Hyperion Research
SC25 Advanced Computing Update

November 2025

www.HyperionResearch.com
www.hpcuserforum.com

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

Welcome

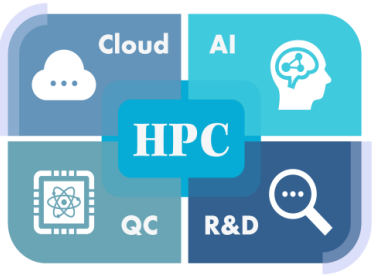
- **Today's Presentation Will Be Sent to All Registrants**
- **A Video Recording Will Be Posted at**
  - www.cpcworldwide.com/HyperionSC25
- **For Follow-Up Details or Discussions, Please email mthorp@hyperionres.com**
- **The Briefing Will Conclude by 8:20**

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:

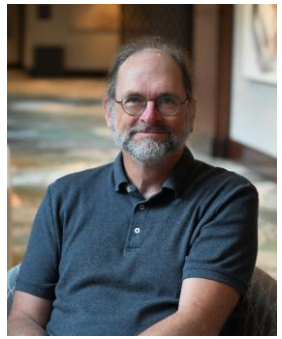
- *To improve the health of the HPC/AI/QC industry*
  - *Through open discussions, information sharing and initiatives involving HPC users in industry, government and academia along with HPC vendors and other interested parties*

Hyperion Research Team

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

Earl Joseph	Jean Sorensen	Bob Sorensen	Mark Nossokoff	Jaclyn Ludema	Thomas Sorensen	Mike Thorp
• CEO Hyperion Research • Executive Director HPC/AI User Forum	• Chief Operating Officer • HR Director	• Senior Vice President of Research • Chief Quantum Computing Analyst • Lead AI Analyst	• Research Director • Chief Storage Analyst • Lead Cloud Analyst	• Lead Sustainability Analyst • Cloud Analyst	• Principal Analyst AI/HPC	• Senior Global Account Executive



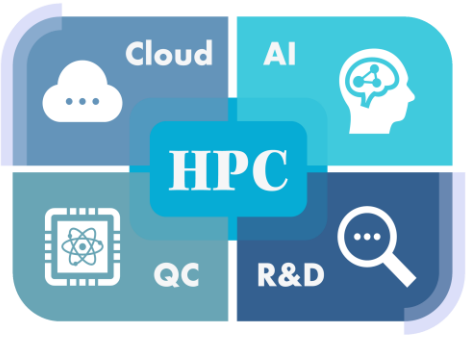




Hyperion Research Partners

KC Associates

Example Research Areas



- Traditional HPC
- All types of AI for technical computing
- Cloud Computing
- Storage & Data
- Interconnects
- Software & Applications
- ROI and Scientific Returns from HPC
- Tracking all Processor Types & Growth rates
- Quantum Computing
- R&D and Engineering -- all types
- Supply Chain Issues
- Sustainability, Power & Cooling

Today's Agenda

- **Mike Thorp, Senior Global Account Executive**
  - Introduction
- **Earl Joseph, CEO**
  - HPC and AI Market Update
- **Bob Sorensen, SVP, Chief QC & AI Analyst**
  - QC, AI, ToM
- **Mark Nossokoff, Research Director, Chief Storage & Cloud Analyst**
  - Perspective on HPC-AI Storage and Interconnects
  - HPC-AI Cloud
- **Innovation Awards Announcement**
- **Conclusions**

HPC/AI Market Update

Major Trends and Market Activities

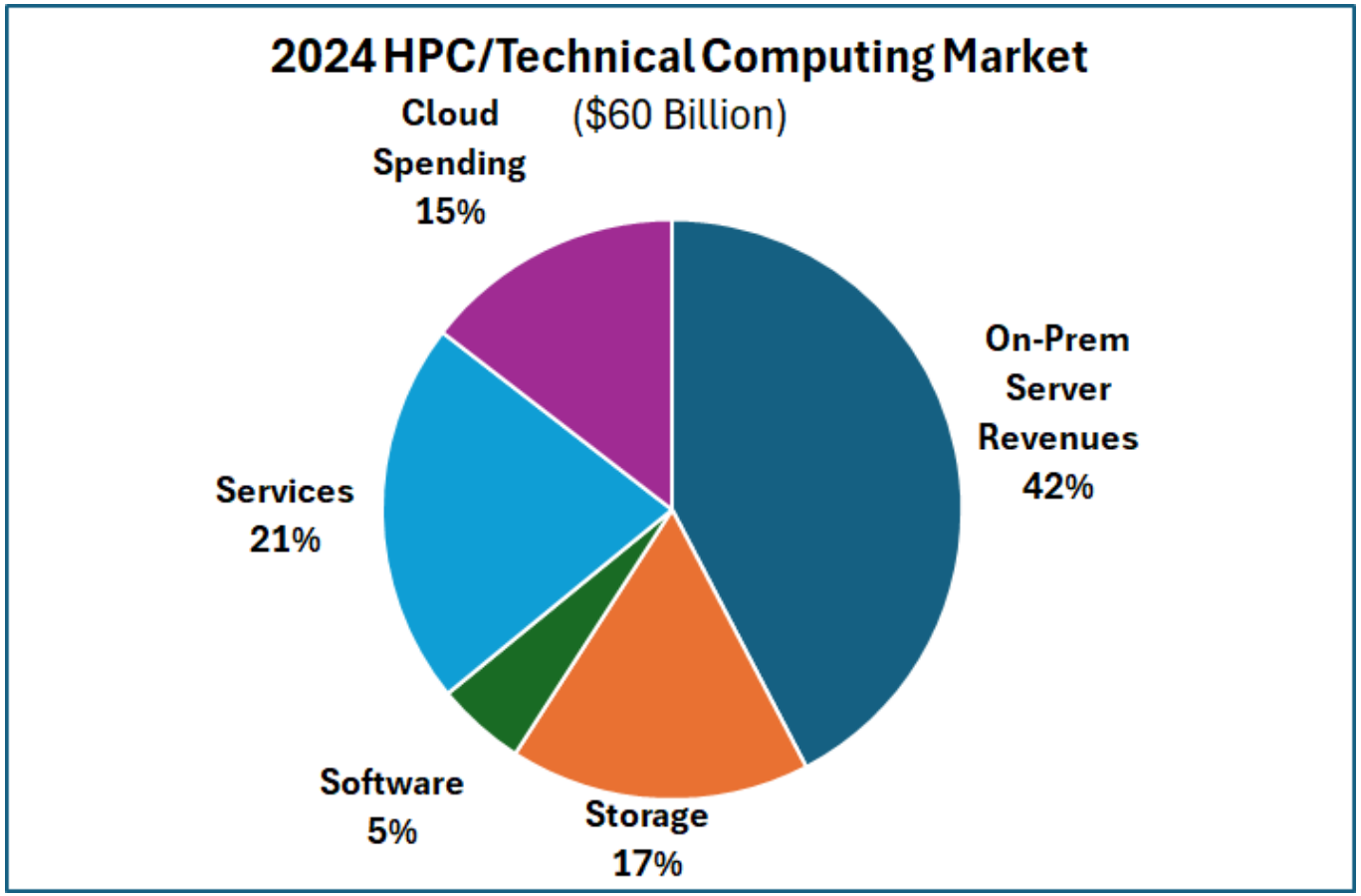
HPC has entered a new high growth mode

- **Now growing at over 20%, compared to the historic 7% to 8%**
- **New technologies and new use cases are expanding quickly**
  - AI and QC systems are being installed at a high rate
- **Governments around the world are making major AI investments – well above anything in the past**
  - AI factories & giga factories in Europe, new AI sites in the US, Japan's AI initiative, …
  - Data centers are now being measured in Gigawatts – makes 25 or 50 MW look small
  - Power requirements are growing at a major rate
  - Leads to some difficult questions:
    - Will investments continue to grow at these rates?
    - When will ROI become important?
    - Is AI at scale safe? How can we make it safe?
- **Cloud computing is being used at most HPC sites**
  - The questions are now when, where and how much – as opposed to if they should be used

2024 Was a Strong Growth Year

The highest growth in over two decades (23.5%)!
The first half of 2026 grew by another 22%

- **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

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

Strong growth in systems that sell for over $1 million USD

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

Five sectors are now over $2 billion USD a year

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

The HPC/AI Market Should See Strong Growth in 2025

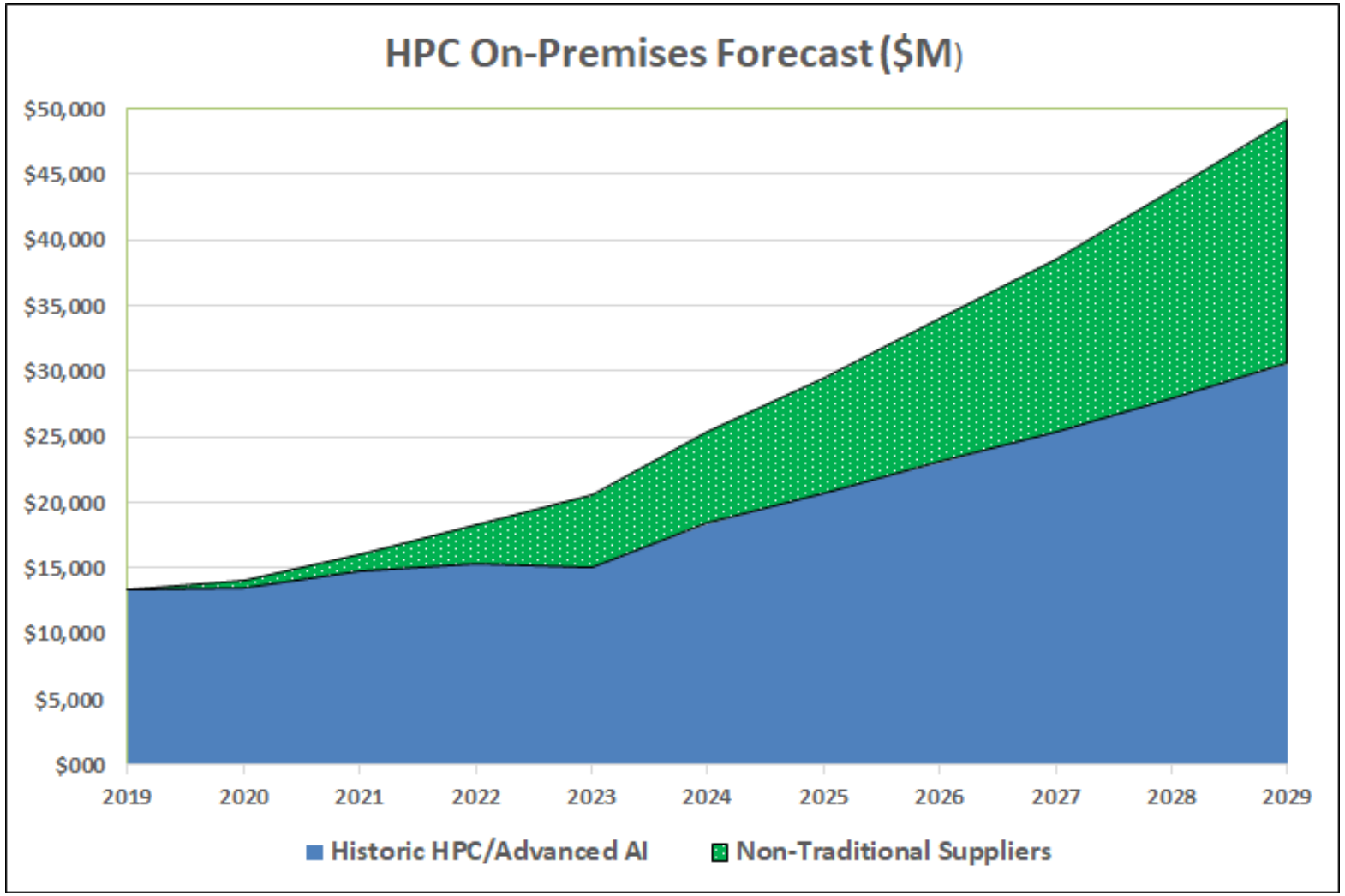
… but there are some major concerns

- **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
  - Clouds are being more broadly used

Updated HPC On-Prem Server Forecast

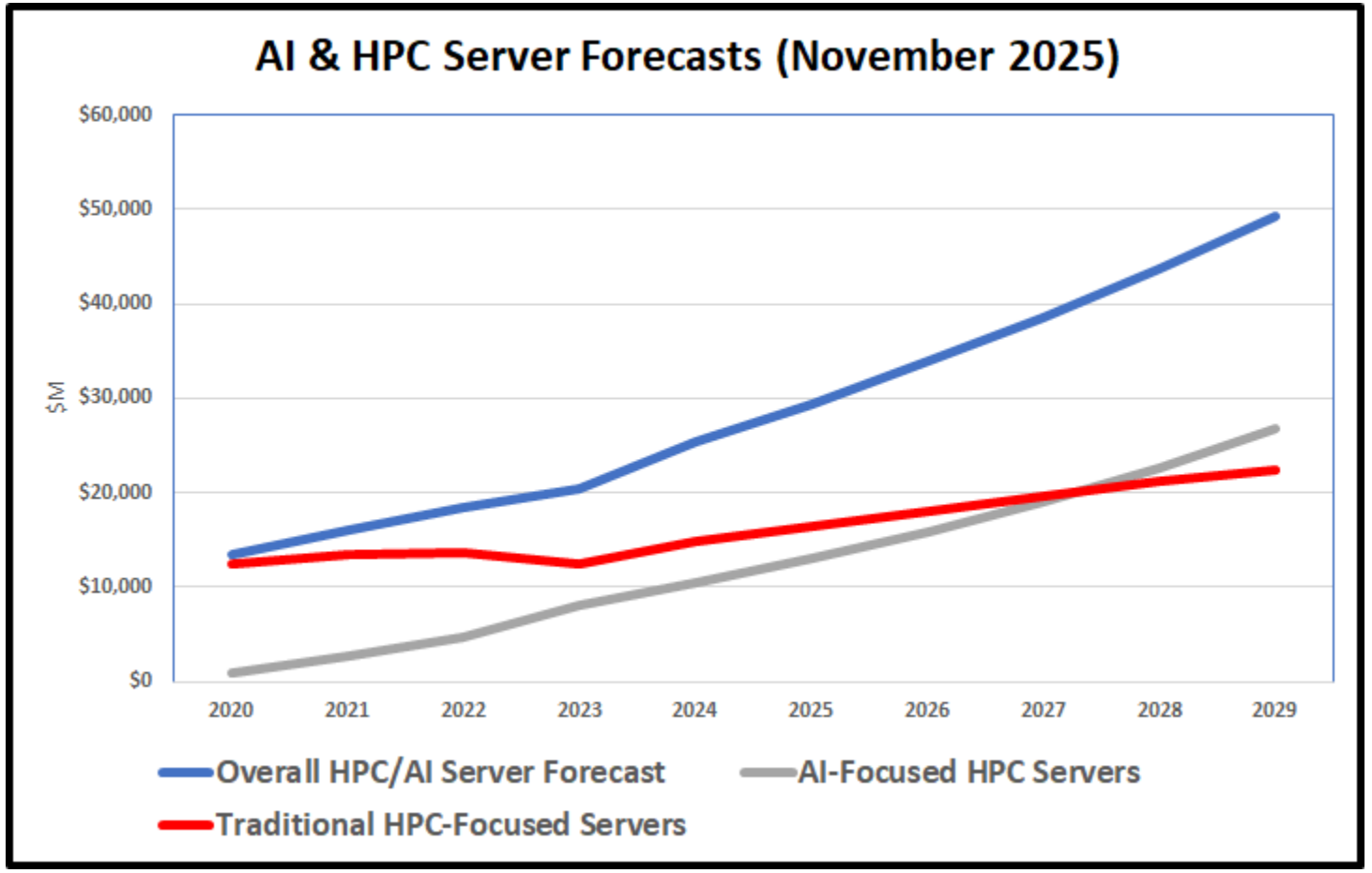
The on-prem HPC server market is projected to reach ~$50 billion by 2029



