Welcome To The ISC25 Hyperion Research Market Update

June 2025

Earl Joseph, Bob Sorensen,
Mark Nossokoff,
Tom Sorensen and Jaclyn Ludema

www.HyperionResearch.com
www.hpcuserforum.com

About Hyperion Research

(www.HyperionResearch.com & www.HPCUserForum.com)

Hyperion Research Mission:

- *Hyperion Research helps organizations make effective decisions and seize growth opportunities*
  - *By providing research and recommendations in high performance computing and emerging technology areas*

HPC User Forum Mission:

- *To improve the health of the HPC/AI/QC industry*
  - *Through open discussions, information sharing and initiatives involving HPC users in industry, government and academia along with HPC vendors and other interested parties*

The Hyperion Research Team

Analysts

- Earl Joseph, CEO
- Bob Sorensen, SVP Research
- Mark Nossokoff, Research Director
- Jaclyn Ludema, Analyst
- Thomas Sorensen, Analyst

Executive

- Jean Sorensen, COO

Survey Specialist

- Cary Sudan, Principal Survey Specialist

Global Accounts

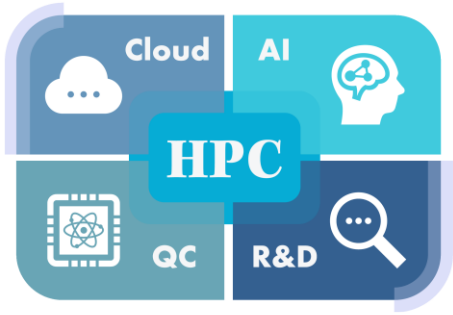
- Mike Thorp, Sr. Global Sales Executive
- Kurt Gantrish, Sr. Account Executive
- Brian Eccles, Client Services Specialist

Consultants

- Katsuya Nishi, Japan and Asia
- Kirsten Chapman, KC Associates
- Andrew Rugg, Certus Insights
- Jie Wu, China and Technology Trends
- Mara Jacob, HPC User Forum Support

Example Research Areas

(www.HyperionResearch.com & www.HPCUserForum.com)



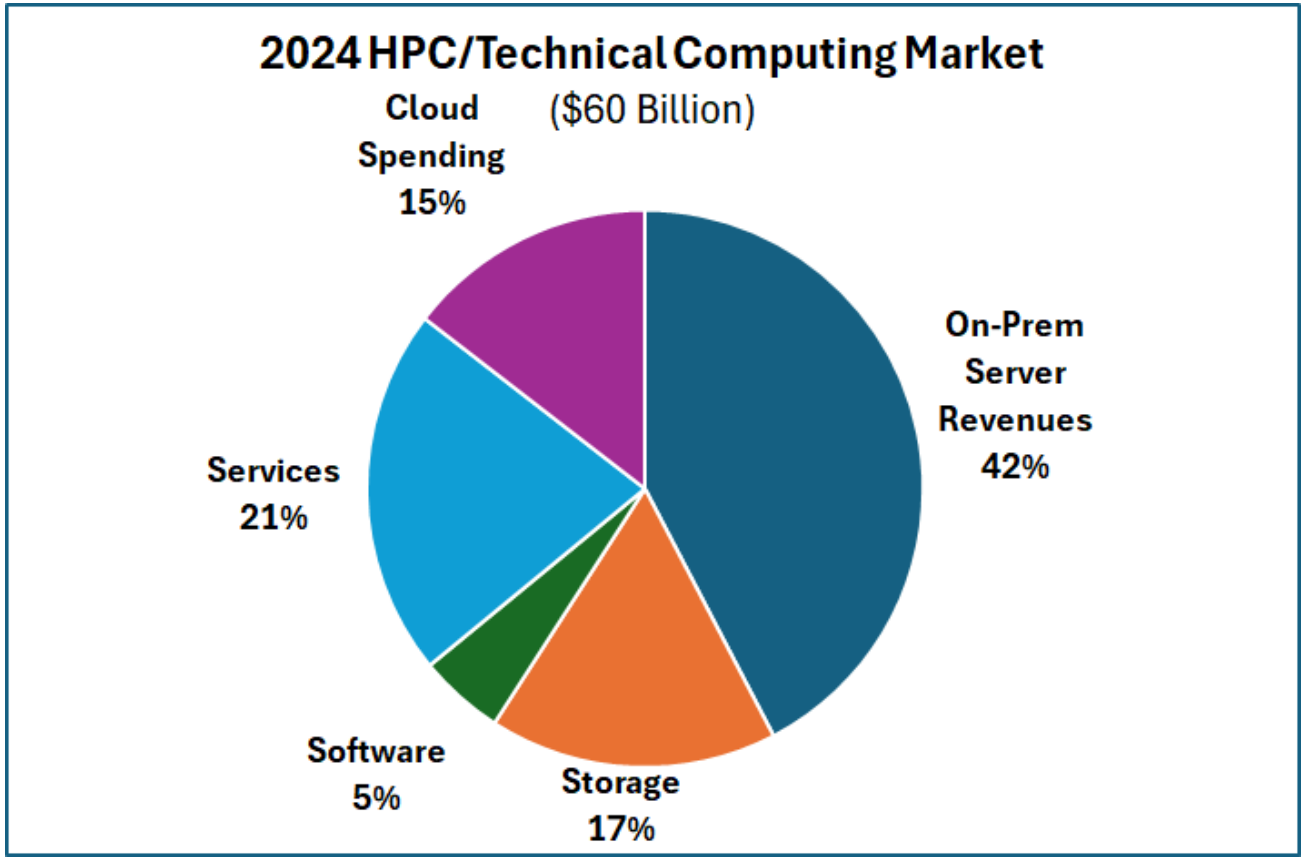
- Traditional HPC
- AI, ML, DL, LLMs, Graph
- Cloud Computing
- Storage & Data
- Interconnects
- Software & Applications
- ROI and Scientific Returns from HPC
- Power & Cooling
- Tracking all Processor Types & Growth rates
- Quantum Computing
- R&D and Engineering -- all types
- Edge Computing
- Supply Chain Issues
- Sustainability

HYPERION RESEARCH

HPC/AI Market Update

2024 Was a Strong Growth Year

The highest growth in over two decades (23.5%)!



- **23.4% growth in on-premises servers**
- **21.3% growth in the use of clouds**
- **Over $60 billion in total spending**

2024 HPC/AI Market By Vendor

The highest growth in over two decades (23.5%)!

2024 HPC/AI Market By Vendor		
Vendor	**2024 Server Revenues**	**2024 Market Shares**
HPE	7,151	28.2%
Dell Technologies	3,916	15.5%
Lenovo	1,450	5.7%
Inspur	1,082	4.3%
Atos	708	2.8%
Sugon	619	2.4%
IBM	332	1.3%
Penguin	356	1.4%
Fujitsu	233	0.9%
NEC	213	0.8%
Other HPC	2,337	9.2%
Non-Traditional Suppliers	6,934	27.4%
Total	**25,332**	**100.0%**
Source: Hyperion Research, 2025		

2024 HPC/AI Market By Segment

The highest growth in over two decades (23.5%)!

2024 HPC/AI Market By Segment		
2024 New Segments	**2024 Server Revenues**	**2024 Market Shares**
Leadership Computers (>$150M)	1,190	4.7%
Supercomputers ($10M-$150M)	6,921	27.3%
Large HPC ($1M-$10M)	7,078	27.9%
Medium HPC ($250K-$1M)	3,985	15.7%
Entry HPC (<$250K)	6,159	24.3%
Total	**25,332**	**100.0%**
Source: Hyperion Research, 2025		

2024 HPC/AI Market By Vertical

The highest growth in over two decades (23.5%)!

WW High-Performance Systems Revenue by Applications			
	2023	2024	2023 to 2024 Growth
Bio-Sciences	$1,883	$2,279	21.0%
CAE	$2,319	$2,729	17.7%
Chemical Engineering	$236	$301	27.5%
DCC & Distribution	$1,143	$1,389	21.5%
Economics/Financial	$1,044	$1,323	26.7%
EDA / IT / ISV	$1,196	$1,480	23.7%
Geosciences	$1,300	$1,543	18.6%
Mechanical Design	$058	$061	4.4%
Defense	$2,151	$2,563	19.2%
Government Lab	$4,446	$6,114	37.5%
University/Academic	$3,482	$4,012	15.2%
Weather	$940	$1,127	20.0%
Other	$350	$412	17.6%
Total Server Revenue	$20,550	$25,333	23.3%

Source: Hyperion Research, 2025

2024 HPC/AI Market By Region

The highest growth in over two decades (23.5%)!

2024 HPC/AI Market By Region		
2024 New Segments	**2024 Server Revenues**	**2024 Market Shares**
North America	13,421	53.0%
EMEA	6,168	24.3%
Asia/Pacific (All)	5,467	21.6%
Rest of World	276	1.1%
Total	**25,332**	**100.0%**
Source: Hyperion Research, 2025		

HYPERION RESEARCH

Forecasts

The HPC/AI Market Should See Growth in 2025

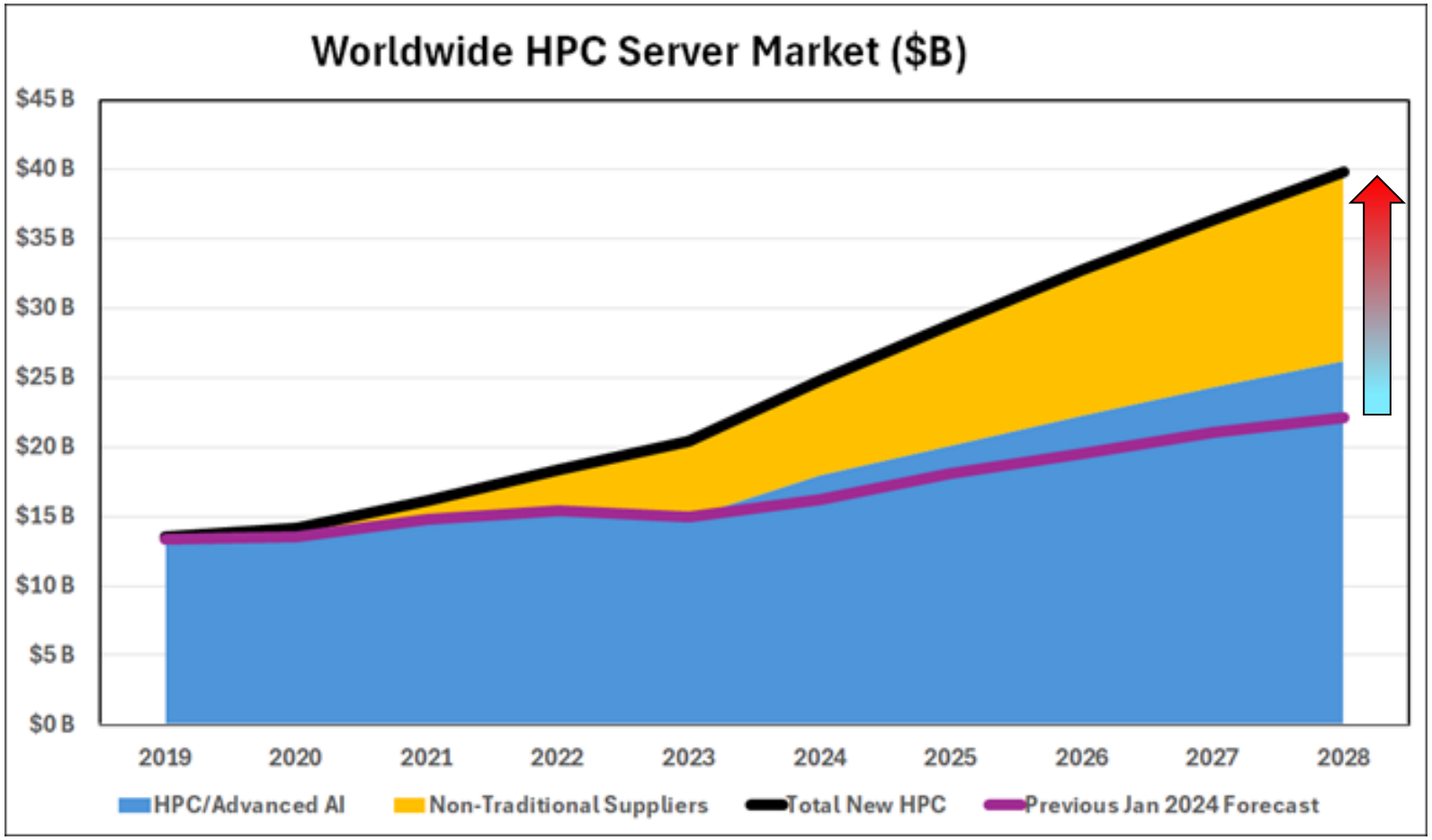
… but there are some major concerns

- **The global economic situation and changing trade rules could have a major impact to IT build outs in 2025**
- **Supply chain issues are still impacting installations (e.g., GPUs)**
- **Exascale system acceptances are seeing delays**
- **The lower end of the on-premises market continues to struggle**

