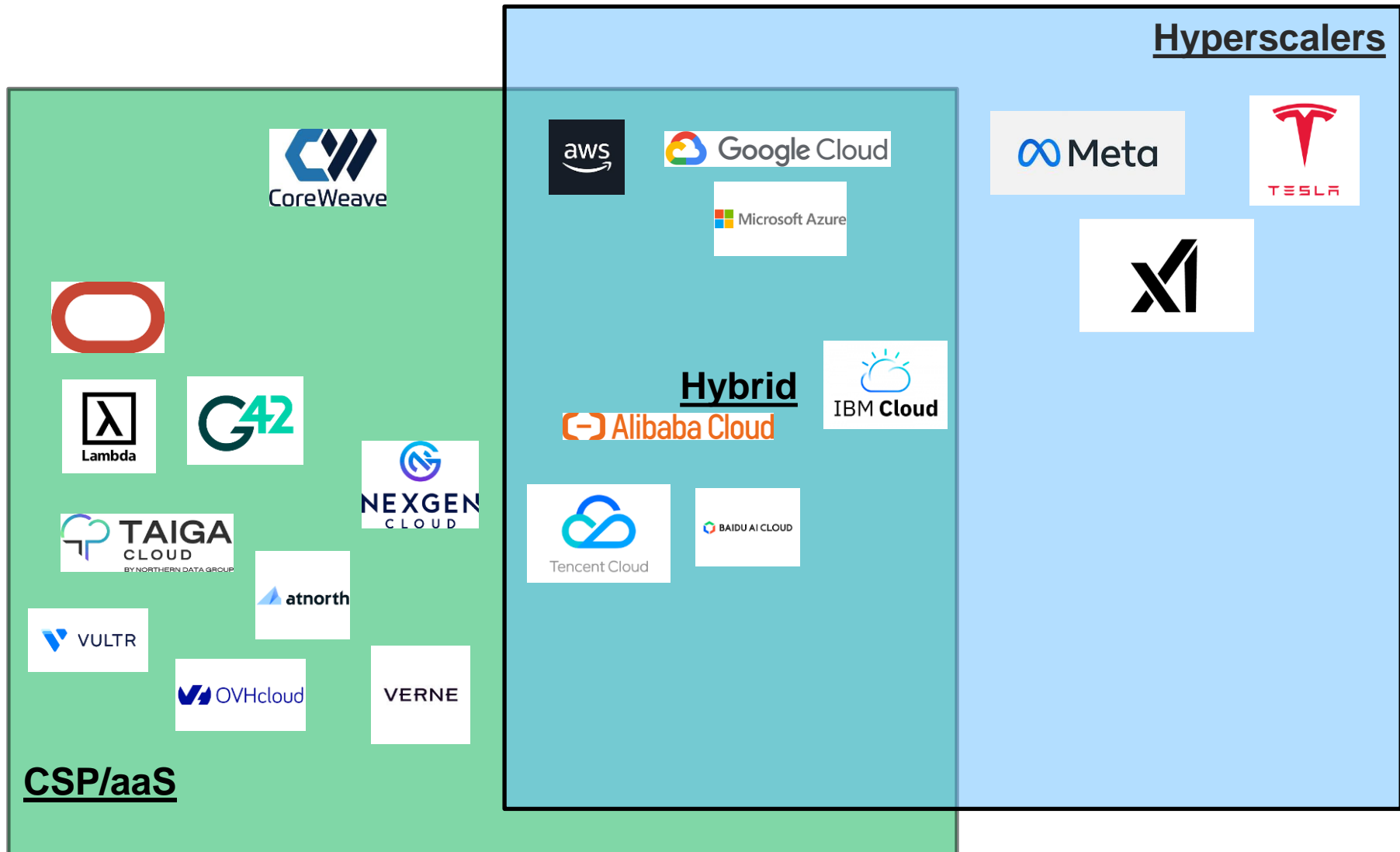
- Foster a site-wide consensus among key science drivers, user base technical requirements, and IT staff strategies on new technology strategies
- Establish a scientific advisory panel to highlights key science drivers for the next five to ten years and related impact on future workload requirements
- Collaborate on HPC operational KPIs from a science and technical perspective
- Stand up a team of SME and senior management to develop an overall strategic plan that considers science, workload, systems specifications, and related KPIs
- Participate in benchmarking/procurement activities
- Encourage interactions between the HPC and SME staff through co-location schemes, periodic rotations between HPC and SME positions
- Have in-house technical staff periodically update the user base on key technology trends not only currently available but looking out to those projected to come online in the next 3-5 years

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	
	Reduced 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 resources for internal consumption		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?



mnoskoff@hyperionres.com