HPC Compared to AI-centric Servers

Many servers are running both traditional HPC and AI workloads

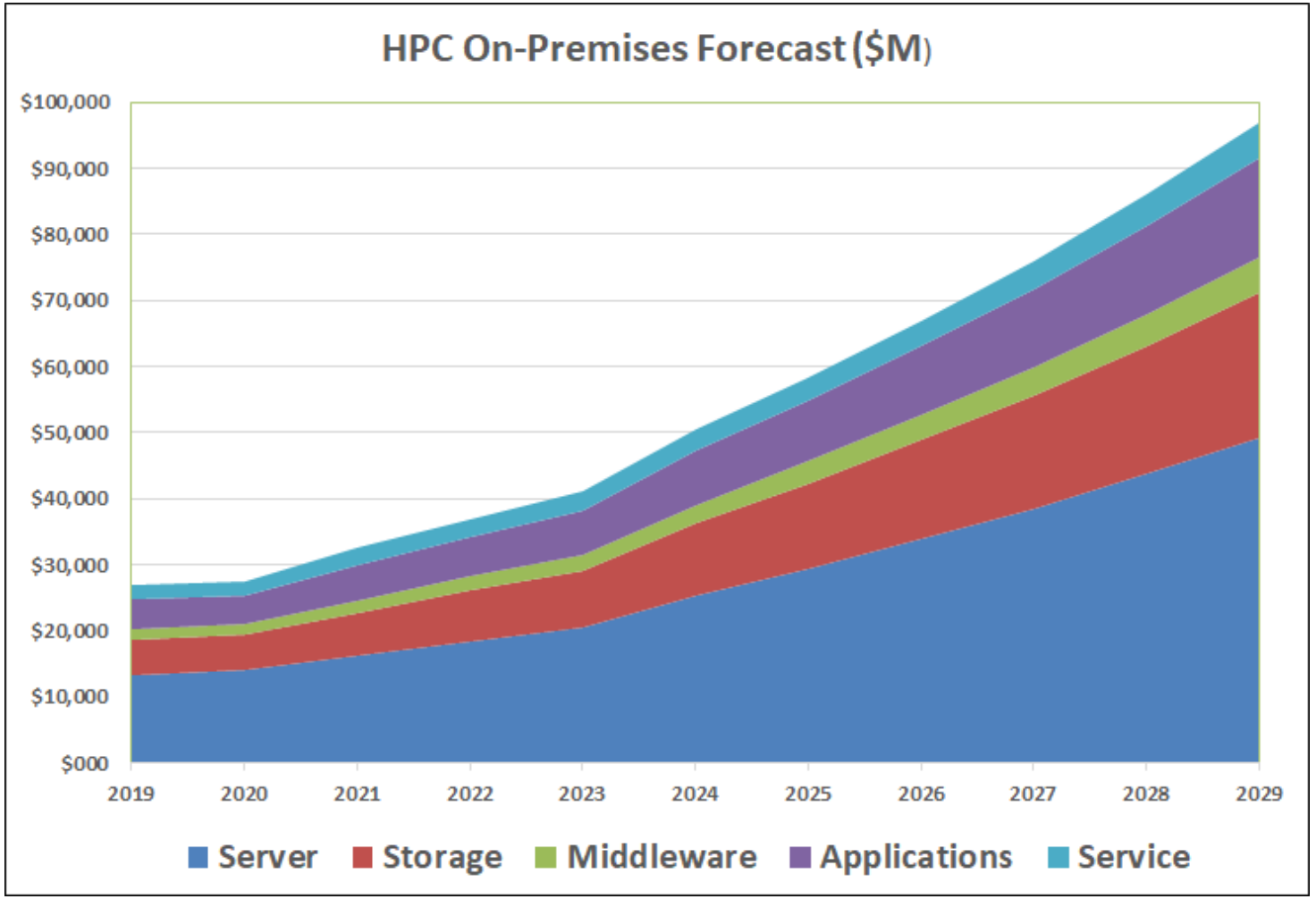
- **AI systems often run some traditional HPC jobs (<50% of workload)**
- **Likewise, traditional HPC systems often run some AI jobs (<50% of workload)**



Updated HPC On-Prem Broader Market Forecast

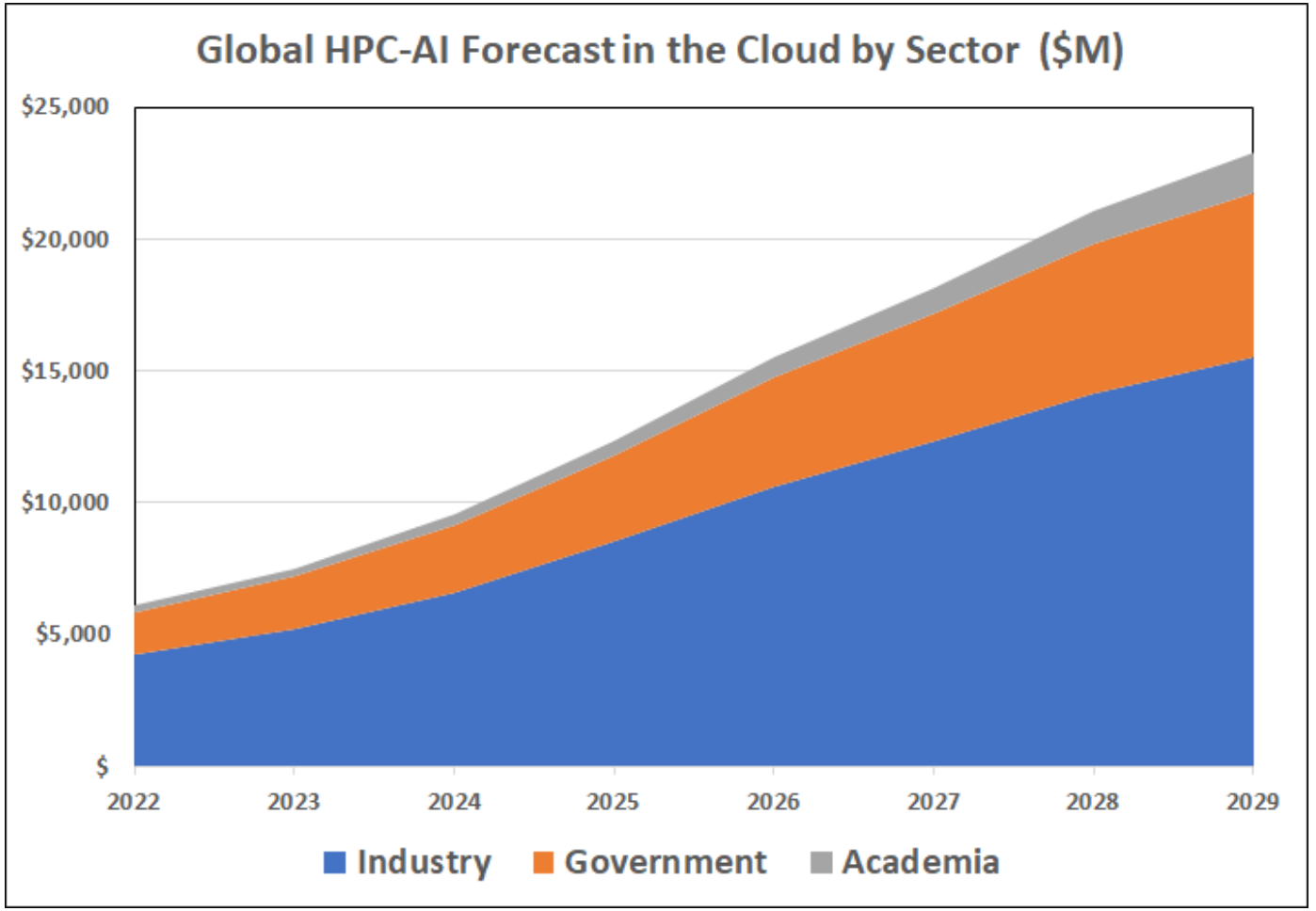
The on-prem HPC broader market (servers, software, storage & maintenance) is projected to reach close to $100 billion by 2029

Almost doubling over the next 5 years



HPC Cloud Forecast

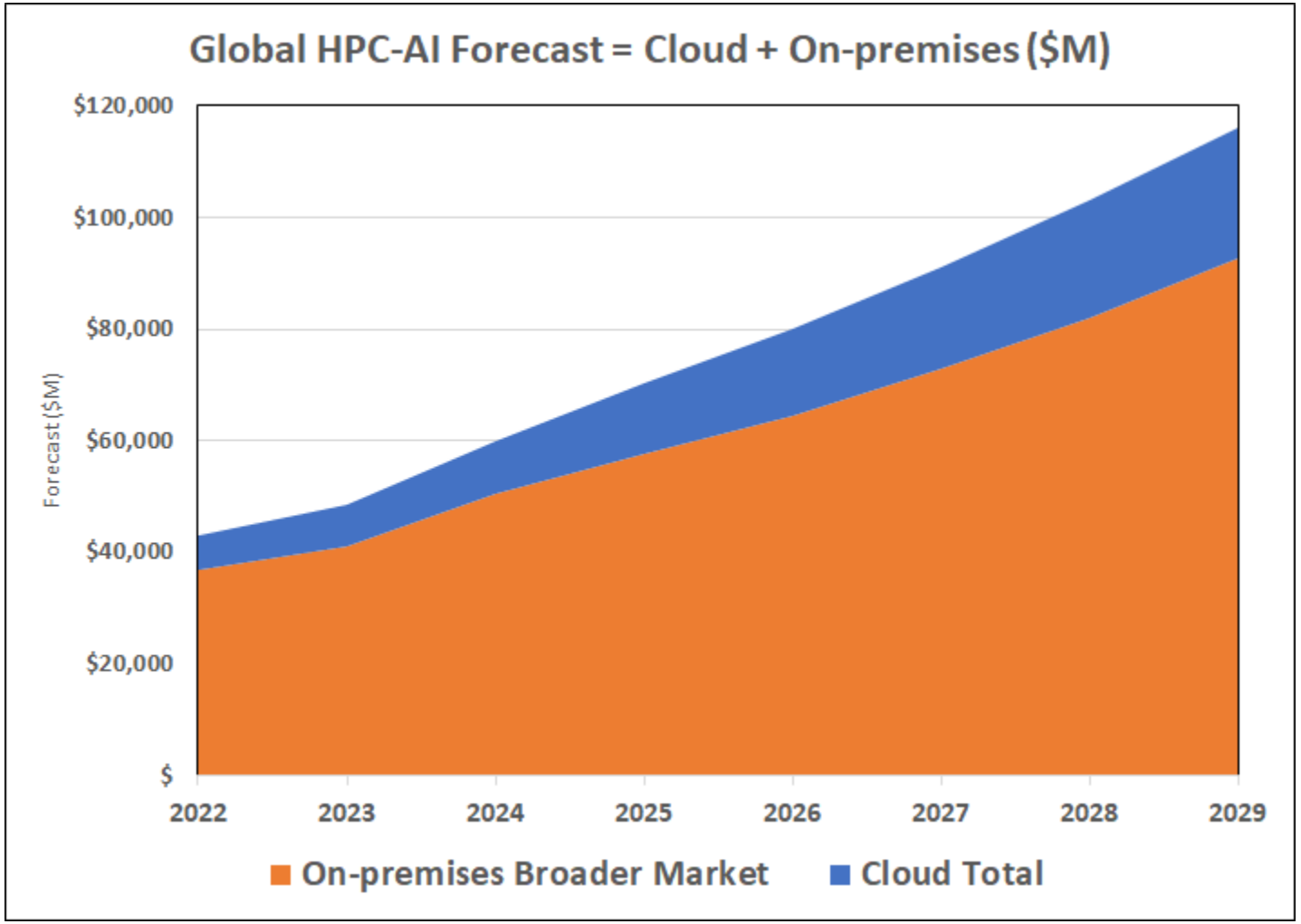
HPC cloud spending is projected to reach ~$23 billion by 2029
~2.5X increase over the next five years (2024 to 2029)



Note: This forecast hasn't yet fully incorporated the new DOE AI systems.