- **Growth drivers include:**
  - New use cases especially in AI/LLMs/Generative AI/Smarter AI are providing new areas for users to advance their research
  - Countries and companies around the world continue to recognize the value of being innovative and investing in R&D to advance society, grow revenues, reduce costs, and become more competitive

Updated View of the On-Prem Server Market

- *Hyperion Research just announced a 36.7% increase in the HPC/AI server market size (now growing at 15% CAGR)*
- *Added tracking of non-traditional AI/HPC suppliers*



Updated View of the HPC/AI Market

On-prem HPC/AI servers are projected to exceed $47 billion in 2029

Worldwide Overall HPC Server Market Forecast ($M)								
	2023	2024	2025	2026	2027	2028	2029	CAGR 24-29
Total HPC	20,550	25,333	29,159	32,713	36,909	41,681	47,115	13.2%
Historic HPC/Advanced AI	14,768	17,875	19,288	21,120	23,295	25,695	28,341	9.7%
Non-Traditional Suppliers	5,782	7,458	9,872	11,593	13,614	15,987	18,774	20.3%
Source: Hyperion Research, 2025								

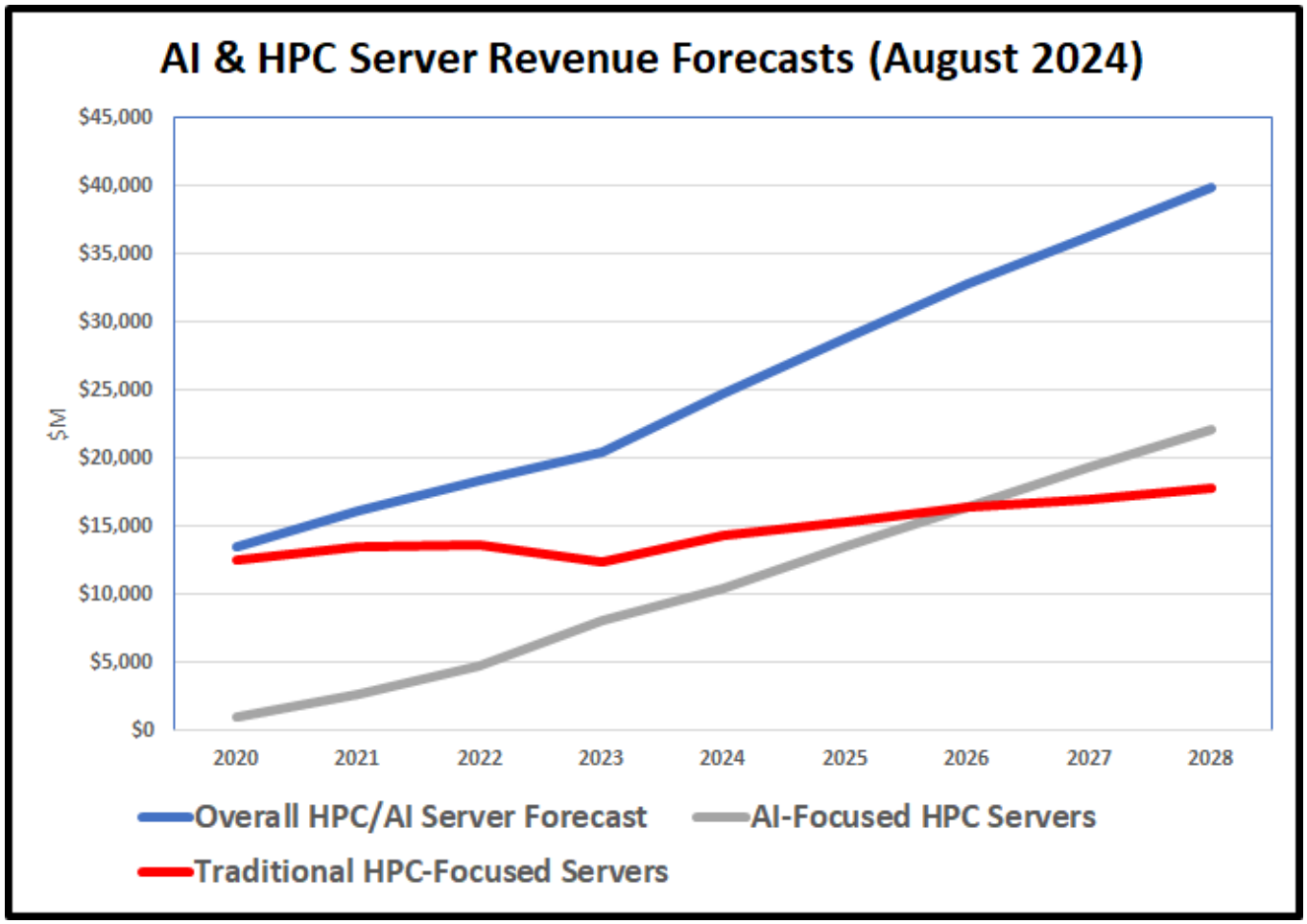
Market Segment Definition: *Non-Traditional Suppliers* *(new revenues added to the previous HPC market sizing)*

These are on-premises AI-centric HPC servers that are provided by non-traditional HPC suppliers like NVIDIA, Cerebras, SambaNova, SuperMicro, etc. These servers are designed primarily to run AI and AI-related workloads

- These servers are a subsegment of the overall HPC market but haven't historically been accounted for within prior HPC market numbers

HPC Compared to AI-centric Servers

Many servers are running both traditional HPC and AI Workloads



Note: AI systems may still run some traditional HPC jobs (<50% of workload).
Likewise, traditional HPC systems often run some AI jobs (<50% of workload).

The Exascale Market (System Acceptances)

Over 45 systems and over $12 billion in value

Year Accepted	China	Europe	Japan	US	Other Countries*	Total Systems	Total Value
2020			1 near-exascale system ~$1.1B			1	$1.1B
2021	2 exascale ~$350M each	1 pre-exascale system ~$180M	--	1 pre-exascale system ~$200M	--	4	$1.1B
2022	1 exascale ~$350M	2 pre-exascale systems ~$390M total	--	1 exascale system ~$600M (2/3 accepted 2022)	--	4	$1.1B
2023	--	2 pre-exascale systems ~$150M each	1 near-exascale system ~$150M	Remaining 1/3 of Frontier system	--	3	~$0.5B
2024	1 exascale system ~$350M	1 pre-exascale ~$150M	--	2 exascale system ~$600M each	--	4	~$1.7B
2025	1 or 2 exascale systems ~$300M each	2 or 3 exascale systems ~$350M each	1 exascale system ~$200M	1 or 2 exascale systems ~$350M each	1 near-exascale system ~$125M	6-9	$1.7B - $2.7B
2026	2 exascale systems ~$300M each	2 or 3 exascale systems ~$325M each	?	1 or 2 exascale systems ~$325M each	1 or 2 exascale systems ~$150M each	6-9	$1.7B - $2.5B
2027	2 exascale systems ~$275M each	2 or 3 exascale systems ~$300M	1 exascale system ~$150M	1 or 2 exascale systems ~$275M each	2 or 3 exascale systems ~$130M each	8-11	$1.8B - $2.5B
2028	2 exascale systems ~$250M each	2 or 3 exascale systems ~$275M	1 or 2 exascale systems ~$150M each	1 or 2 exascale systems ~$275M each	2 or 3 exascale systems ~$125M each	8-12	$1.7B - $2.6B
Total	**11-12**	**14-18**	**5-6**	**8-12**	**6-9**	**44-57**	**$12.4B - $16.8B**

* Includes S. Korea, Singapore, Australia, Russia, Canada, India, Israel, Saudi Arabia, etc.

Note: After 2023, many exascale systems will be 2-10 exascale.

Source: Hyperion Research, March 2025

Conclusions

- **2024 was a strong growth year**
  - GPUs, cloud, AI/ML/DL/LLM were high growth areas
- **There are many high growth areas**
  - Using clouds to run HPC & AI workloads
  - All types of AI workloads
  - QC systems are being installed around the world
  - Storage will see major growth driven by AI, big data and the need for much larger data sets
- **New technologies are showing up large numbers:**
  - Generative AI, smarter AI, LLMs and SLLs are fueling a new level of growth
  - Processors, AI hardware & software, memories, new storage approaches, etc.
  - The cloud has become a viable option for many HPC workloads
- **There are growing concerns around power & talent**

HYPERION RESEARCH

We Welcome Questions, Comments and Suggestions

Please contact us at:

info@hyperionres.com

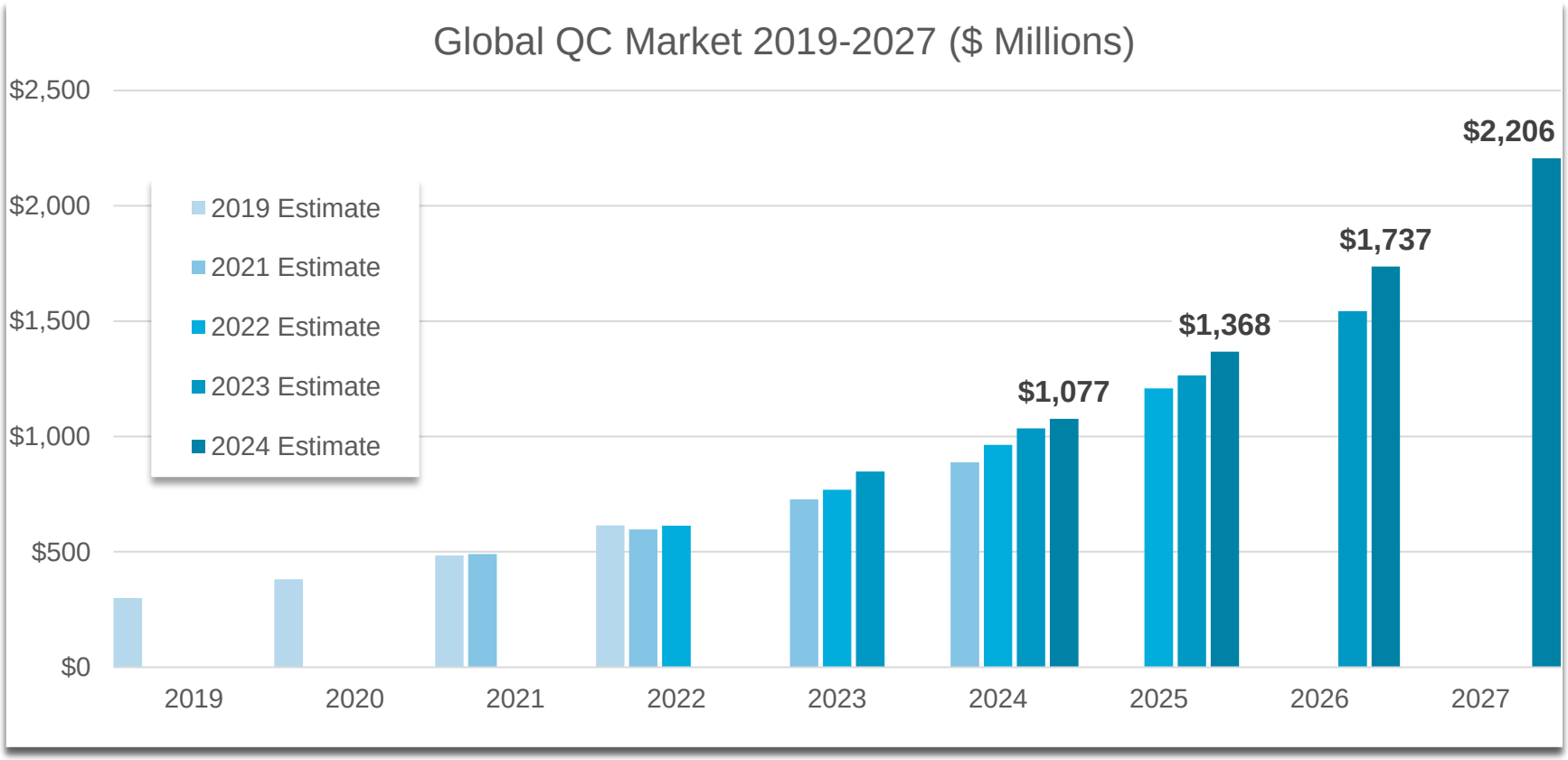
5th Annual Global QC Market Survey: Continued Progress But Changes in the Air

Bob Sorensen
Chief Analyst for Quantum Computing
Hyperion Research, LLC

www.HyperionResearch.com
www.hpcuserforum.com

QC Market Estimate: $1.07 billion in 2024

27% annual growth rate drives global QC market to $2.2 billion in 2027



- Exponential curve begins to dominate growth
- Consistently underestimating growth?

QC Market 2027: Top End User Sectors

Chemicals and QC R&D on top, but broad applicability envisioned

- Chemical/Chemistry sector hits number 1
  - Up from #2 last year, #4 year before that
  - Reinforces early emphasis on mod/sim, especially computational chemistry, as major algorithm
- Likewise, pharmaceuticals continues its upward climb
  - 21% last year, 35% this year
- Applicability spans academic, commercial, and government spaces
- Finance drops from 30% to 21%
  - Optimization issues, saturation or contrived lack of visibility?
- Government labs hold steady, for now
- Although nearly every sector choice deemed important by some, there are clear concentrations in key areas

QC Market 2027: Primary End User Motivations

New algorithms and future classical performance concerns lead

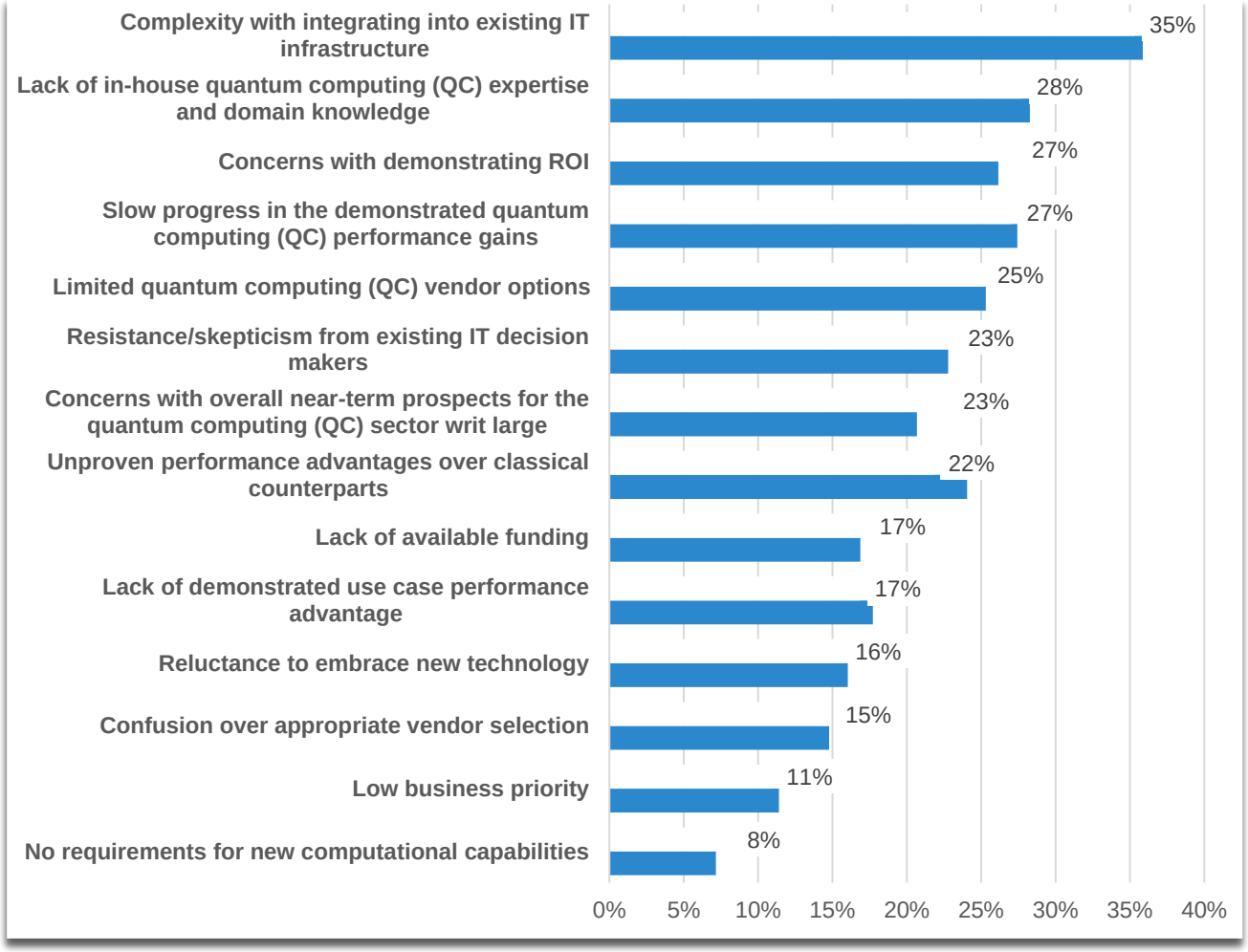
Option	% Selected
Implement new algorithm(s) not possible on classical counterpart systems	56.5%
Address concerns with future performance capabilities of classical computing systems	51.3%
Explore organizationally relevant QC use case potential with no expectations of near-term advantage	47.0%
Develop in-house familiarization with QC skills with no expectations of near-term end use deployment	45.2%
Engage with the QC vendor community for future activities	31.3%
Enable better real-time computational capabilities	27.8%
Realize faster turnaround time on existing classical counterpart systems	27.0%
Reduce overall computing systems costs	23.5%
Reduce overall computational power and cooling requirements	14.8%
Don't know/Not sure	5.2%
Other	2.6%

N = 115, Select all that apply

- Implement new algorithms and address concerns with future classical performance selected by majority of respondents
  - Classical developments could impact QC uptake
- An average of 3.3 options selected per respondent
- Many are still exploring for the sake of exploration
- One in four are looking at real-time compute opportunities
- Reduce overall compute systems cost:
  - 2023 Survey: 9.0%
  - 2025 Survey: 23.5%
- Reduce Power/Cooling Costs
  - 2023 Survey: 17.3%
  - 2025 Survey: 14.8%

Greatest Hurdles to QC Adoption

Led by complexity with IT integration and a lack of in-house QC expertise



- These are typical concerns found in many major advanced computing surveys going back five years or more
  - Spanning classical HPC, GPUs, AI, and now QC
- An average of ~3 identified hurdles per respondent
- No requirements for new compute capability: a perennial low number
  - Signifying pent-up demand for new solutions

Respondents could select all options that apply.

QC Partnerships: With QC End Users

Most respondent organizations have a range of partnerships with QC end users

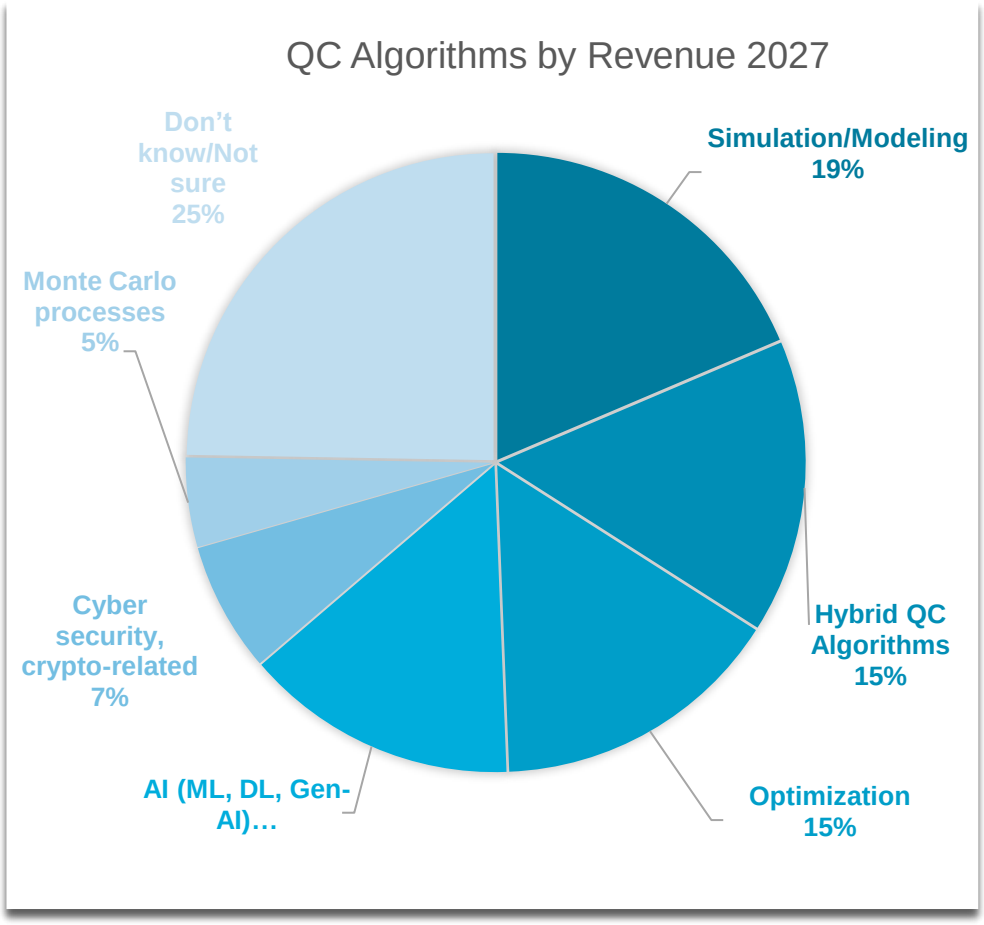
Option	% Selected
Explore new QC sector/vertical-specific QC-related opportunities	74.1%
Field test/evaluate new QC hardware	44.8%
Field test/evaluate new QC software	44.8%
Explore key performance gains over classical counterpart	43.1%
Establish sector-specific capabilities	41.4%
Foster public attention	39.7%
Encourage follow-on sales	36.2%
Explore QC/classical integration issues	31.0%
Access QC end user QC expertise	29.3%
Explorer QC sector/vertical-specific performance opportunities on existing classical workloads	27.6%
Access QC end user classical IT expertise	8.6%
Other	5.2%

N = 58, Select all that apply

- 71% of respondent organizations have a partnership with at least one QC end user
- Average respondent selected 4.2 options
- Field testing QC hardware and software both selected 44.8%
- Exploring sector-specific opportunities was overwhelming justification (74.1%)
- Building sector-specific skills (41.4%) and exploring QC performance advantages (43.1%) also key drivers

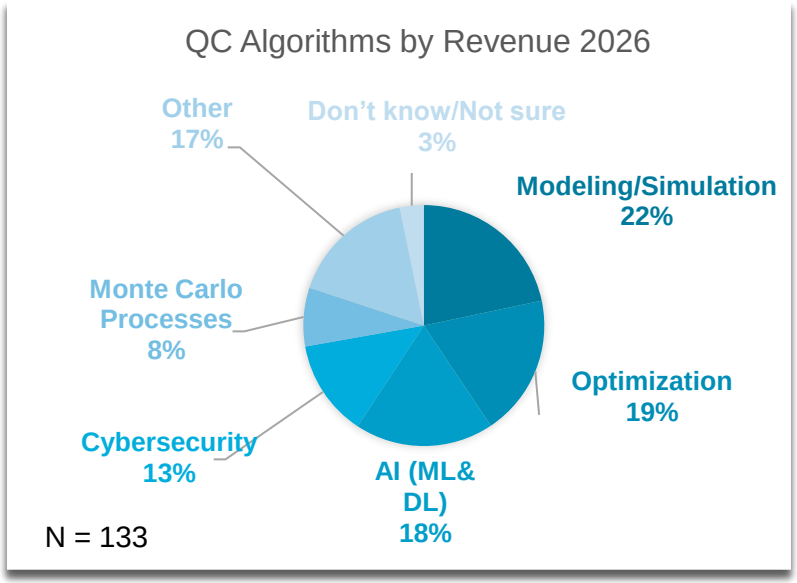
QC Market 2027: Major Algorithms by Revenue