HPC On-Prem Plus Cloud Forecast

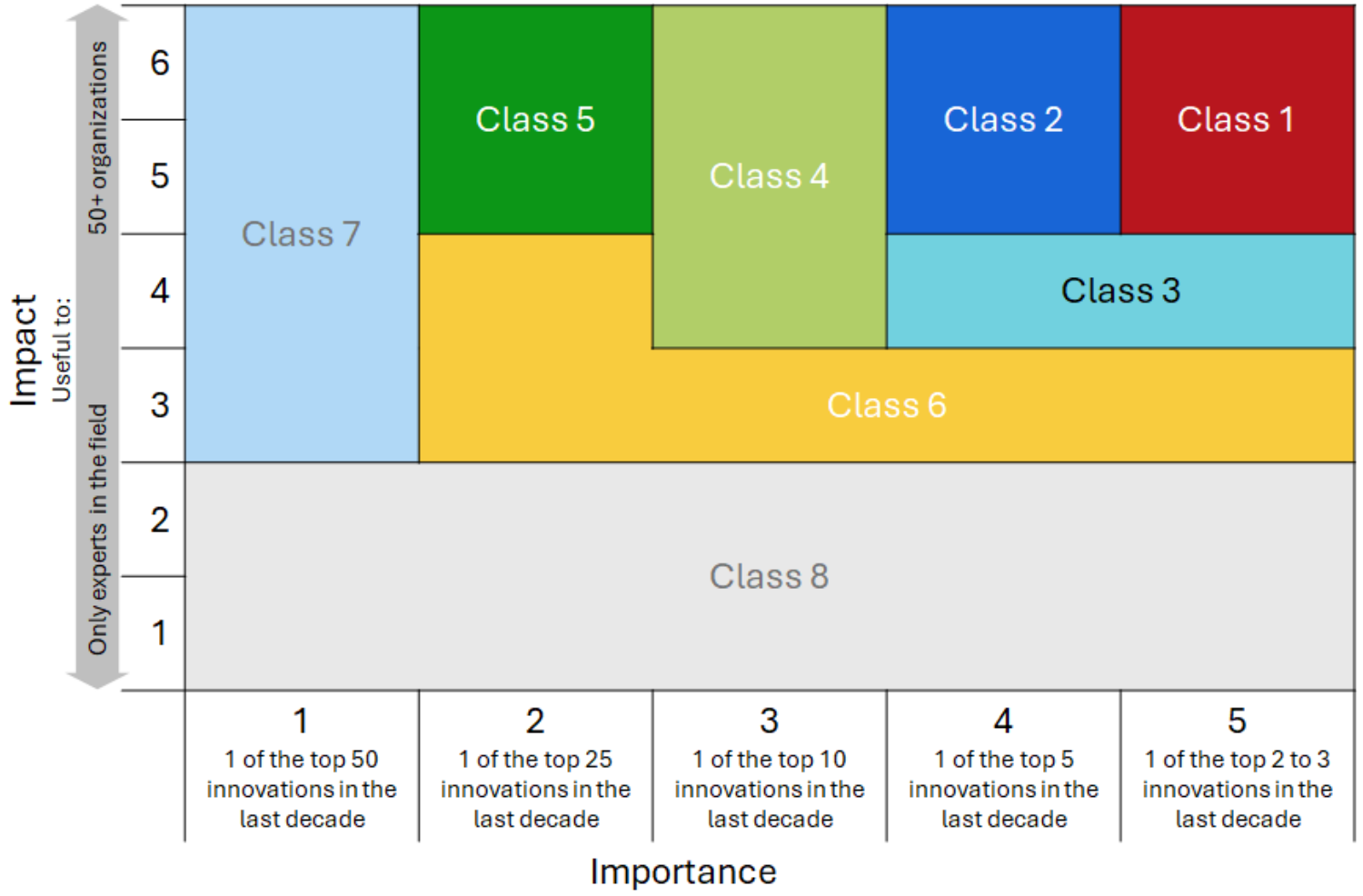
The overall HPC market is projected to reach close to $120 billion by 2029
Doubling over the next 5 years



Measuring and Comparing Leadership Computing

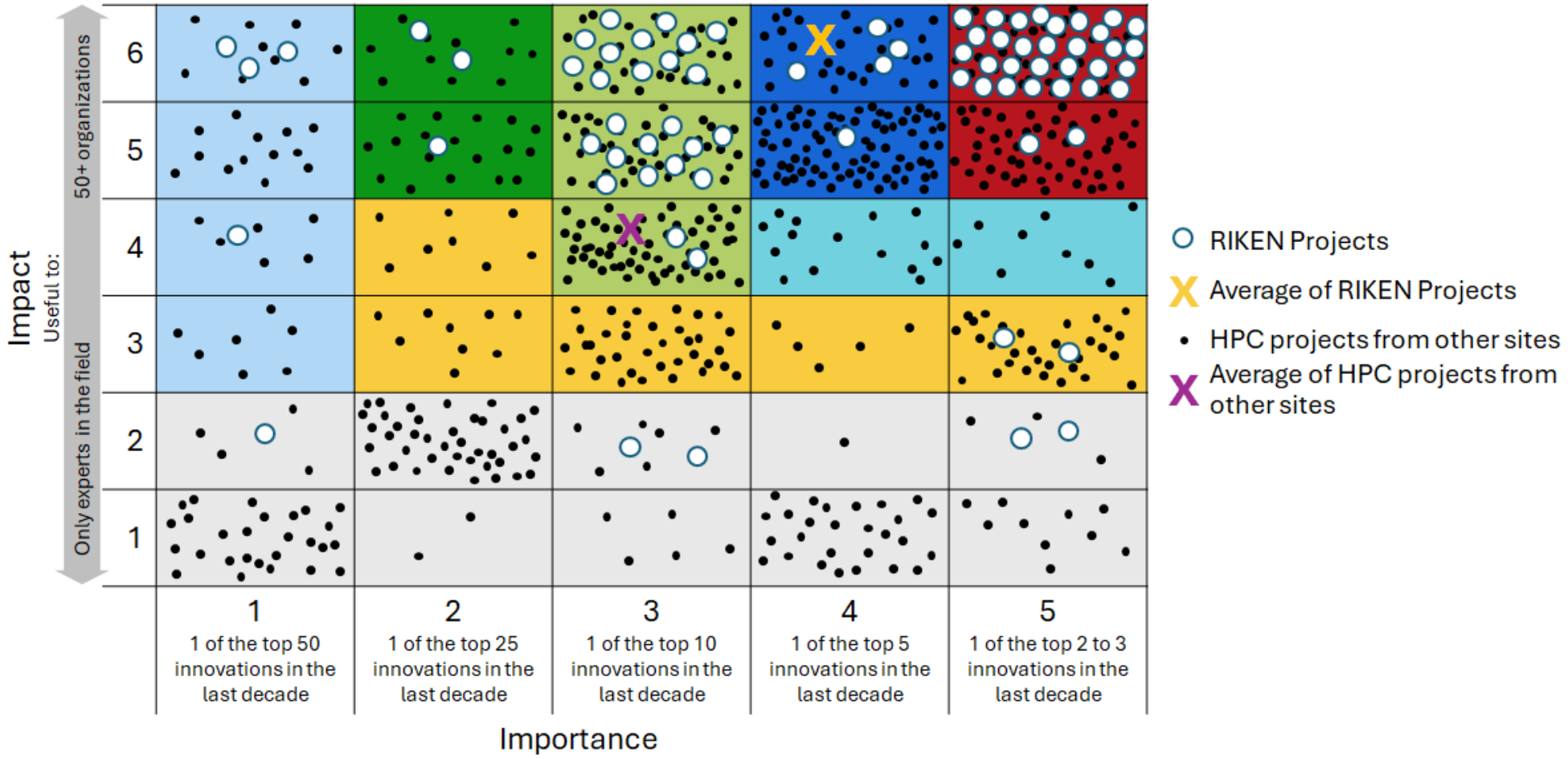
A New Way to Show the Value of Leadership Computing

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



Showing the Value of Leadership Computing: RIKEN

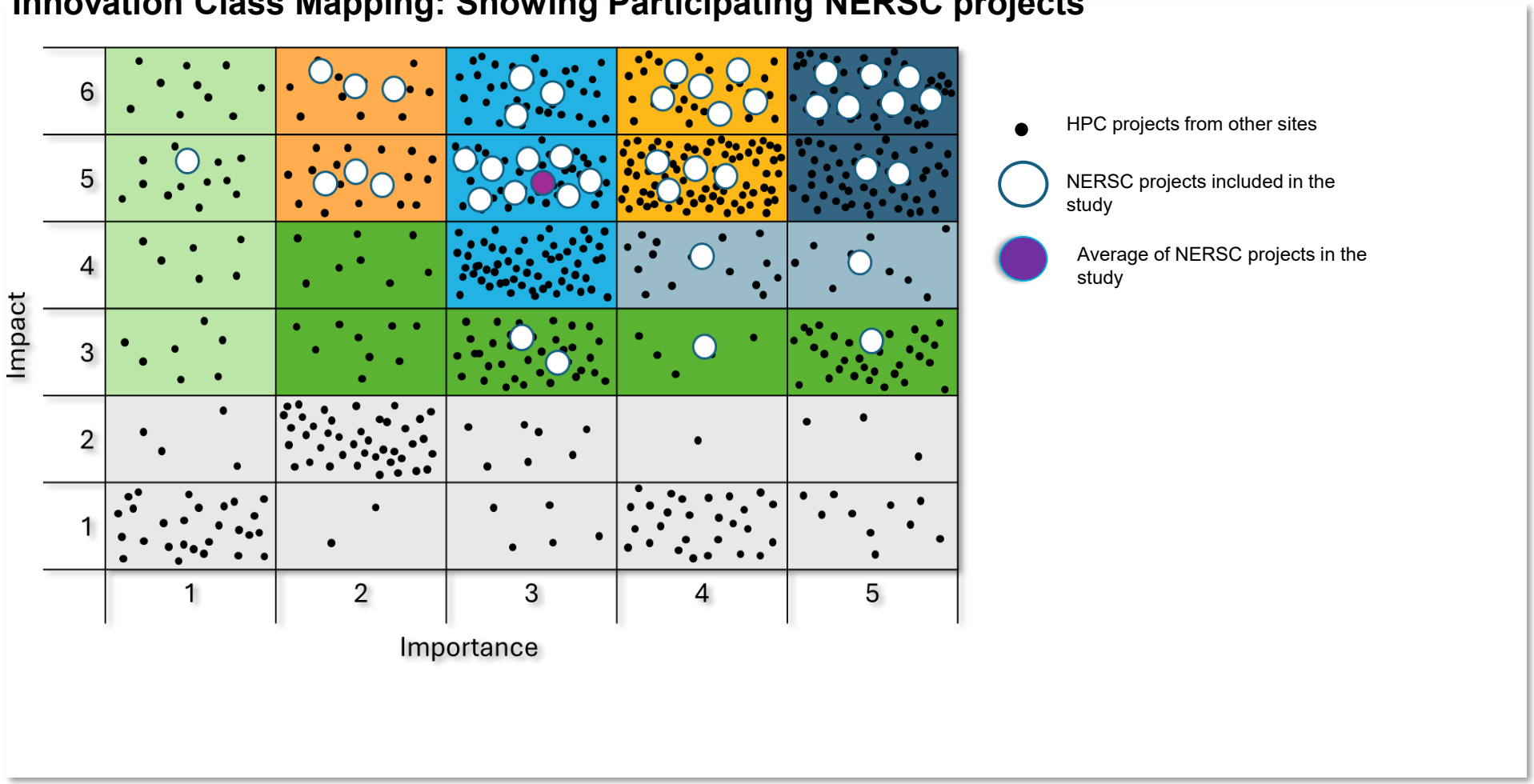
An example from a 2024 study compared to 650 other projects



Showing the Value of Leadership Computing: NERSC

An example from a 2024 study compared to 650 other projects

Innovation Class Mapping: Showing Participating NERSC projects



QC/AI/ToM

QC Market Executive Summary/Highlights

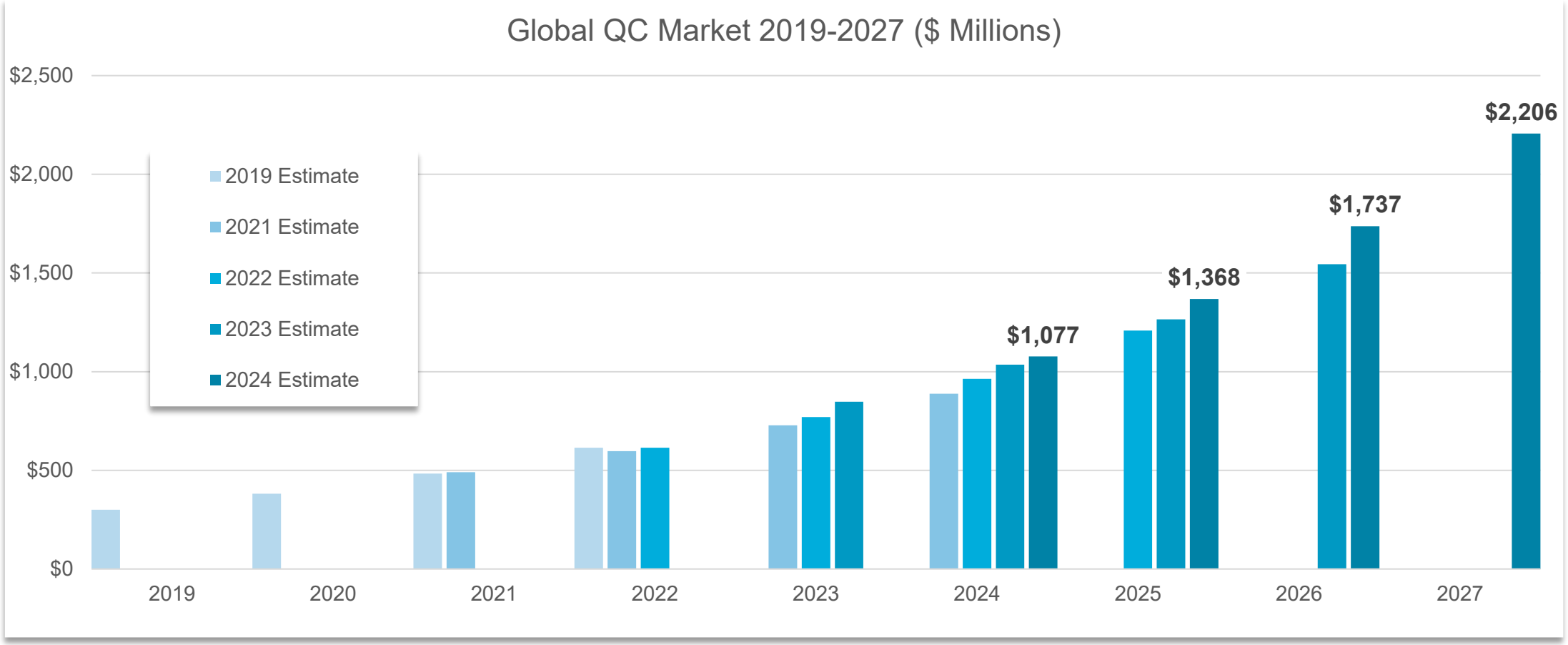
- The global quantum computing (QC) market is estimated to have been worth $1.07 billion in 2024, with a projected average annual growth rate of 27%, driving the global market to $2.2 billion on 2027
- Based on a survey of 115 respondents representing 82 QC companies:
  - 41% estimate their organization's revenues will increase by more than 25% in 2025, up from 36.6% in 2024
  - 18% are looking to 2025 gains to be on the order of 11-25%, almost double last year's percentage of 9.8%
  - The number of firms that see flat or nearly the same revenues dropped from 25.6% in 2024 to 9.8% in 2025
- Partnerships have become a fundamental activity within QC supplier sector
  - 83% had partnerships with other QC supplier(s)
  - 74% had partnerships with at least one government research organization
  - 71% had partnerships with at least one QC end user
- Anticipated cloud and on-premises revenues will be nearing parity in 2027
  - Total on-prem activities projected to account for 46% of QC market in 2027
  - Up from 31% in last year's market study
- Modeling/simulation anticipated to remain the #1 algorithm by revenue in 2027
  - Optimization and AI remain major algorithms

QC Market Executive Summary/Highlights (cont.)