Mod/sim #1 algorithm, but hybrid comes to the fore



N = 115

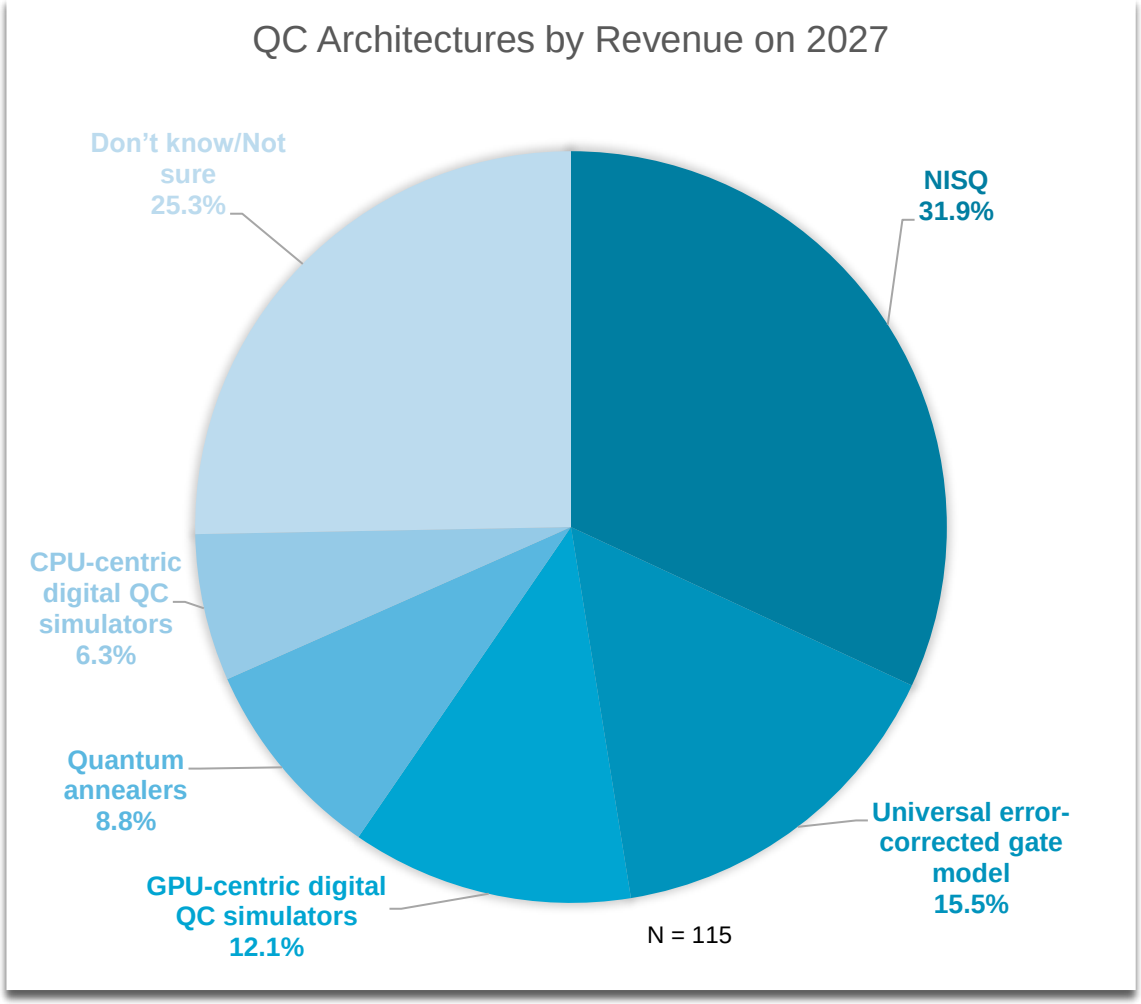
- Modeling/simulation remains at #1
- Appearance of hybrid QC algorithms follows refinement of Others option from previous years
- Don't know/Not Sure dominates responses
  - Is this a problem for the QC supplier base?



N = 133

QC Market 2027: QC Architectures

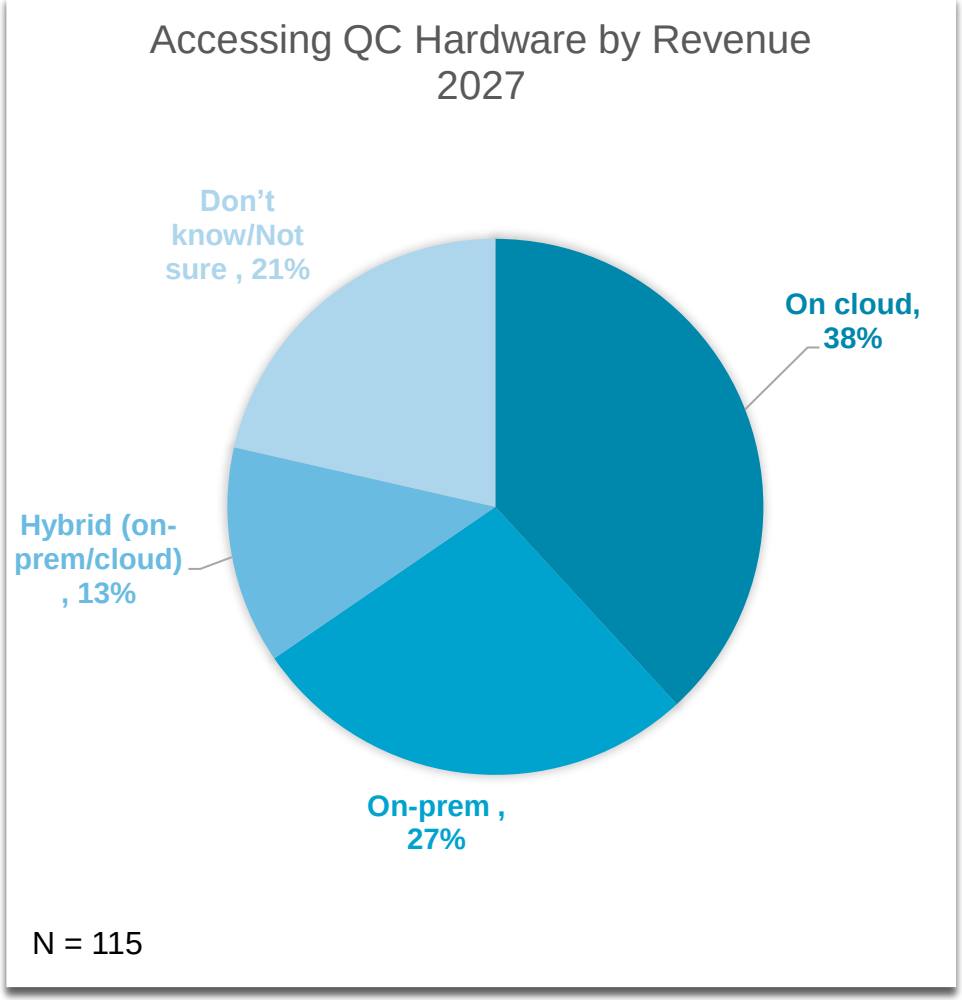
NISQ maintains lead, QC simulators still major element of QC architecture



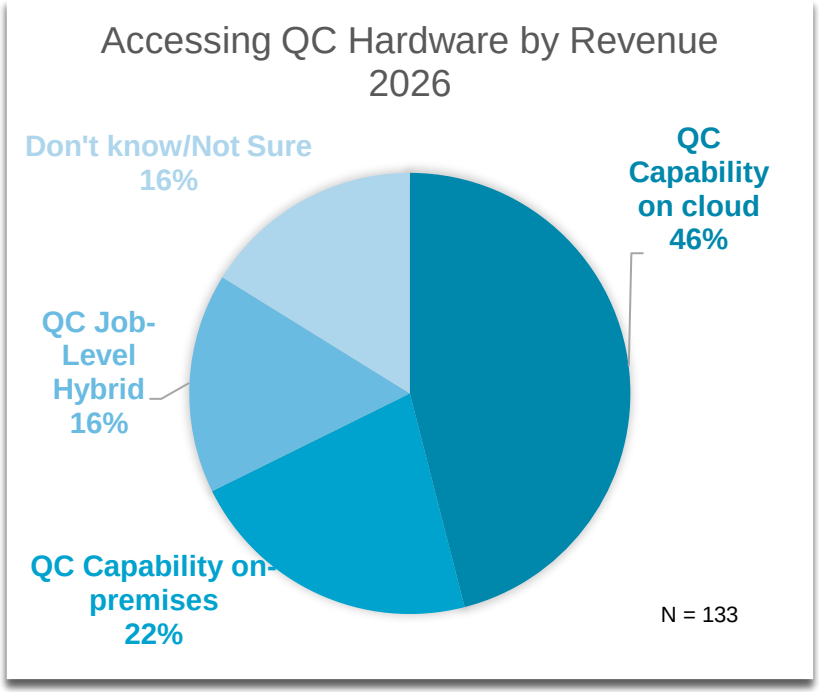
- NISQ dominates QC architecture in 2027
  - 2X universal error corrected gate model alternative
- Digital simulators (CPU and GPU based) combine for almost 19% of hardware market
  - But GPUs are preferred at 2x CPU rate
  - Room for options here
- Many Don't Knows/Not Sures
  - More fence sitting or lack of information?

QC Market 2027: Access to QC Hardware

Cloud continues to dominate but on-prem moving up

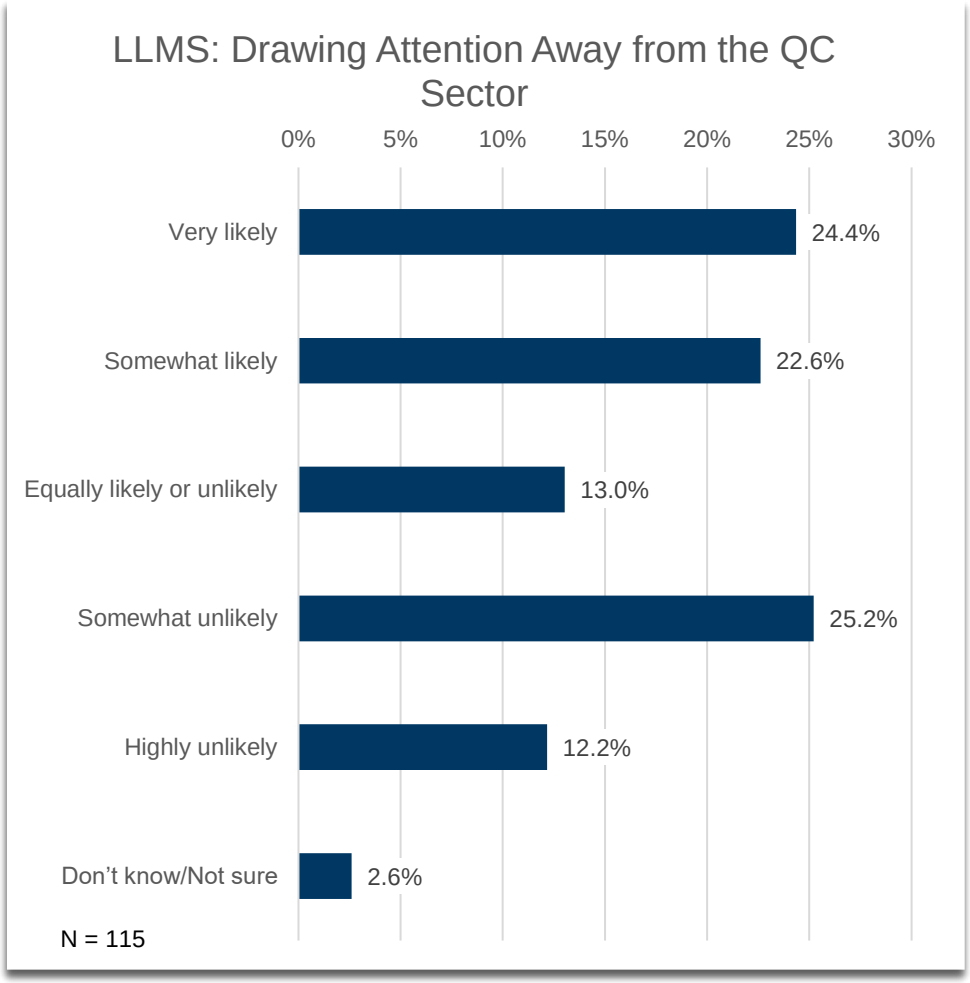


- Cloud revenues move from 46% to 38%
- Biggest move since tracking this number
- More Don't know/Not sure: more fence sitting?



QC Distractions and LLMs

How likely is it that the emergence of large language models like ChatGPT and BERT will draw attention away from end user interest in quantum computing?



- LLMs - and likely generative AI in general - seen as near-term competitor for end user interest in QC by 47% of respondents
  - Up from 42% last year
- 37% not overly concerned
  - Down from 42% last year
- Demonstrates need for QC to continue to deliver on technology/performance gains
- Highlights perceived end user interest in performance gains no matter how it is delivered

QUESTIONS?

bsorensen@hyperionres.com

Float to the top or sink to the bottom. Everything in the middle is the churn.