- Aspiring QC end users are looking for new algorithms and ways to address concerns with future classical performance
  - But many are still exploring for the sake of exploration
  - One in four are looking at real-time compute opportunities
  - Interest in lowering total compute costs is gaining traction
    - 2023 Survey: 9.0%
    - 2025 Survey: 23.5%
  - Not so with reducing power/cooling costs
    - 2023 Survey: 17.3%
    - 2025 Survey: 14.8%
- About half of the respondents (52%) expect the availability of utility class QC in the next 2-5 years
  - About one in three say five years or more
  - About 8% say they are already here or will be in the next year
- Almost half of the respondents see a chance for a quantum winter
  - Significant jump from last year in ‘very high’ from 14% to 24%
  - Combined with drop in ‘somewhat unlikely’ from 33% to 25%
- LLMs – and likely generative AI in general – are seen as near-term competitor for QC end user interest by 47% of respondents, up from 42% last year

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 Projection Considerations

Key factors contributing to growing QC market complexity

- Increased interest/orders/installations for on-premises systems will drive significant revenues for some
  - D-Wave FY 2024 bookings exceeded $23 million, a 120% increase over FY 2023, due in large part to a single system purchase by Davidson Technologies
  - QuEra was awarded a $42 million contract in 2024 by Japan's AIST to deliver an on-premises QC
- Losses generally, and sometimes significantly, outweigh revenues
  - Company A posted revenues of $10.8 million with a net loss of $201.0 million for 2024
  - Company B recognized revenue of $43.1 million with a net loss of $331.6 million for 2024
- Most QC companies are private, and most intend to stay that way for at least the next few years
  - Limited financial reporting requirements
  - The bulk of funding inputs are not revenue based but instead come from VC or government sources

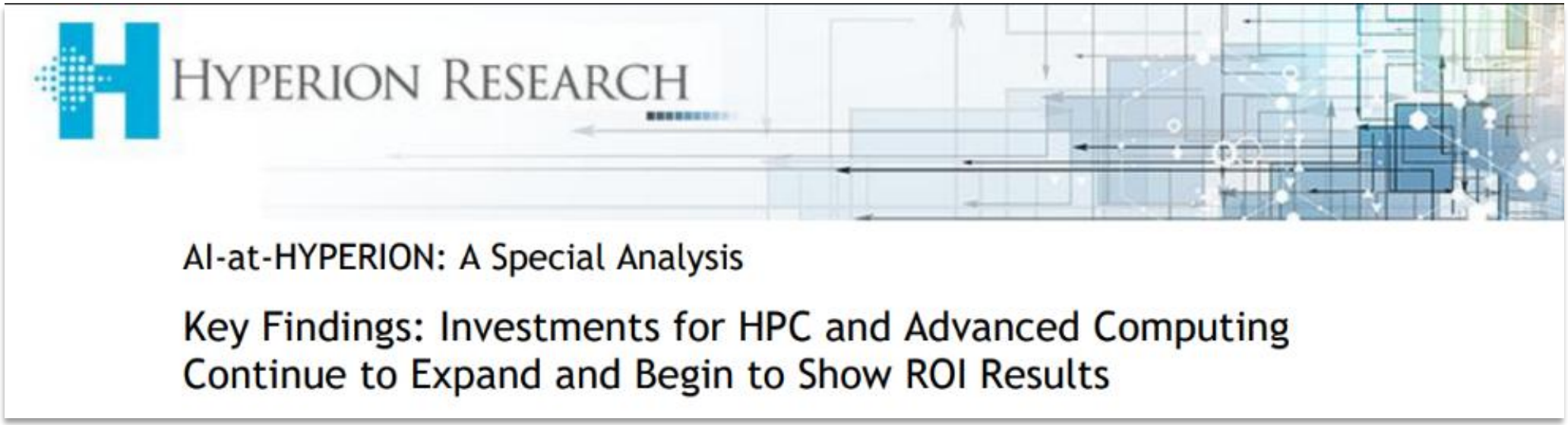
QC Market Projection Considerations (cont.)

External funding numbers are growing substantially and will likely impact market winners and losers

- **PsiQuantum**: Raised $1 billion in September 2025 in a Series E round led by BlackRock,
  - Includes participation from Nvidia's venture arm and significant backing from the Australian government.
  - Nvidia recently took part in funding rounds for the Quantinuum, PsiQuantum, and QuEra
- **Quantinuum:** Secured $600 million in new funding in September 2025, giving the company a $10 billion valuation
- **QuEra Computing**: Raised over $230 million in a financing round in February 2025, with investors including Google's Quantum AI business unit and SoftBank Vision Fund
- **IQM Quantum Computers:** The Finnish company raised a €200 million (approximately $220 million) Series B funding round in April 2025
- **Classiq** secured $110 million in Series C funding in May 2025
- **Alice & Bob** closed a $104 million Series B round in January 2025

Is there going be an M&A outbreak, a culling of the herd, or at least a slowing of new entrants?
More QC hardware suppliers right know than the sum of all HPC vendors spanning the last 60 years

HYPERION RESEARCH



Tom Sorensen, Principal Analyst for AI/HPC

HPC-related Gen-AI End Use

End Use	% Selected
Scientific data analysis	80.6%
Time series data analysis	61.2%
Text generation	60.2%
Code generation	56.3%
Synthetic data generation	54.4%
Image creation	43.7%
Audio or music generation	14.6%
Other (please specify)	3.9%
Don't know/Not Sure	0.0%

Notes: N = 103. Respondents could select multiple options.
Source: Hyperion Research, 2025

- Key Finding: Data Analysis (scientific and time series) is most prevalent end use, text generation (clear and code) follows

Fulfillment of Expectations

Extent to Which Integrated Generative AI Models Met Performance Expectations Over the Last 12-18 Months	% Selected
Greatly exceeded expectations	10.7%
Moderately exceeded expectations	21.4%
Met expectations	43.7%
Somewhat failed to meet expectations	15.5%
Significantly failed to meet expectations	1.9%
Have not yet integrated generative AI into our HPC workloads	4.9%
Don’t know/ Not sure	1.9%

N=103
Source: Hyperion Research, 2025

- Key Finding: Most users reported LLM integration into HPC/advanced computing workloads met or exceeded expectation

Plans to Expand Gen-AI Use

Plans to Move Forward or Expand Gen-AI Development	% Selected
Significantly expand generative AI use to support HPC/advanced computing workloads	28.2%
Moderately expand generative AI use to support HPC/advanced computing workloads	46.6%
Continue at the same level of generative AI use	18.5%
Moderately contract or decrease use of generative AI for HPC/advanced computing workloads	2.9%
Significantly contract or decrease use of generative AI for HPC/advanced computing workloads	0.0%
Entirely discontinue use of generative AI for HPC/advanced computing workloads	0.0%
Don't know/Not sure	3.9%

N=103
Source: Hyperion Research, 2025

Key Finding: Nearly 75% of respondents plan to moderately or significantly expand generative AI use to support HPC workloads

Expected Times to ROI

Anticipated Demonstrable Measurable Monetary Return On Investment by Year	% Selected
Less than one year	29.1%
One year to less than two years	21.4%
Two years to less than three years	20.8%
Three years to less than four years	7.6%
Four years to less than five years	4.2%
More than five years	4.1%
Never	7.8%
Don't know/ Not sure	5.0%

N=103
Source: Hyperion Research, 2025

- Key Finding: Roughly half of respondents expect a measurable monetary return on investment within two years, with 30% expecting it in less than one year

Cost Expectations and Realities

Extent to Which Integrated Gen AI Models Met Cost Expectations Over the Last 12-18 Months	% Selected
Significantly more cost than expected	11.7%
Moderately more cost than expected	40.8%
Met expectations	34.0%
Somewhat less costly than expected	4.9%
Significantly less costly than expected	1.9%
Have not yet integrated generative AI into our HPC workloads	3.9%
Don’t know/ Not sure	2.9%

N=103
Source: Hyperion Research, 2025

- Key Finding: Most users reported their AI integration exceeded cost expectations at least moderately, but they intended to continue investing in the technology

Challenges and Roadblocks

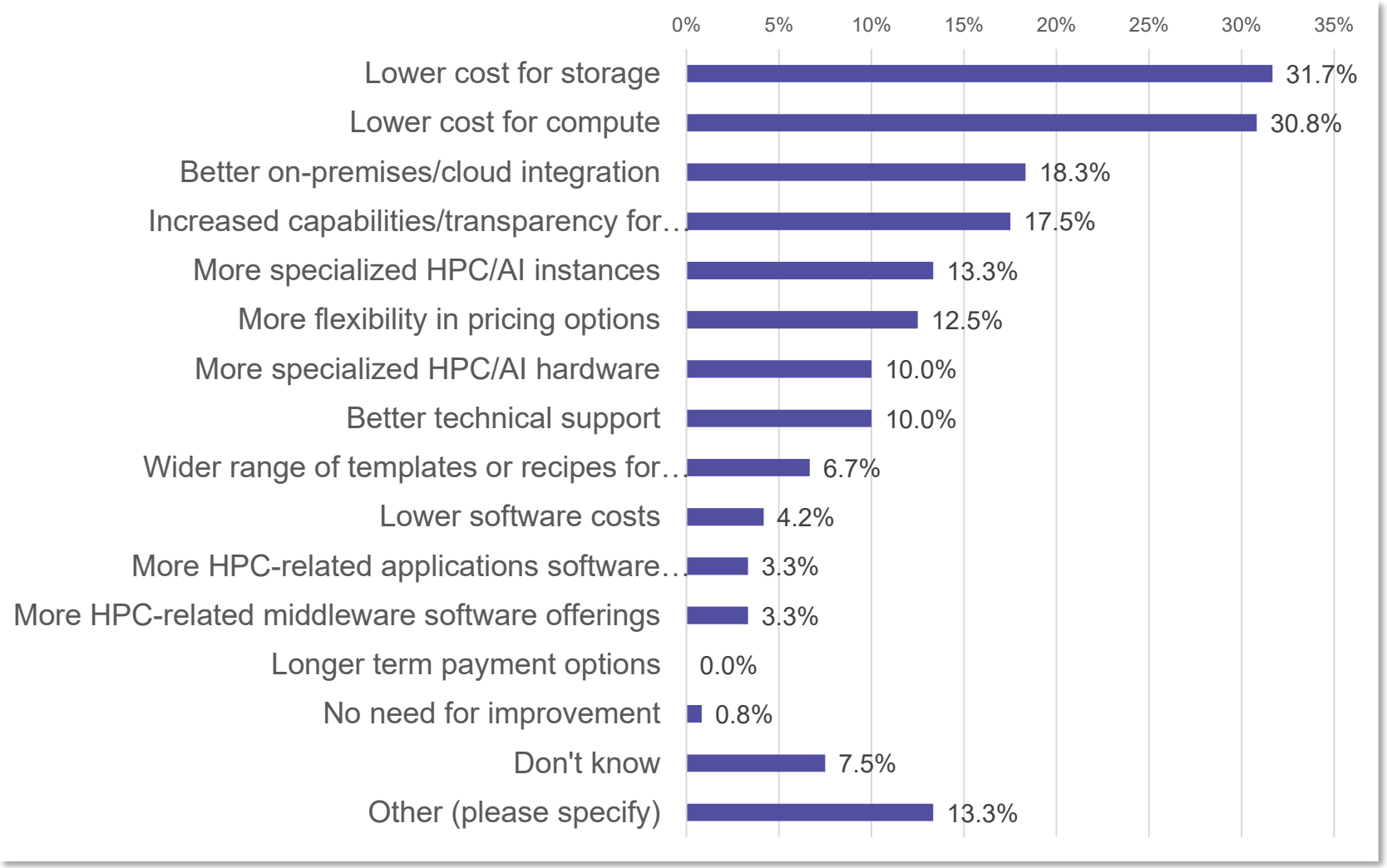
Key Challenges Faced in Integrated AI-Based Models into HPC/Advanced Computing Environment	% Selected
Complexity with integrating generative AI-based models into existing HPC-based scientific and engineering workloads	51.5%
Concerns with technical issues surrounding generative AI-models such as explainability or hallucinations	31.1%
High generative AI computational demands	29.1%
Lack of generative data precision	23.3%
Lack of in-house generative AI expertise	22.3%
Technology is moving too fast for credible assessment of value	17.5%
Lack of required hardware onsite	16.5%
Lack of return on generative AI investment	16.5%
High/uncertain generative AI development costs	14.6%
Confusion/uncertainty with generative AI hardware vendor selection	11.7%
Confusion/uncertainty with generative AI software vendor selection	11.7%
High/uncertain generative AI operational costs	10.7%
Uncertainty of demonstrated computational performance improvements	9.7%
Lack of required hardware in the cloud	6.8%
Lack of required software onsite	6.8%
Long/uncertain generative AI implementation times	5.8%
Lack of required software in the cloud	2.9%
We have not faced any challenges	2.9%

Key Finding: Respondents indicated integration complexity, technical issues, and high computational demands as their top challenges

N=103
Source: Hyperion Research, 2025

Top of Mind

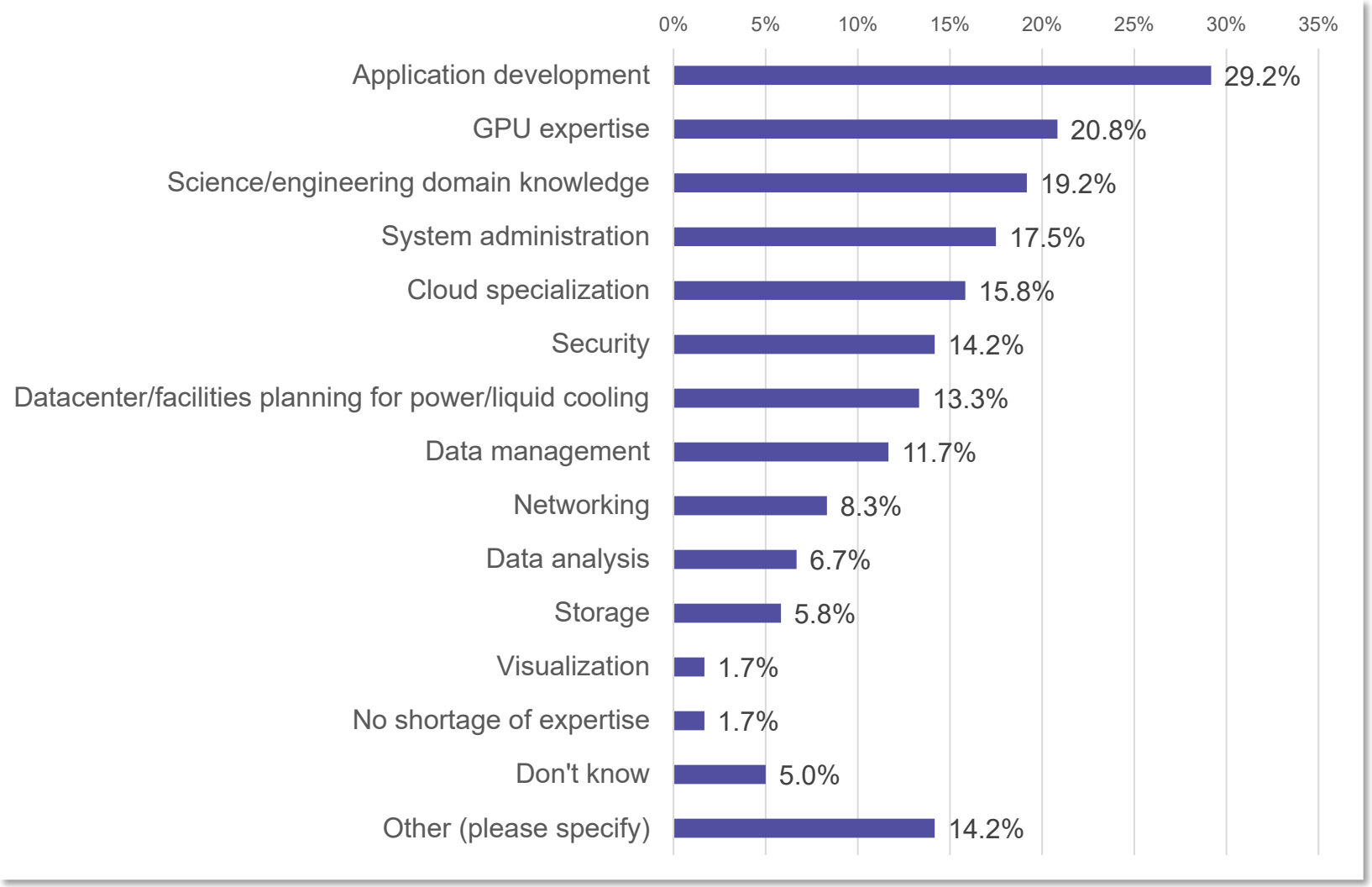
In what areas do the major cloud service providers need the most improvement?



- ~30% mentioned lower cost at least once
  - 15% mentioned lower cost twice
- However, there appears to be no one-size-fits-all solution
- 13.3% of respondents wrote in "other" responses including:
  - Better cost overrun guardrails
  - Faster/easier cloud formation and cross-cloud flexibility
  - More integrated HPC/Secure-Encrypted/Cloud-Storage-to-On-Prem Solutions
  - More transparency on price predictability
  - Address environmental impacts/ways to reduce enormous power demands

N = 120, Source: Hyperion Research, 2025

What areas of HPC have the greatest shortage of expertise?



- These finding align with the persistent challenge of maintaining a HPC talent pool that works in the interdisciplinary spaces
- 14.2% wrote in other responses including:
  - Quantum (x4)
  - AI adoption for scientific applications
- Several respondents used the “other” to express the need for more than two options

N = 120, Source: Hyperion Research, 2025

SC25 Advanced Computing Market Update Briefing
Cloud, Storage, Interconnects, and Sustainability