- Amos Burton, Engineer *The Expanse*

Perspectives on HPC-AI Storage and Interconnects

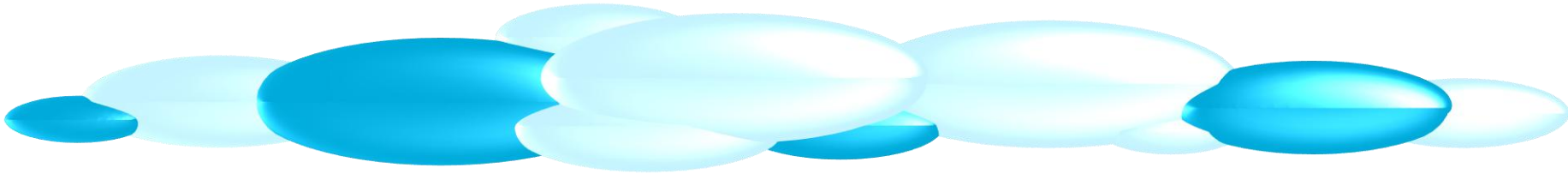
ISC25 Market Update Briefing
June 2025

www.HyperionResearch.com
www.hpcuserforum.com

Mark Nossokoff

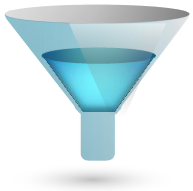
AI Data Pipeline

Diverse range of profiles and requirements



Prep

Checkpoint



Ingest

Train

Inference

AI Data Pipeline Storage Considerations

	Ingest	Prep (ETL)	Train	Checkpoint	Inference
Access Pattern	Sequential	**Sequential** or Random	Random	Sequential	Sequential
Access Type	Writes	**Reads** and Writes	Reads	Writes	Reads
Access Frequency	Idle ←→ Intense	Moderate	Idle ←→ Intense	Idle ←→ Intense	Moderate to Intense
Data Size	Small to Large	Small to Large	Mostly Small	Small to Large	Small to Large
Locality	Edge	Edge, Cloud, On-premises	Cloud, On-premises	Cloud, On-premises	Edge, Cloud, On-premises

*ETL – Extract, Transform, Load
Source: Hyperion Research, 2024

- **Training frequency (new foundation, RAG, pre-trained)**
- **Model type and size**
- **Data type (structure, unstructured; file, block, object)**
- **Data mode (text, image, video)**
- **Security**
- **Compliance (what data to save and for how long?)**
- **Parallel file system – is one a requirement?**

What's Happened in Storage and Interconnects Since SC24?

Storage

- **Blackstone $300M investment in DDN**
- **Vendor announcements**
  - System vendors
    - Dell
    - HPE
  - Data platform vendors
    - DDN
    - Hammerspace
    - VAST
    - Weka

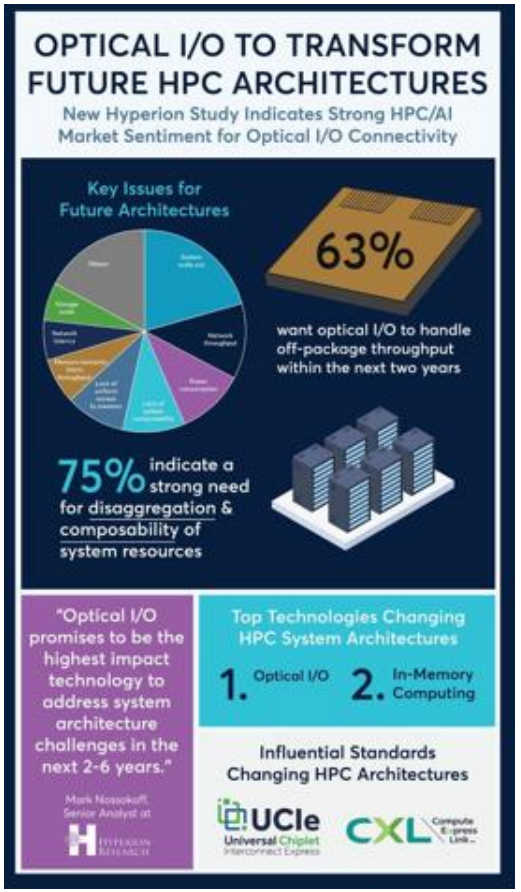
Interconnects

- **GTC items**
  - On-chip optics
  - Heavy emphasis on ethernet; limited mention of InfiniBand
- **Progress in standards**
  - UltraEthernet Consortium (UEC)
    - Contributions to Linux kernel
  - UltraAccelerator Link (UALink) Consortium
    - Released version 1.0 of spec
  - EuroHPC JU NET4EXA
    - Deploy BXIv3 exascale integration
    - Intent to develop BXIv4 roadmap
- **NVLink Fusion for 3rd party integration**

Strong Sentiment Toward Optical I/O

Both users and vendors eagerly anticipating optical I/O

- **Predominant system issues for future architectures**
  - System scale-out
  - Lack of system composability
  - Network throughput
- **Optical I/O was rated as technology that has highest potential to improve HPC architectures in next 2-6 years**
- **75% of respondents felt that there is a strong need for disaggregation of system resources**



Graphic: Ayar Labs, 2023

Prior Study

Future Research Direction

The more things change, the more they stay the same

- **Continued impact of the AI data pipeline on storage architectures**
- **Evolution of data platform SW stack**
- **AI workflow impacts on interconnect architectures**
- **Evaluating and determining optimized utilization of on-premises and cloud storage resources**
- **Maturity and adoption of optical interconnects**
- **Convergence or differentiation between interconnects (InfiniBand, Ethernet, OmniPath, BXI) as a result of standards activities (UEC, UAL) and changes in vendor strategy (NVIDIA NVLink Fusion)**

HYPERION RESEARCH

Questions?

mnossokoff@hyperionres.com

Perspectives on HPC-AI in the Cloud

ISC25 Market Update Briefing
June 2025

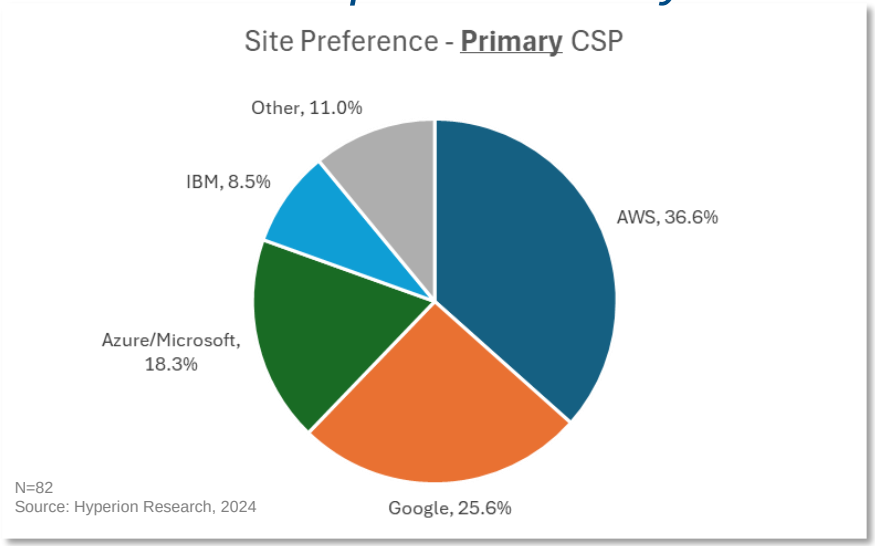
www.HyperionResearch.com
www.hpcuserforum.com