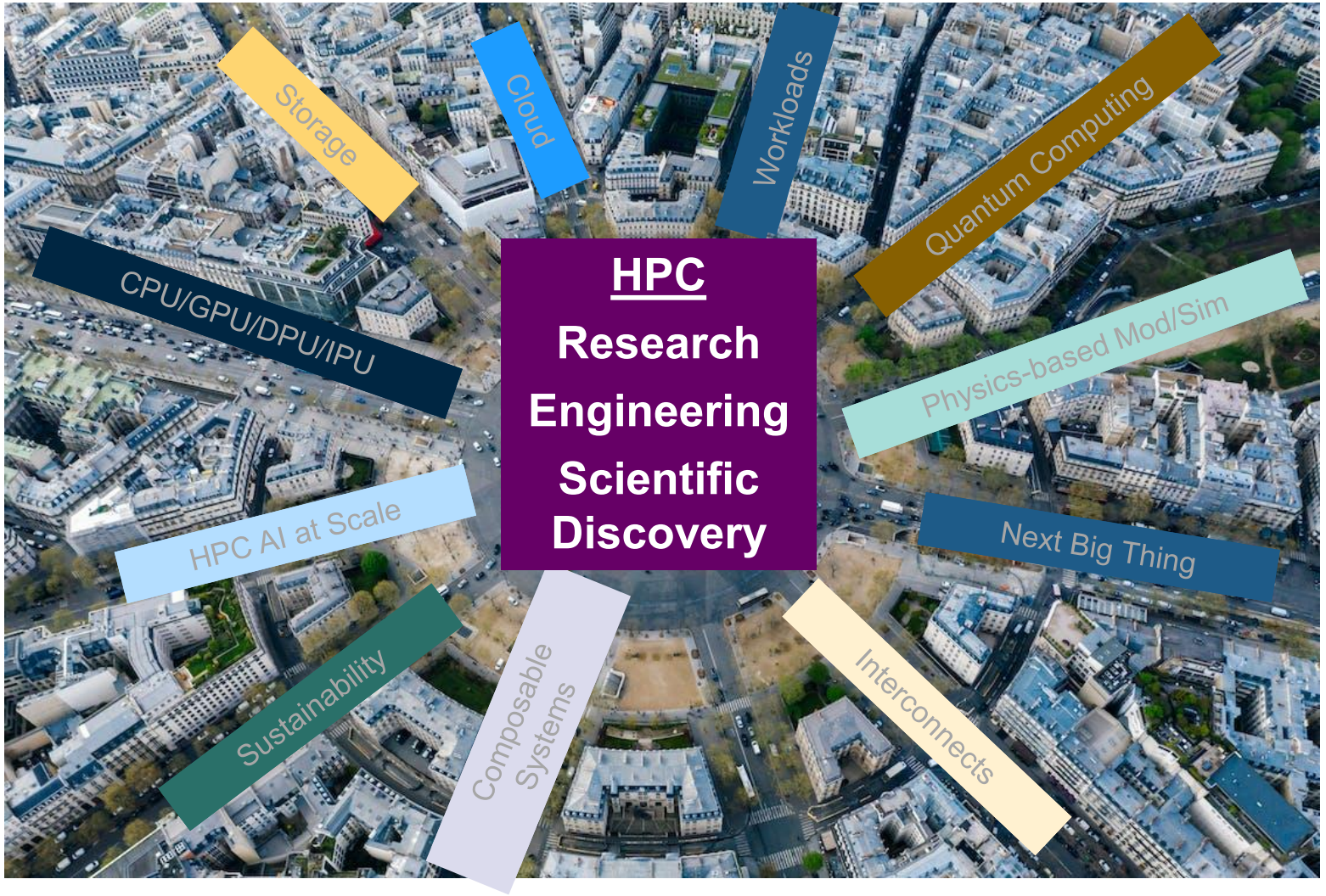
November 2025

www.HyperionResearch.com
www.hpcuserforum.com

Mark Nossokoff and Jaclyn Judema
Research Director

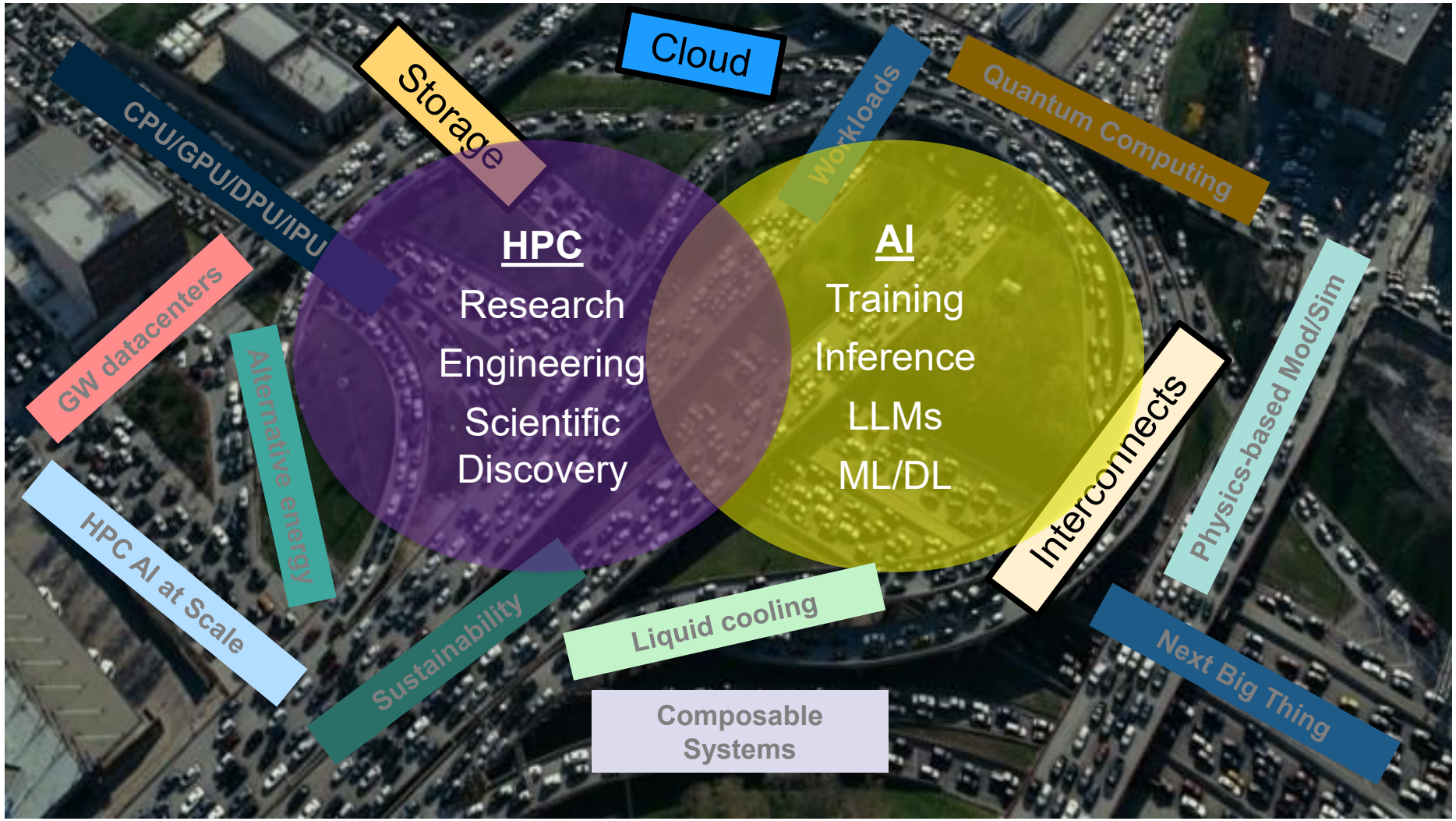
Not Your Father's HPC

A Busy Intersection of Complex Challenges



Not Your Father's HPC... ...now add AI

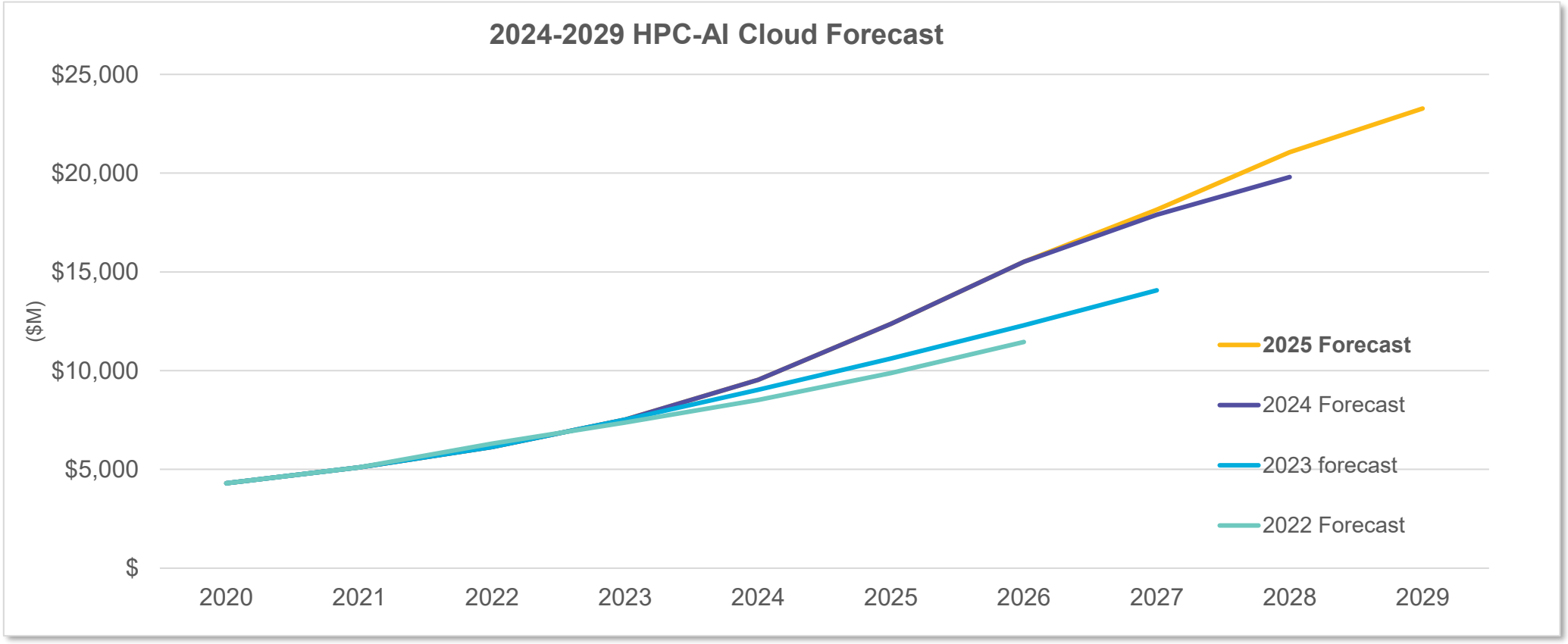
A Busy Intersection of Complex Challenges



Cloud

2024-2029 Global HPC-AI Cloud Forecast

Forecast continues to be revised upward from previous forecasts

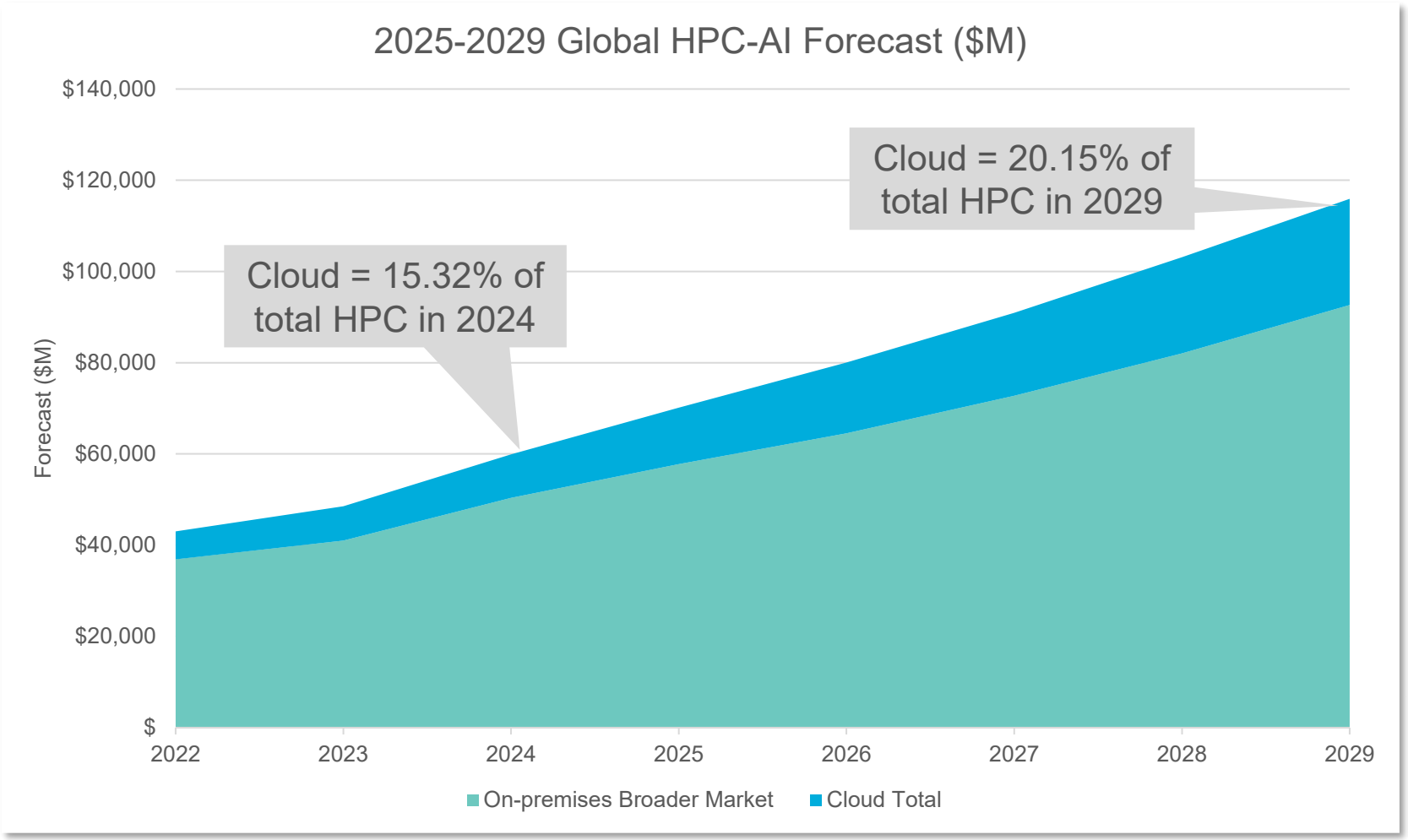


Source: Hyperion Research 2025

Note: Cloud revenues are what users spend on utilization HPC-AI resources in the cloud, as opposed to what cloud providers are spending to provision HPC-AI resources

2025-2029 Global HPC-AI Forecast

Projected 19.5% 5-year CAGR to $23B in 2029

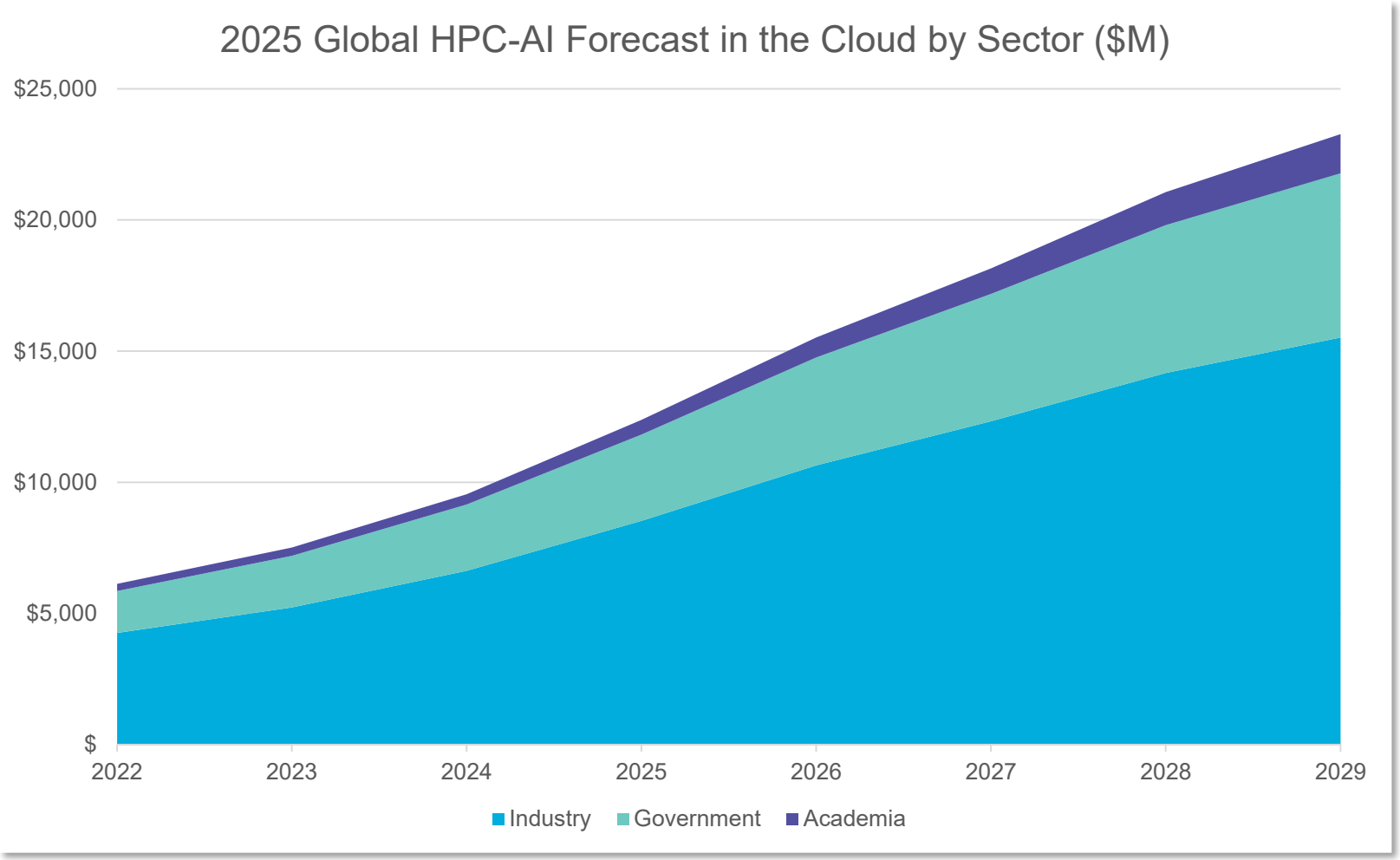


23.5% growth from 2023-2024

	Cloud Total	On-premises	Total
2022	$6,132	$36,900	$43,032
2023	$7,516	$41,008	$48,525
2024	$9,540	$50,391	$59,931
2025	$12,376	$57,751	$70,128
2026	$15,519	$64,507	$80,025
2027	$18,157	$72,779	$90,936
2028	$21,062	$82,074	$103,137
2029	$23,274	$92,656	$115,930
CAGR 24-29	19.50%	13.00%	14.10%

2025-2029 Global HPC-AI Cloud Forecast - By Sector

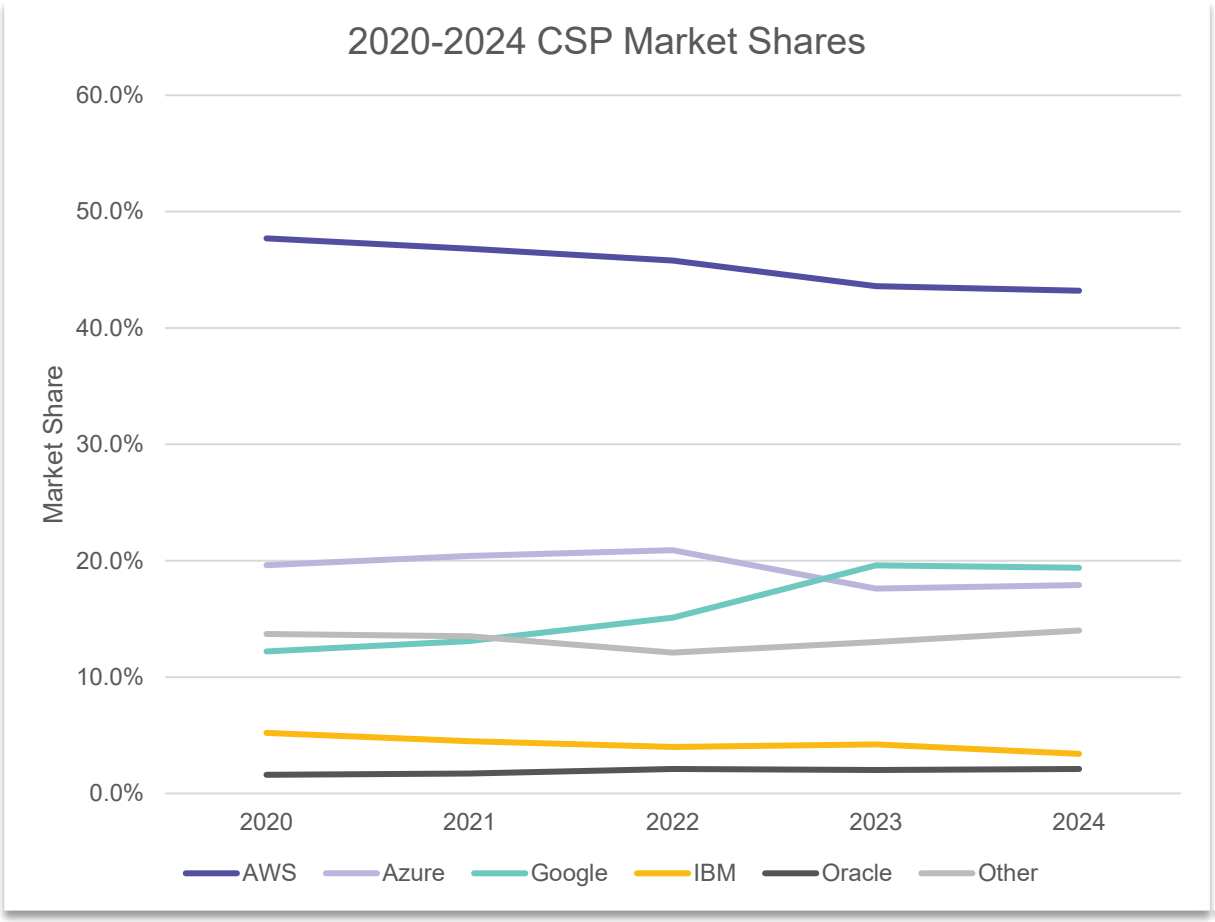
Industry is the largest sector overall, and academia has the highest CAGR

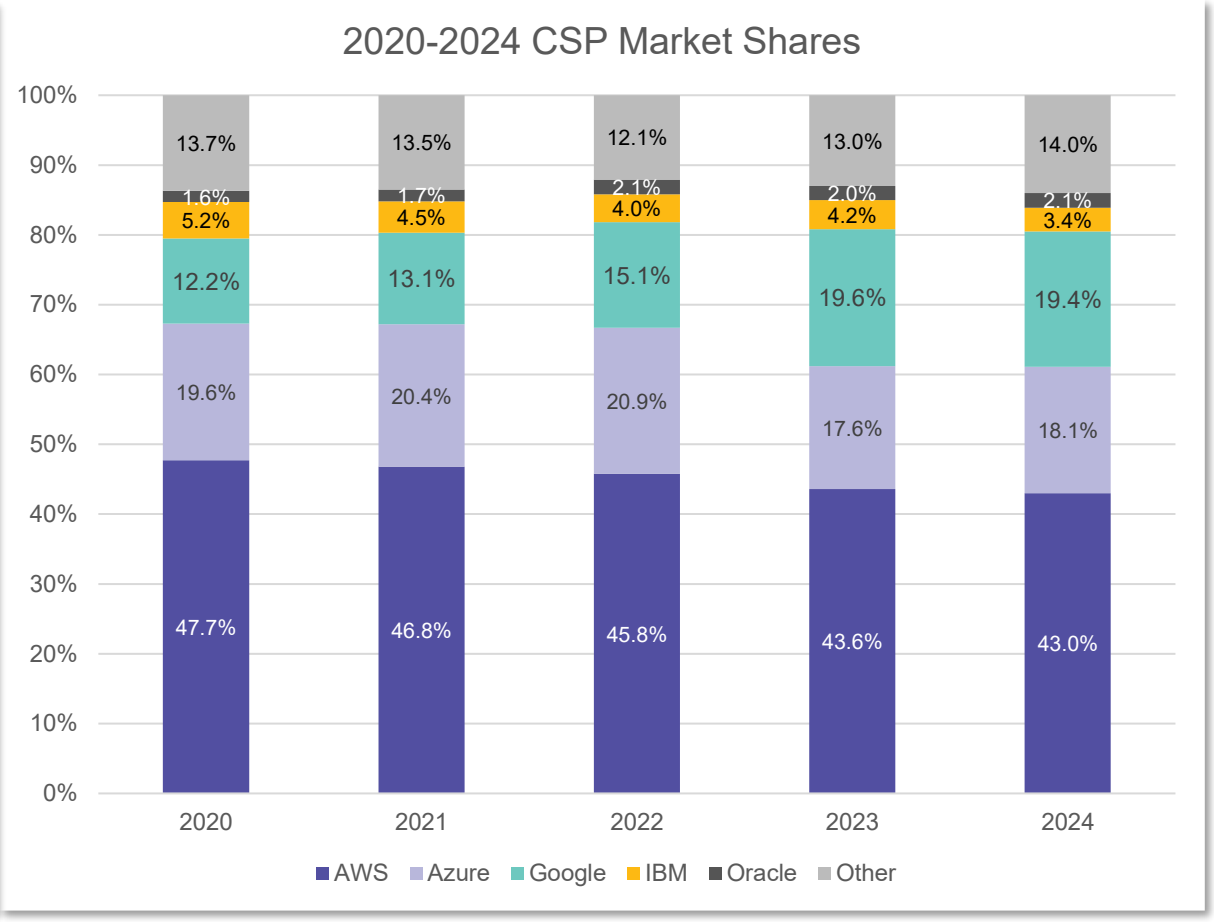


	Industry	Gov	Academia	Total
2022	$4,259	$1,597	$276	$6,132
2023	$5,232	$1,962	$322	$7,516
2024	$6,625	$2,526	$389	$9,540
2025	$8,526	$3,293	$557	$12,376
2026	$10,640	$4,118	$760	$15,519
2027	$12,324	$4,852	$980	$18,157
2028	$14,165	$5,633	$1,264	$21,062
2029	$15,519	$6,265	$1,490	$23,274
CAGR 25-29	18.60%	19.90%	30.80%	19.50%

2020-2024 CSP HPC-AI Market Shares

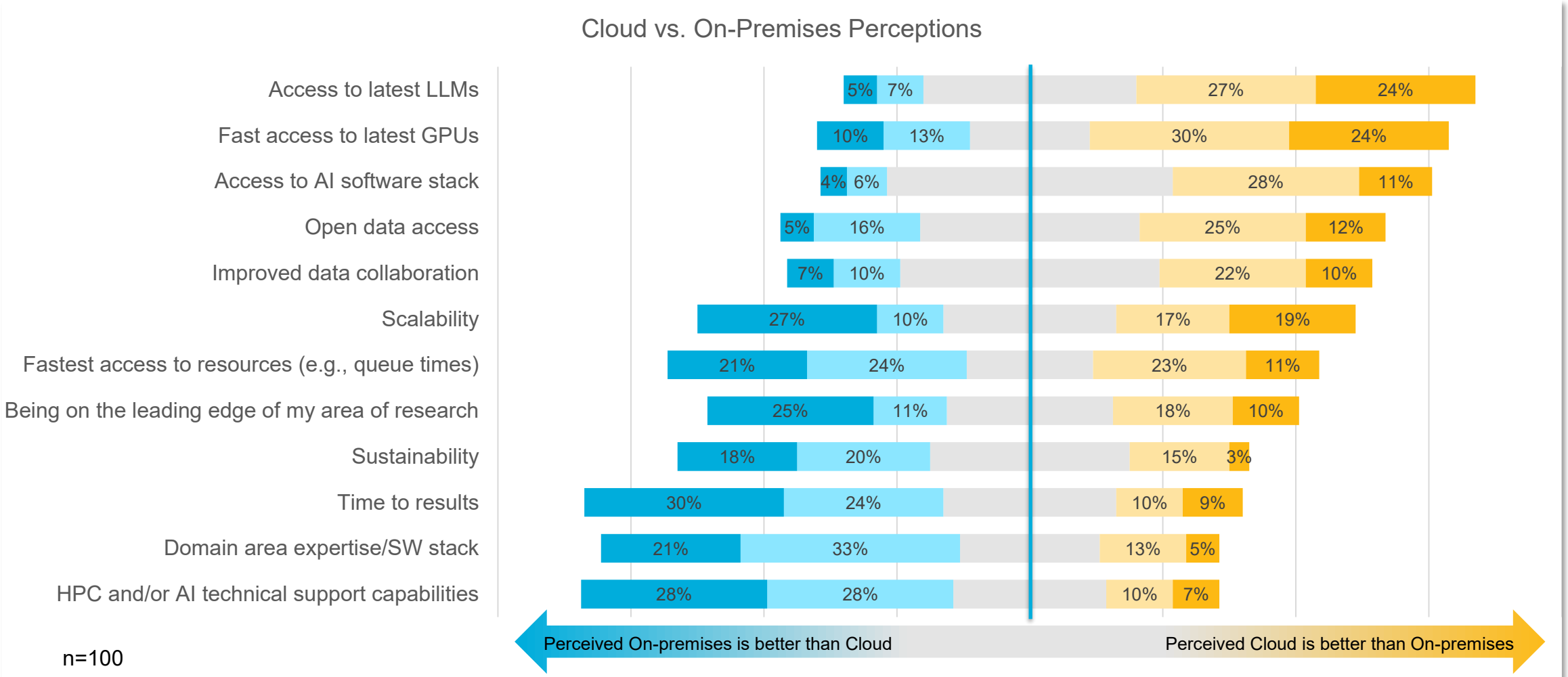
AWS maintains the share lead; Google Cloud gains the most share in recent years; Microsoft Azure gaining traction with AI offerings (watch for our press release)





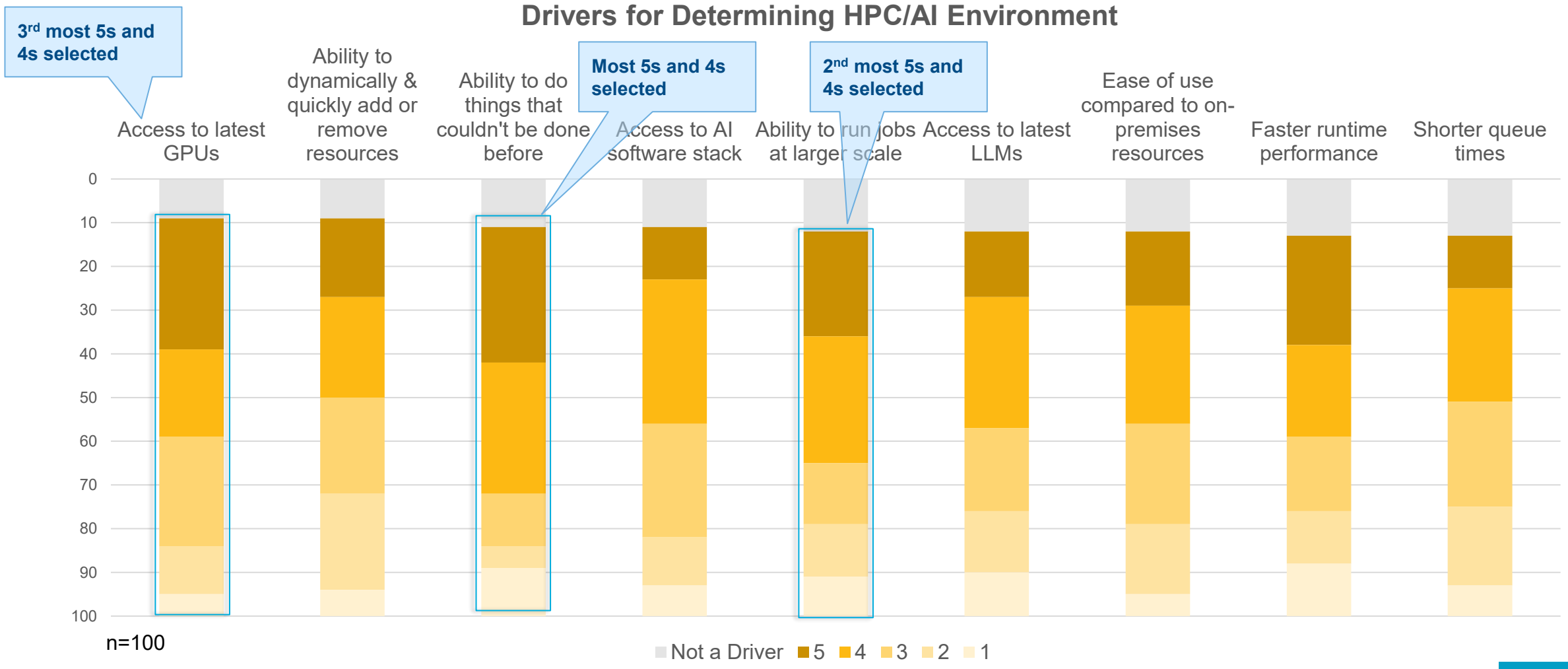
Customer Perceptions of Cloud versus On-premises

More customers are looking to the Cloud as a means of exploring cutting-edge methodologies, including LLMs, GPUs, and AI software stacks



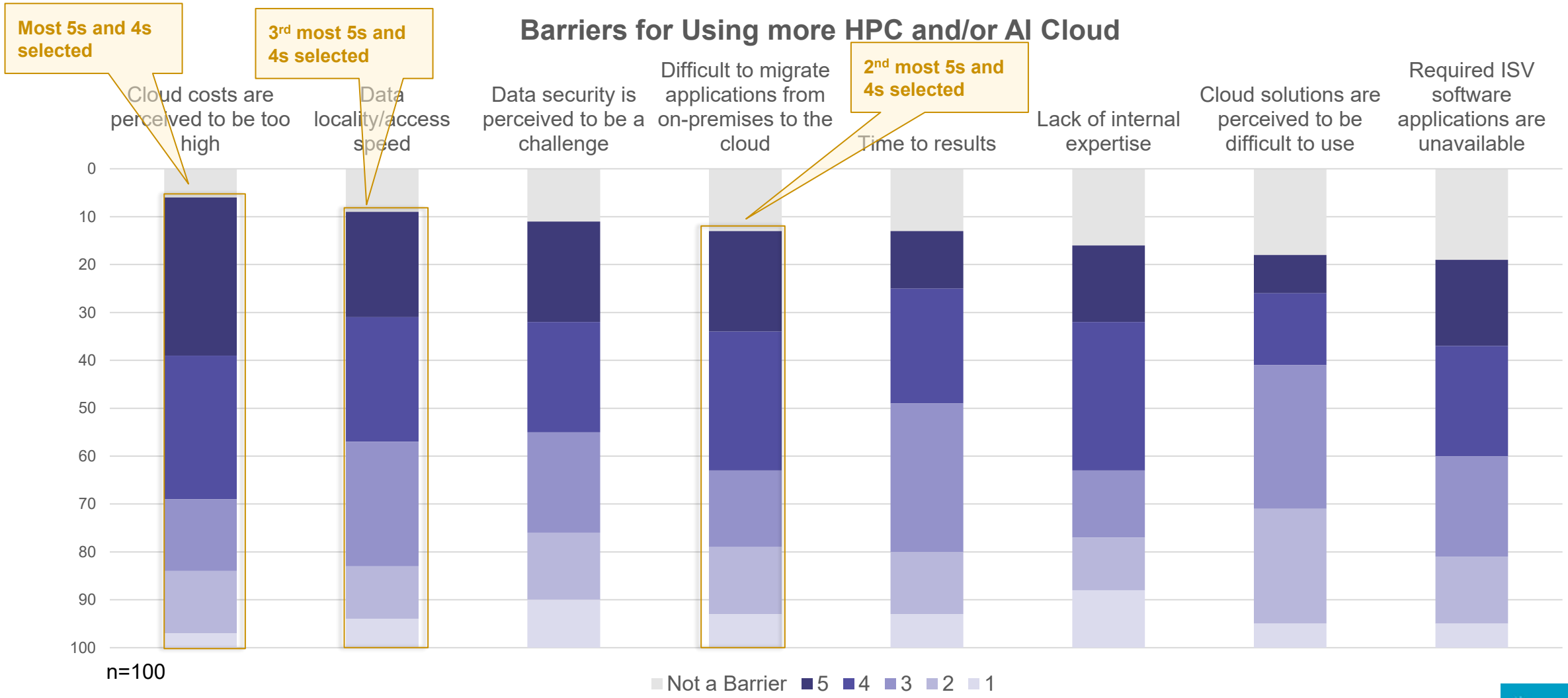
Cloud Drivers

On a scale of 1-5, with 1 being the least impactful and 5 being the most impactful, please indicate the impact of the following drivers in your choice to use HPC and/or AI cloud