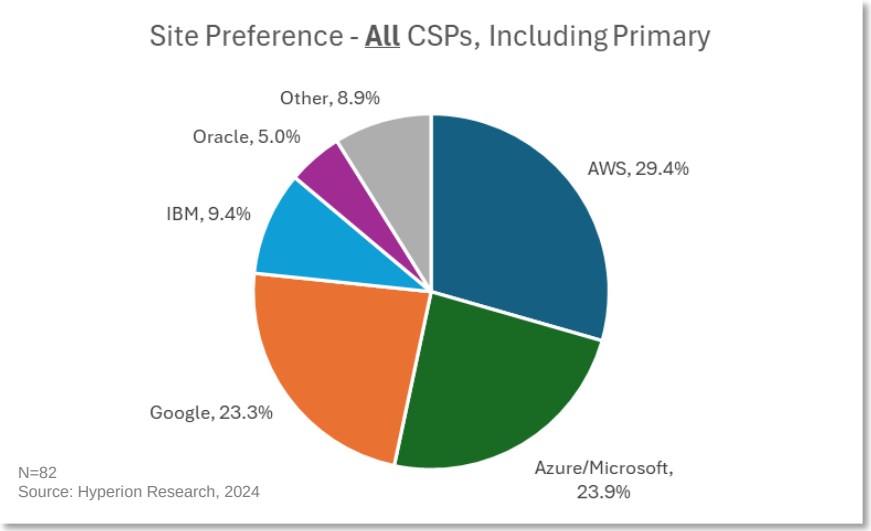
Mark Nossokoff

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

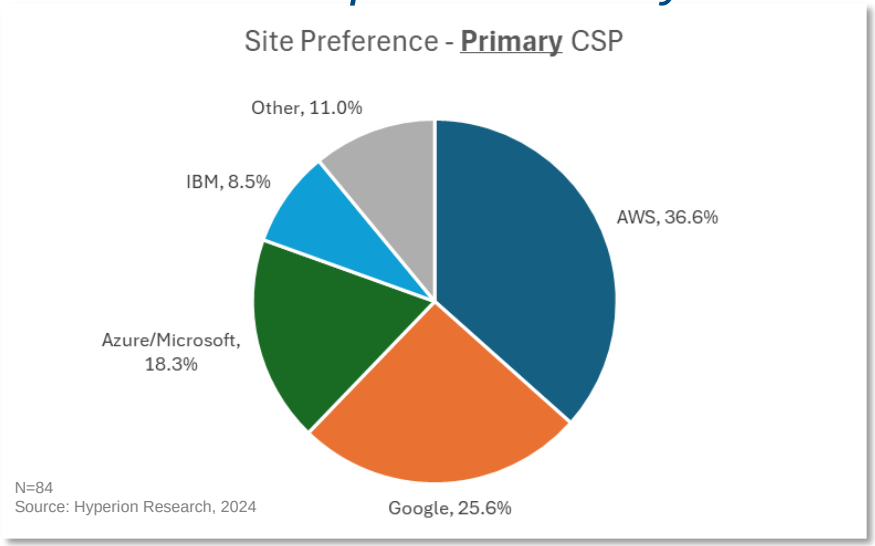


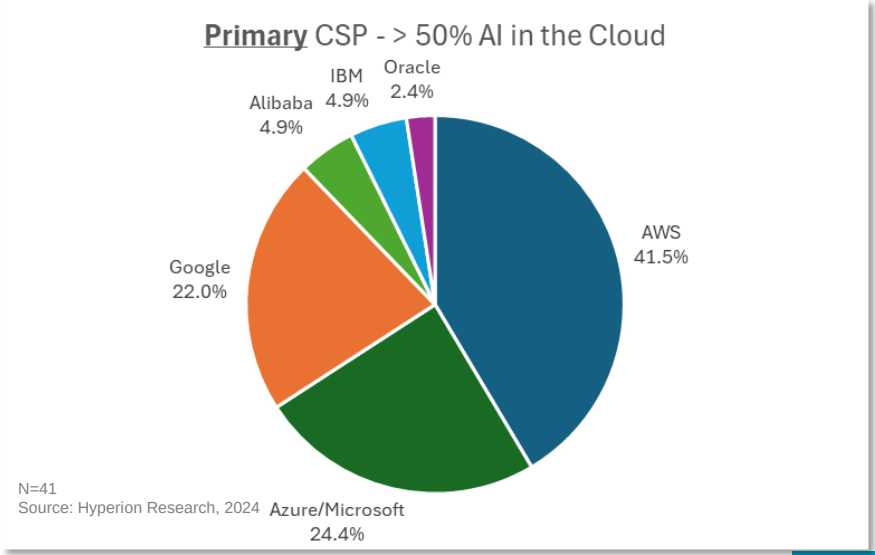


CSP Preferences – AI Workload Crosscut

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**

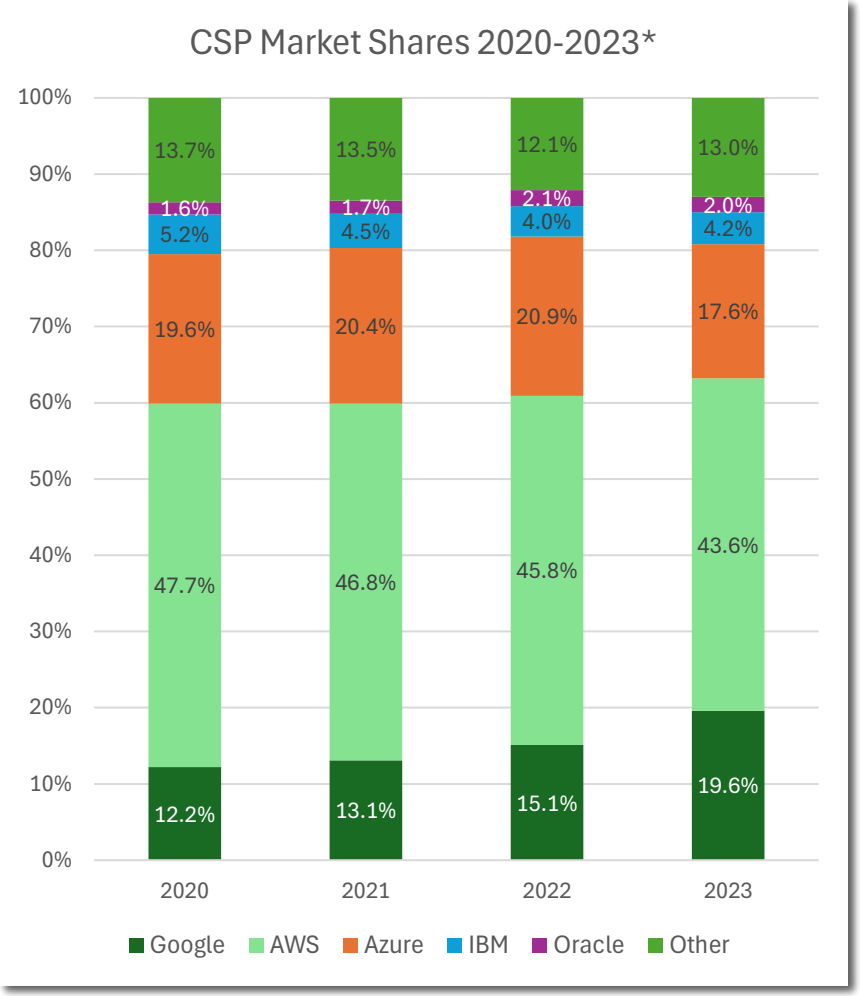




Estimated CSP HPC-AI Market Shares

AWS maintains highest share

- **Google is gaining share**
- **"Other" is also gaining share**
  - European clouds
  - China clouds
  - Neo-clouds (AIaaS, 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?

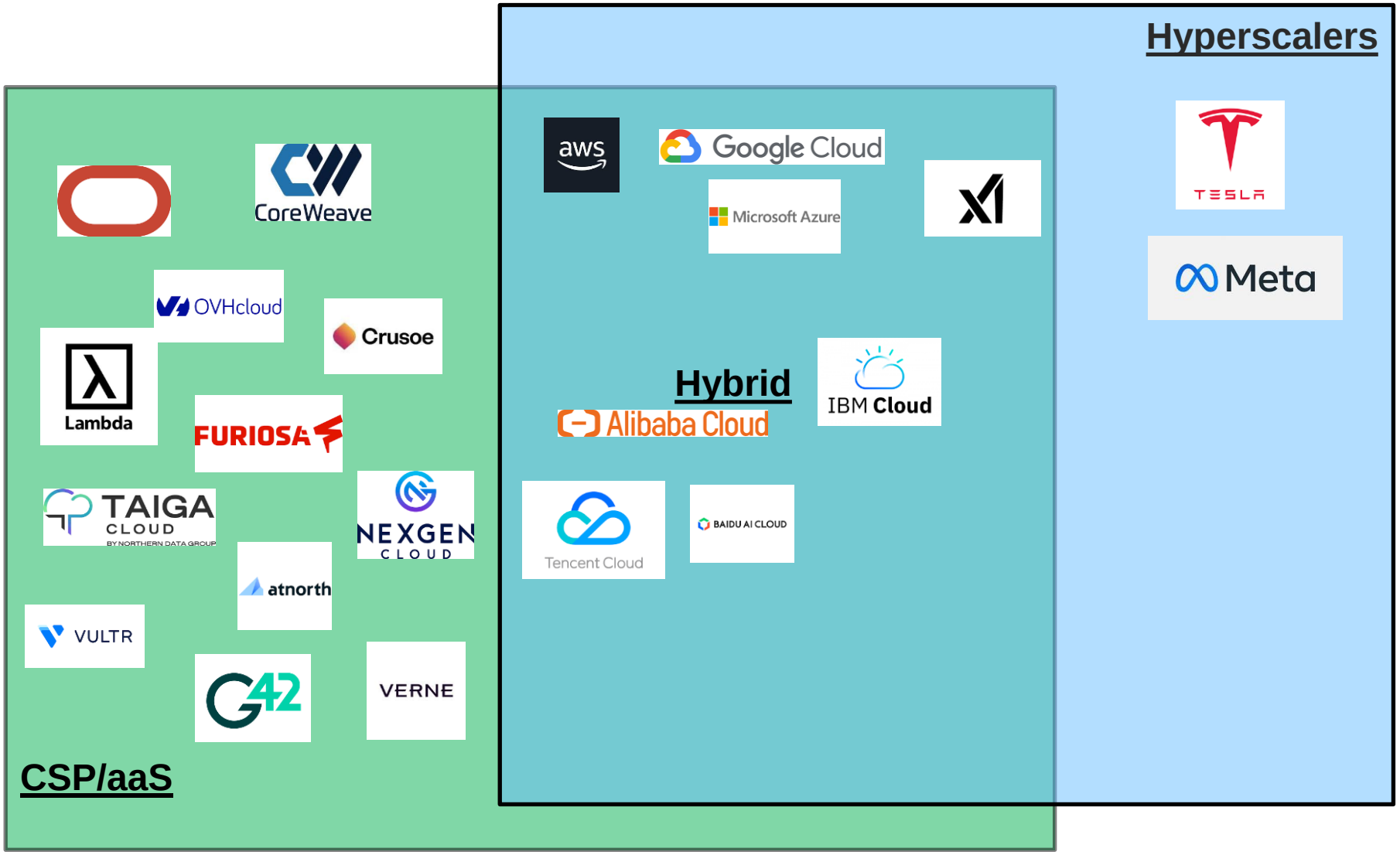
- **Google Cloud Platform**
  - New TPU7
  - New H4D CPU VMs
  - NVIDIA Blackwell support
  - Cluster Toolkit and Cluster Director
  - Google Cloud Managed Lustre
  - Agent Engine in Agent Space
- **AWS**
  - Trainium (GA for T2; preview announcement for T3; EC2 instances)
  - EC2 P6-B200 NVIDIA Blackwell 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 AIaaS and GPUaaS providers, to meet their compute needs

- **Acceleration of Cloud Adoption for AI Workloads**
  - AIaaS 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**

HYPERION RESEARCH

Questions?

mnossokoff@hyperionres.com

HPC Data Center Energy Challenges and Sustainable Solutions

May 2025

www.HyperionResearch.com
www.hpcuserforum.com

Jaclyn Ludema

The Energy Challenge

Growing energy demand is heavily influencing global market

- **Data centers consumed 240-340 TWh in 2022 (IEA estimate)**
  - That's 1-1.3% of total global demand
  - This is expected to double by 2026
- **The US is seeing rapid growth in electricity demand**
- **Many European markets are shifting**
  - Ireland-halted new data center developments near Dublin until 2028 to ensure grid stability
  - Amsterdam maintains a similar moratorium to address environmental concerns
  - Shifting investments towards countries like Portugal, Spain, Sweden, and Finland
- **AI and HPC workloads drive higher power densities**

Energy Resource Program

Global initiatives powering sustainable HPC data centers

- **United States: COOLERCHIPS (2023)**
  - Program for energy-efficient cooling
  - Support for solar, wind, and battery storage projects
- **European Union: REPowerEU Plan (2022)**
  - Accelerating clean energy projects to reduce dependency on fossil fuels
- **China: Hydropower Investments for Hyperscale Centers (2023)**
  - Leveraging renewable energy for major data hubs
- **Japan: Cool Japan Initiative (2023)**
  - Encouraging the adoption of advanced cooling technologies in high-density data centers
  - Focuses on R&D funding for next-gen cooling solutions

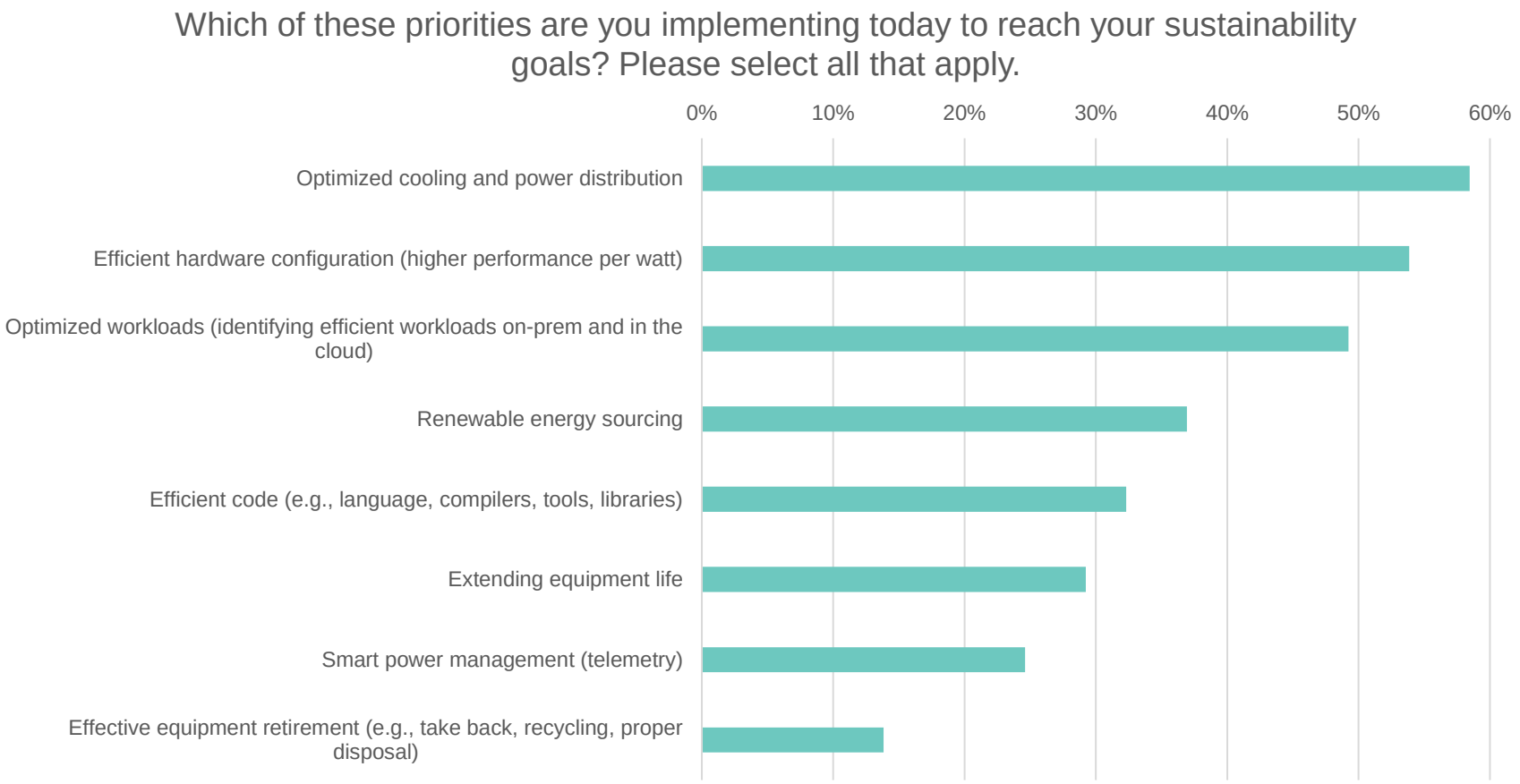
Innovative Solutions for Efficiency

Liquid cooling, heat recovery, and innovative server solutions

- **Liquid cooling adoption in HPC centers**
  - Recent study finds 67% of sites use some form of liquid cooling today
  - Projected to increase to 80% in 12-18 months
  - Incremental cooling upgrades more commonly found in brownfield sites (L2A cooling)
- **Heat recovery systems**
  - Nebius – tripling the capacity of data center in Mäntsälä, Finland. Currently recovers 20,000MWh/yr
  - Microsoft data center cluster – Expected to supply heat for 40% of Espoo, Finland (100,000 homes)
- **More players in single-socket server market: doubling core count while reducing power usage**

Trends in Sustainability Strategies

On average, sites implement 3 priorities for sustainability goals

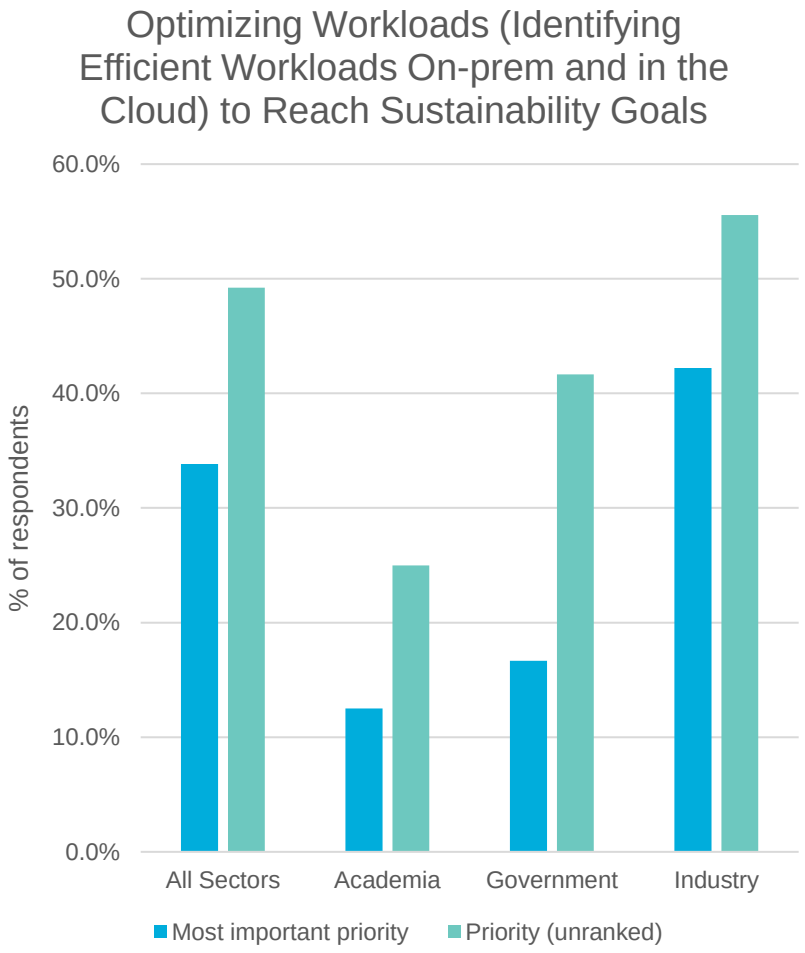


Hyperion Research 2025, n = 65

Adding Cloud in Sustainability Strategies

Sites are using Cloud as another tool to reach ESG goals

- **“Neo-clouds” – avoiding costly upgrades while reducing carbon footprint**
- **Hyperscalers (AWS, Google, Microsoft) lead with 100% renewable energy goals**
  - NVIDIA’s Reno, NV facility (solar and hydroelectric)
  - Google’s Groningen, Netherlands center (renewable energy infrastructure)



Small Modular Reactors (SMRs)

Harnessing nuclear innovation for HPC energy needs

Movement Towards SMRs

- **Consistent, carbon-free power suitable for 24/7 HPC operations, in compact, near-site design**
- **Addresses energy demands of AI and HPC workloads**
- **Recent Initiatives:**
  - **Google**- Kairos Power
  - **Amazon-** Energy Northwest
- **DOE Support: $900M in federal funding for next-gen nuclear**

Challenges

- **High initial costs: NuScale's Utah project failed due to cost escalation**
- **Regulatory hurdles: Complex permitting and compliance requirements**
- **Need for LARGE stable customer commitments to ensure project viability**

HYPERION RESEARCH

THANK YOU!

Questions or comments are welcome!

Please contact me:

jludema@hyperionres.com

Understanding the Evolving Use of AI in HPC

June 2025

www.HyperionResearch.com
www.hpcuserforum.com

Tom Sorensen

Maturing AI Use Raises New Questions

As the technology continues to be further integrated into HPC environments, challenges and opportunities expand

- **Continued integration progress of AI among HPC users prompting longer-term perspectives:**
  - How to efficiently procure resources
  - How extensively cloud resources should be used
  - Comprehensiveness of regulatory guidelines
- **Despite realized advantages, users are more realistically assessing challenges:**
  - High cost of upkeep including power & infrastructure
  - Continual education of in-house expertise
  - Management of shifting regulatory demands

Forecasting in a Shifting Environment

Hyperion Research AI forecasts are still being fine tuned

- **Forecasts for server and other hardware procurements is evolving due to major changes in the market**
  - Increased yet often exploratory use of cloud resources
  - Continued assessment of appropriate hardware/software for application
  - Hastened accelerator/GPU release cycles
  - Diversification of language models in domains

Activity	% Selected
Exploring the range of potential performance enhancements by integrating inferencing technology into existing HPC-based scientific and engineering workloads	57.0%
Exploring in-house requirements for integrating inferencing into HPC-based scientific and engineering workloads	52.0%
Testing/assessing inferencing-integrated workload performance	39.0%
Running production level inferencing-enabled workloads	37.0%
Procuring access to necessary inferencing software	30.0%
Procuring access to necessary inferencing hardware	27.0%
Passively monitoring inferencing technology developments	26.0%
Porting inferencing capability into existing workloads	25.0%
Standing up limited inferencing-integrated pilot programs	23.0%
Reaching out to inferencing hardware and software suppliers for information	22.0%
Standing up fully funded inferencing research efforts	17.0%
No current activity or Don't know/Not sure	1.0%
Other	3.0%

N=100
Source: Hyperion Research, 2025

Ongoing Hyperion Research Studies

Hyperion Research continues long-term series of HPC/AI studies

- **Continued series of studies tracking HPC/AI user behaviors, expectations, and challenges**
  - Began with LLM study, continued to integration, inference, and moving on to ROI in June 2025
  - Inference provisioning and management has become an area of heightened focus
  - As “hype” fades, ROI will receive greater attention
- **End User Inferencing: Completed Last Month**
  - Targeted towards the inferencing side of production and near-production integration of advanced AI/LLM
  - The survey dove into the hardware and software requirements of user groups and organizations managing high inferencing demands, as well as related budgetary and infrastructure requirements
  - Survey respondents provided insights on their specific inference types, level of integration and experimentation, and other details of their advanced AI usage including plans and methods of scaling

Inference Study Select Key Findings

Inference is of high importance to HPC users, experimentation continues

- **Users most frequently indicated that they are still exploring/experimenting with AI-centric options both in the cloud and on-premises**
- **Concerns centered on integration complexity, hardware/software cost, and technical issues**
- **HPC users report a nearly even split between on-premises and cloud budgeting**
- **A plurality of the software resources being used to support AI inferencing is open source**

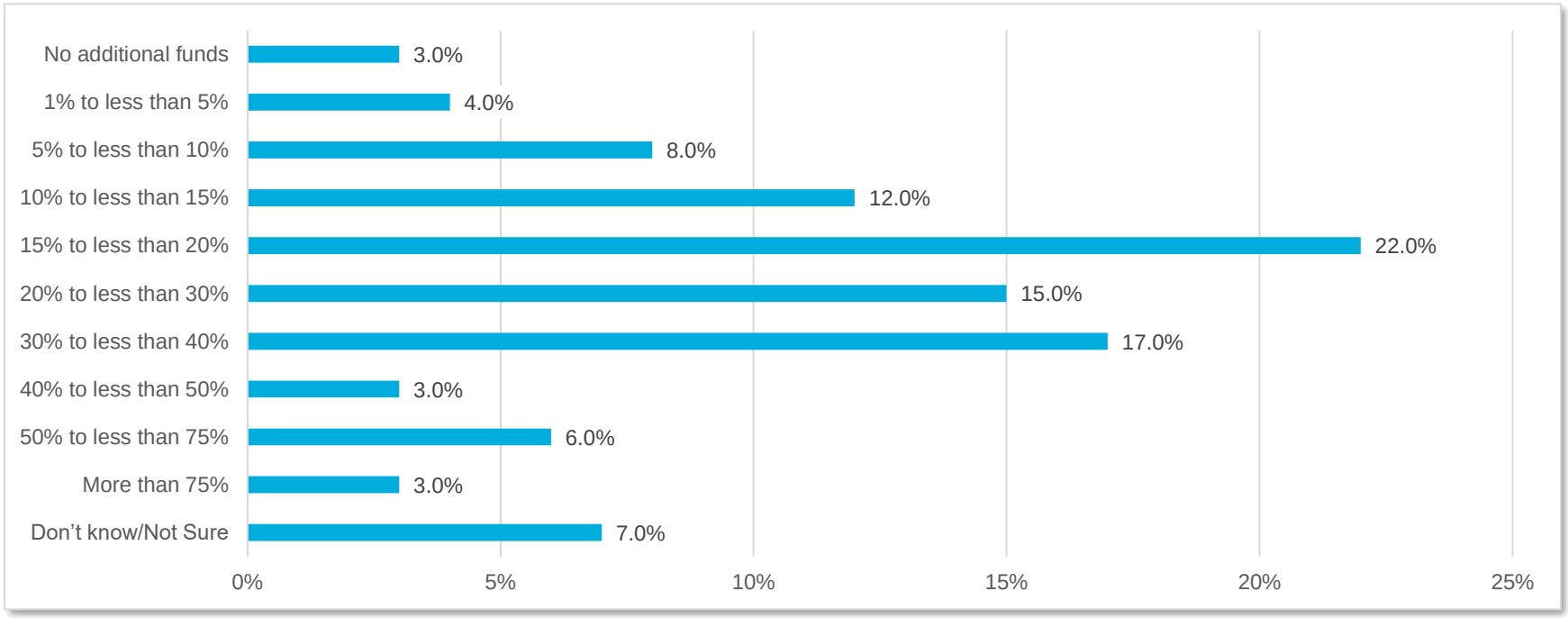
Challenge	Currently
Complexity with integrating inferencing into existing HPC-based scientific and engineering workloads	47.0%
Concerns with cost of inferencing-specific hardware or software	31.0%
Concerns with technical issues surrounding inferencing such as expandability and hallucinations	29.0%
High/uncertain operational costs	24.0%
Uncertainty about the right application or hardware or software to use	24.0%
High/uncertain development costs	23.0%
Too computationally intensive	22.0%
Lack of in-house expertise in inferencing	16.0%
The technology is moving too fast for credible assessment of value	16.0%
Long/uncertain implementation times	15.0%
Lack of demonstrated return on investment	12.0%
Lack of reproducibility	11.0%
Lack of precision	10.0%
Confusion/uncertainty with inference vendor selection	8.0%
Uncertainty of demonstrated computational performance improvements	7.0%
Other	5.0%

N=100

Source: Hyperion Research, 2025

5-Year Anticipated % of Overall Advanced Computing Budget for AI Inferencing

Confidence in considerable efficiency and productivity gains remains very high



N=100
Source: Hyperion Research, 2025

- **The inference spending portion is expected to rise, with some outliers expecting a decrease**
- **Additional outliers expecting to reach the >75% threshold**

110137097

ROI Study Highlighted Survey Questions

Next AI study focuses on return on investment, management of challenges, and shifting allocation of resources

- **To what extent did integrating generative AI models into your HPC workload environment meet performance and cost expectations?**
- **How have budgetary plans to support gen-AI change over the last 12-18 months?**
- **If there have been measurable monetary gains from HPC/AI integration, how long will it be to recover from initial investment?**
- **To what degree will your organization expand or contract gen-AI development moving forward?**

Top of Mind:AI Maturity Brings New Questions

As efforts to adopt and integrate AI gain traction among industry leaders, new use cases, optimization, regulatory developments, and ROI will become a new focus for users

- **HPC/AI integrators have come to expect:**
  - Robust return on investment
  - New levels of efficiency
  - Effective regulatory guidelines
- **As AI integrated systems become the norm, the effectiveness and limitations of the technology will become better understood**
- **Aspirant goals will be realized for many users, but some may face costly challenges of unexpected severity such as:**
  - High cost of upkeep
  - Continual education of in-house expertise
  - Rising emphasis on effective oversight
  - Management of regulatory demands

Top of Mind:LLM Training Needs a Reboot

The rapid rise of compute requirements for large language model training runs will begin to slow with a shift in emphasis on smaller and more efficient models using more focused training data sets

- **Current LLM training requirements 10^{26} total training operations**
  - Projections call for an increase of two to three order of magnitude in the next few years (10^{28} to 10^{29})
  - This is out of reach for all but the most aggressive, well-funded organizations: e.g., Anthropic, OpenAI, Tesla, Meta, Google
- **The mainstream HPC world will instead focus on less demanding LLMs or small language model training**
  - Requires less total compute, perhaps three to four orders of magnitude less
  - Based on training data sets that are smaller, more disciplined or subject focused, appropriately curated, and perhaps even proprietary to a targeted end use or end users

HYPERION RESEARCH

Questions, Comments And Suggestions Are Welcome

Please contact me at:
tsorensen@hyperionres.com

Thank You For Joining Us Today!

June 2025

Earl Joseph, Bob Sorensen,
Mark Nossokoff,
Tom Sorensen, and Jaclyn Ludema

www.HyperionResearch.com
www.hpcuserforum.com

HYPERION RESEARCH

We Invite You To Join Our Future HPC User Forum and Hyperion Research Meetings (www.hpcuserforum.com)

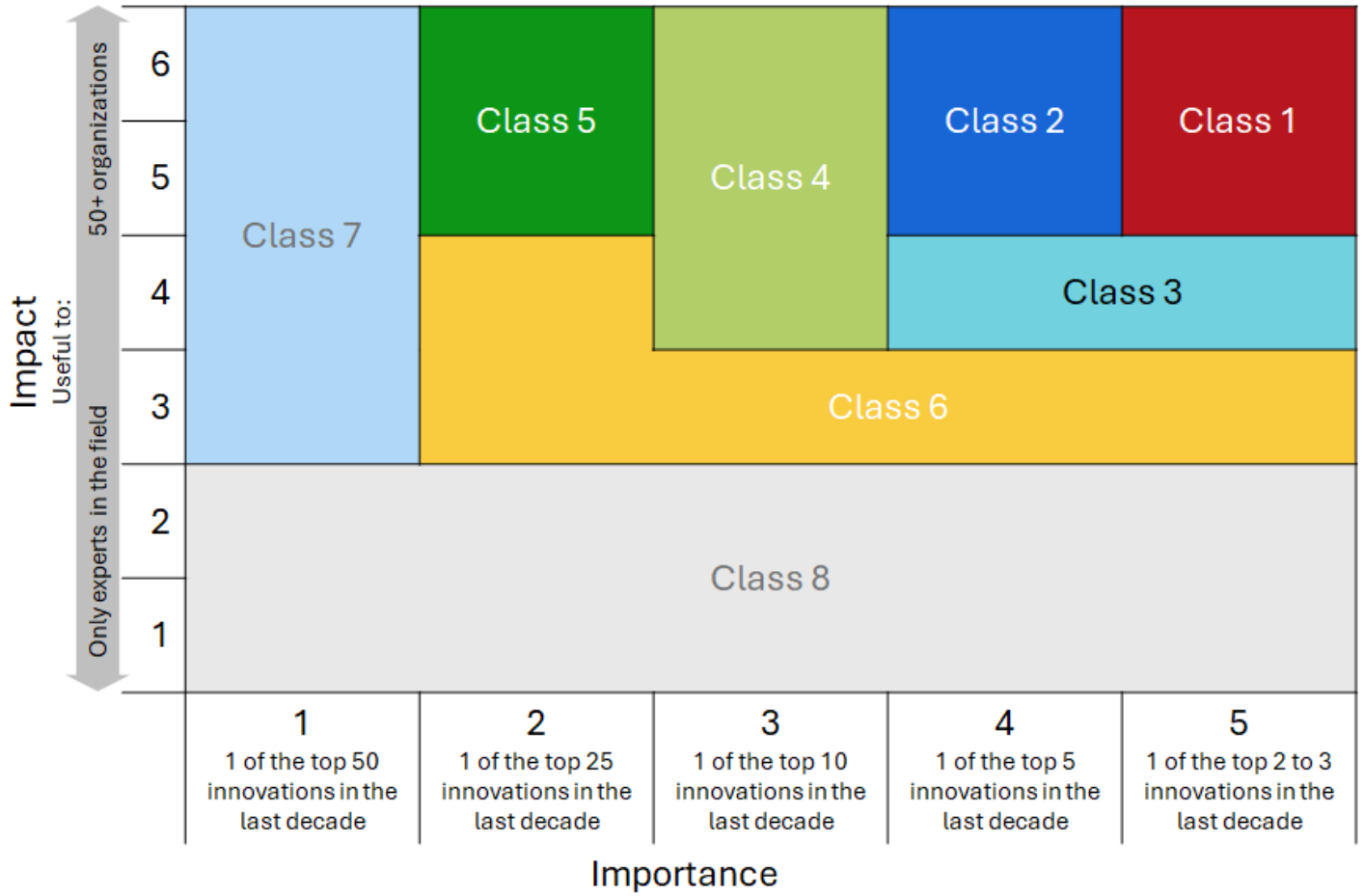
- **June 3 & 4: Pre-ISC Breakfast Briefing**
- **September 3-4: Reston, Virgina**
- **October 7-8: HPC User Forums in Paris**
- **November: SC25 Breakfast Briefing**

HYPERION RESEARCH

A New Way To Measure The Value Of Leadership Computing and R&D Successes

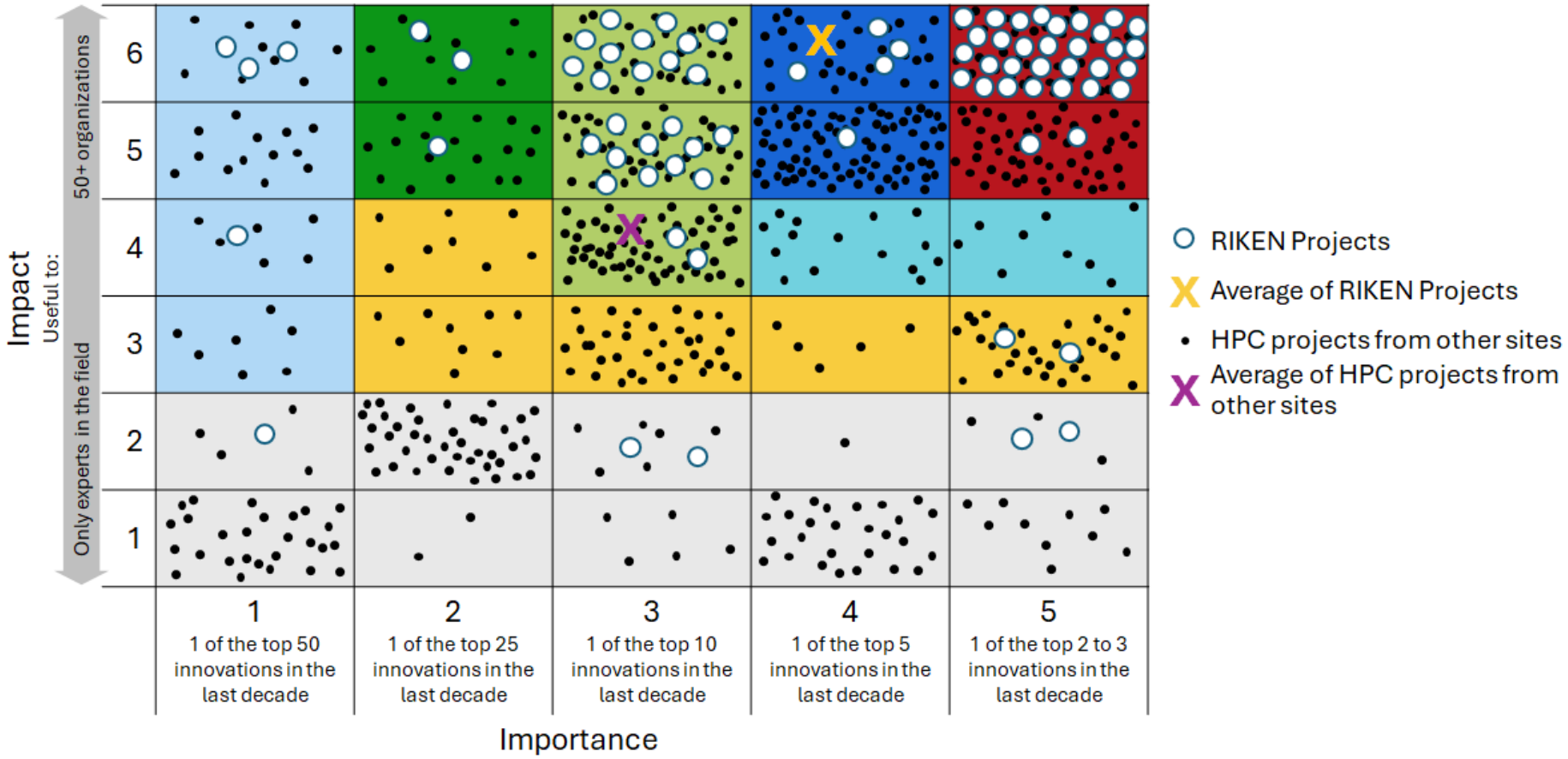
A New Way to Show the Value of Leadership Computing

Using two scales: innovation importance level, and how broadly impactful are the results



Leadership Computing at RIKEN

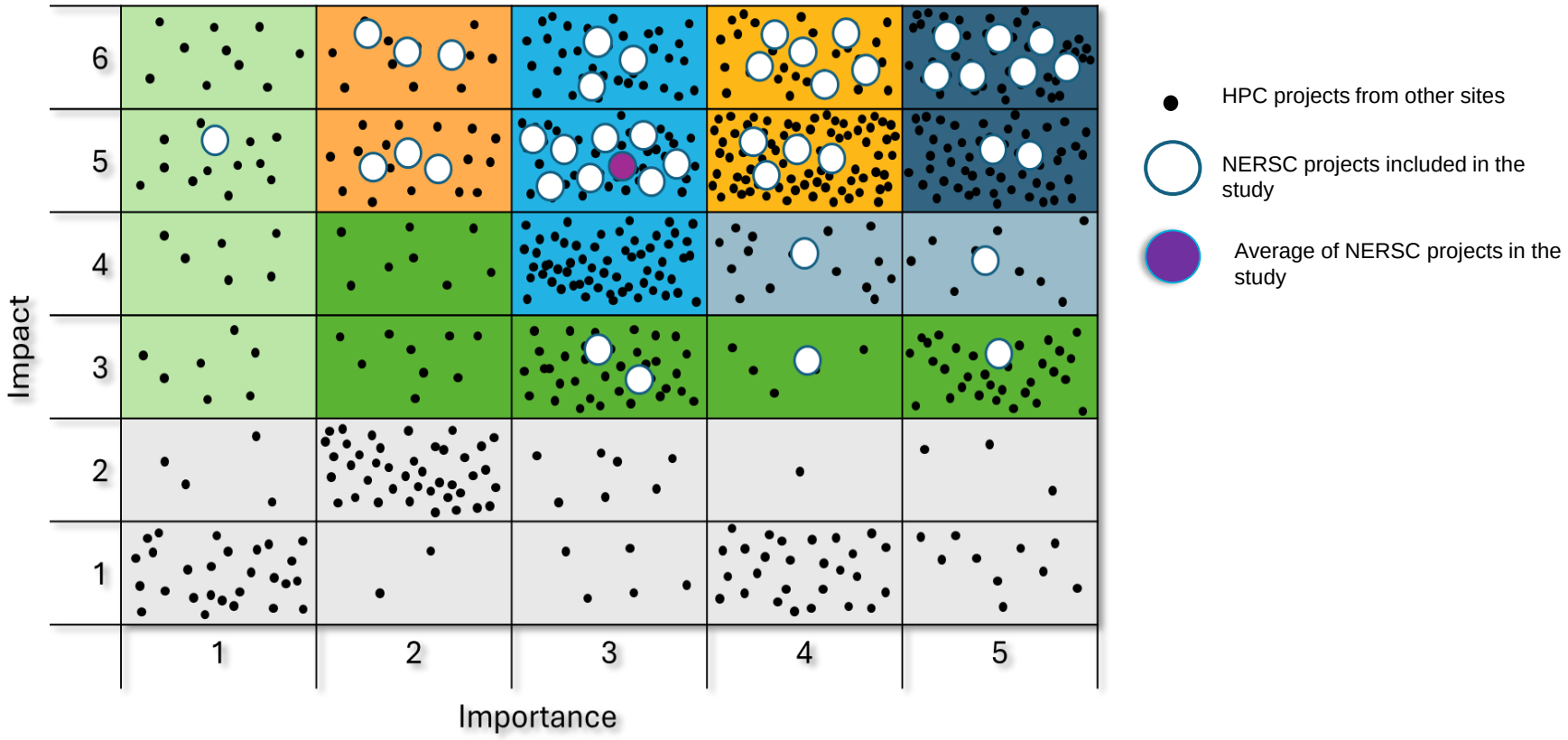
An example from a 2024 study compared to 650 other projects



Leadership Computing at NERSC

An example from a 2024 study compared to 650 other projects

Innovation Class Mapping: Showing Participating NERSC projects



HYPERION RESEARCH

We Welcome Questions, Comments and Suggestions

Please contact us at:

info@hyperionres.com