Perceived Barriers

On a scale of 1-5, with 1 being the least influential and 5 being the most influential, please indicate the influence of the following perceived barriers on your choice to not use more HPC and/or AI cloud



Making Scientific Compute Decisions Transparent

The Challenge

- Hidden costs: on-premises project budgets often overlook indirect costs,
- Planning uncertainty: Early-stage projects lack precise workload or value figures
- Fragmented Dialogue: Technical, budget, and governance teams all look at TCO differently
- Quantifying Value: Difficulty in assigning a monetary value to the societal impact of fundamental research

The Solution

- A Decision-Support Tool: An Excel-based model developed to guide project planning
- Holistic Comparison: Provide a transparent, data-driven comparison of on-premises, cloud and hybrid environments
- Comprehensive Metrics: Calculates and compares TCO, societal value, and ROI
- Fosters Consensus: Creates a common factual basis to streamline conversations and justify compute environment decisions

Continuum Computing TCO and Value Model

Tool to aid in making scientific computing decisions transparent

- **Researchers and project managers project-specific characteristics:**
  - Area(s) of scientific contribution your project plans to accomplish
  - Potential level of importance and impact to the field of research
  - Perceived value of time-to-solution on the project
  - Computing environment readiness
  - Necessary access to technologies
  - Use of AI
  - Compute requirements
  - Data generation

Project Name:	PNNL Project 1					11/13/2025
Model TCO Summary	Cost Range ($K)		Value Range ($K)		ROI	
Compute Options	Typical Min	Typical Max	Typical Min	Typical Max	Maximum	Minimum
Hybrid	$513	$1,667	$262	$388	0.5 X	0.2 X
Primarily On-premises	$790	$1,888	$262	$388	0.3 X	0.2 X
Primarily Cloud	$294	$1,378	$262	$388	0.9 X	0.3 X
Adjusted TCO Summary	Cost Range ($K)		Value Range ($K)		ROI	
Compute Options	Typical Min	Typical Max	Typical Min	Typical Max	Maximum	Minimum
Hybrid	$369	$1,667	$262	$388	0.7 X	0.2 X
Primarily On-premises	$790	$1,888	$262	$388	0.3 X	0.2 X
Primarily Cloud	$349	$828	$262	$388	0.8 X	0.5 X

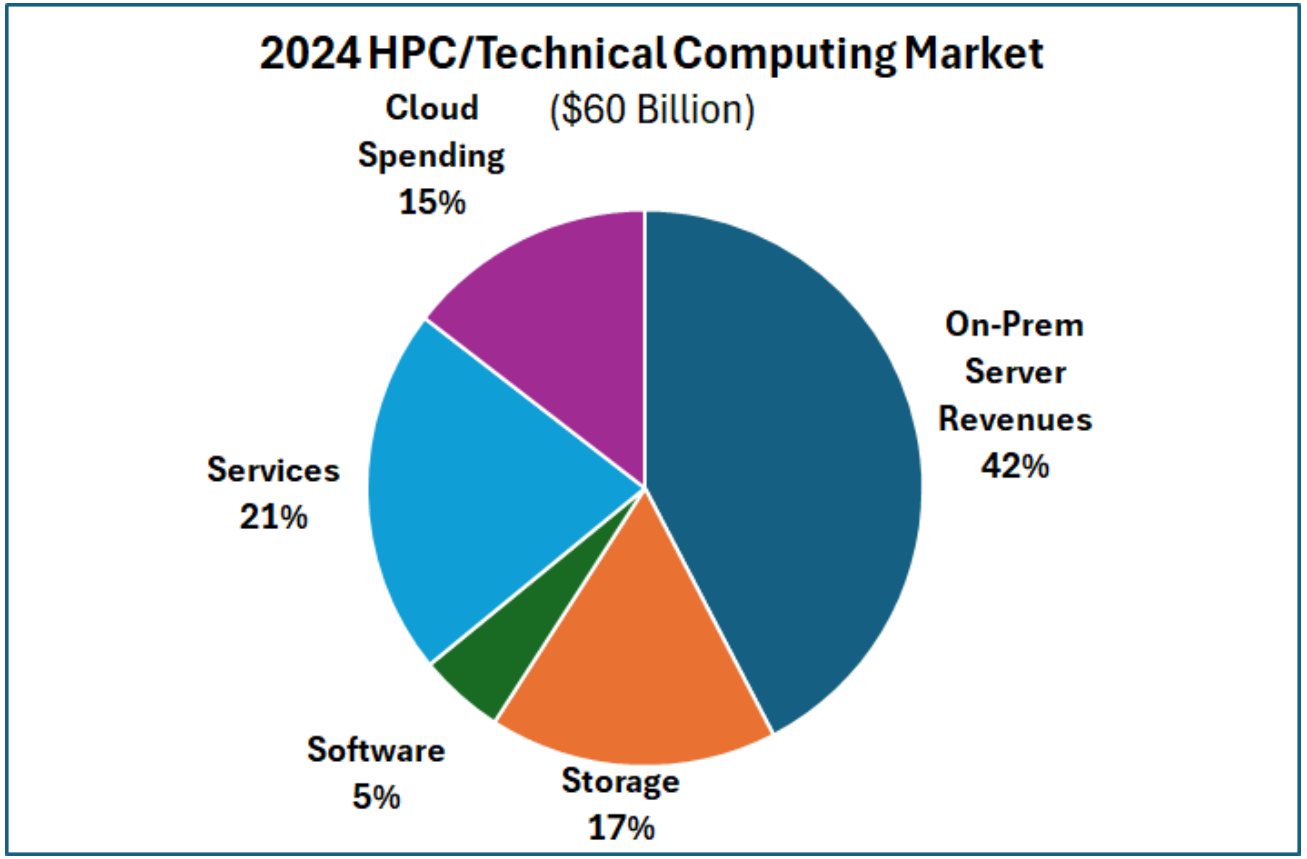
Source: Hyperion Research, 2025

Storage

2024 Was a Strong Growth Year – Total Market

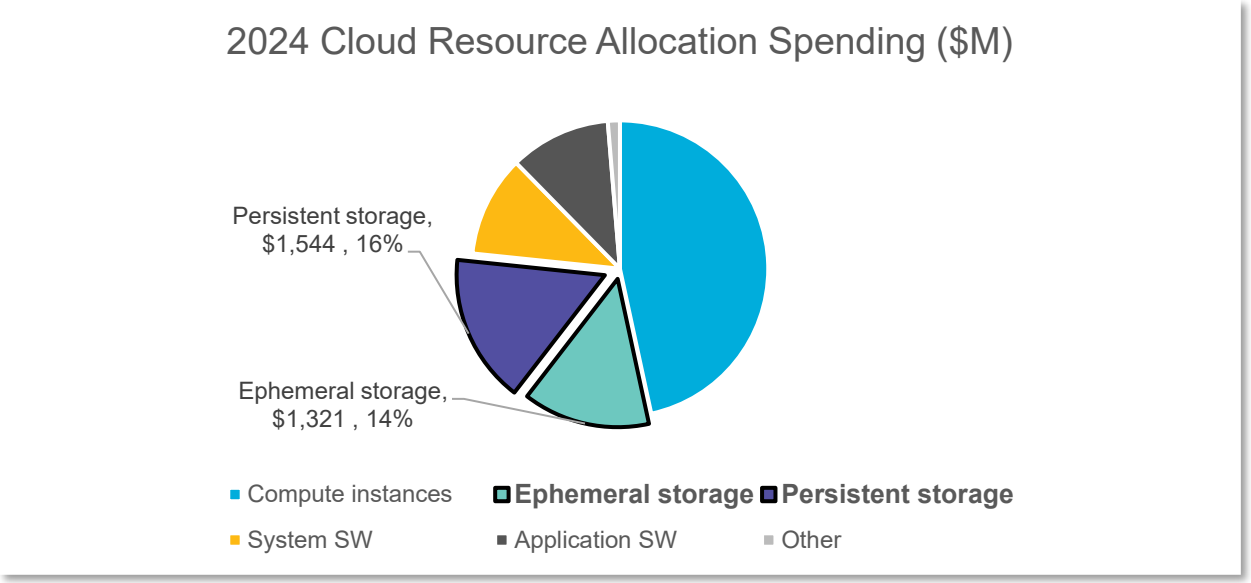
The highest growth in over two decades (23.5%)!

- **Overall market**
  - 23.4% growth in on-premises servers
  - 21.3% growth in the use of clouds
  - Over $60 billion in total spending
- **Storage market**
  - 22.4% growth in 2024
  - 21.4% of total market
    - 30% of cloud spending
    - 21.7% of on-premises spending



HPC-AI Cloud Resource Allocation Spending & Forecast

Cloud storage spending projected to approach ~$7B in 2029



Source: Hyperion Research, 2025

- **Storage comprises ~ 30% of HPC spending in the cloud**

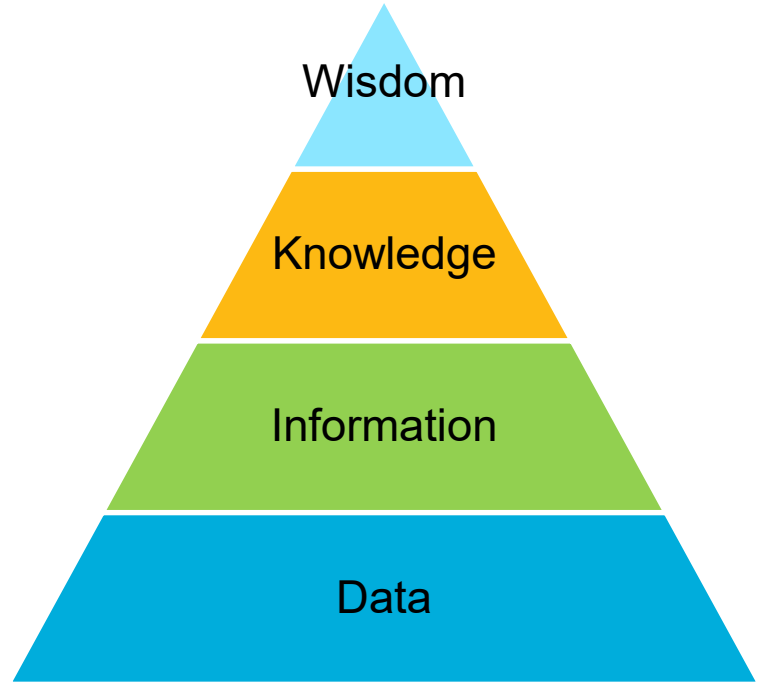
($M)	2024	2025	2026	2027	2028	2029
Compute instances	$4,446	$5,768	$7,232	$8,461	$9,815	$10,846
Ephemeral storage	**$1,321**	**$1,714**	**$2,149**	**$2,514**	**$2,917**	**$3,223**
Persistent storage	**$1,544**	**$2,002**	**$2,511**	**$2,938**	**$3,408**	**$3,766**
System SW	$1,051	$1,363	$1,709	$2,000	$2,319	$2,563
Application SW	$1,052	$1,365	$1,711	$2,002	$2,322	$2,566
Other	$125	$162	$203	$238	$276	$304
Total	$9,540	$12,376	$15,519	$17,892	$19,804	$23,268

Source: Hyperion Research, 2025

Data Platform* Emerges as Key Storage Value Point

While "speeds and feeds" will continue to be important to buyers of storage systems, the primary value point and competitive advantage for data storage solutions will shift to the "Data Platform"

- **Predictable cadence of bandwidth, throughput, and latency performance improvements with each generation of storage system is base table stakes**
- **Speeds and feeds alone do not reflect the value of a storage system's primary asset – the data**
- **Long term storage system business success will be driven by a data platform:**
  - Providing users their data wherever, whenever, however in a performant, reliable manner
  - Delivering capabilities to users to derive value from the data they feed into their scientific, engineering, and business workloads



* Single and global namespace solutions to orchestrate and move data at scale

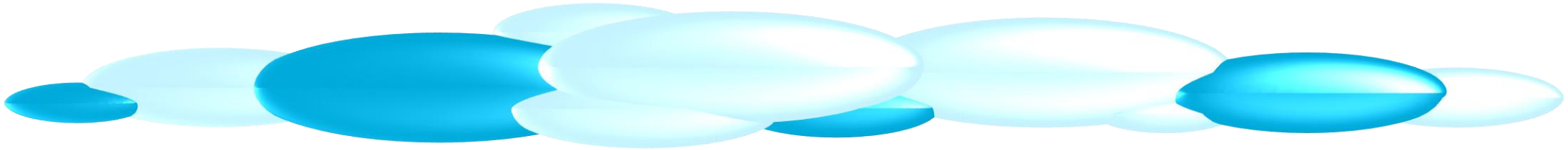
Homogenous HPC Workload Considerations

High bandwidth with fault tolerance to withstand cluster component failures during long traditional modeling/simulation jobs

Checkpoint

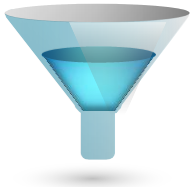
Heterogenous HPC & AI Workload Considerations

High bandwidth with fault tolerance to withstand cluster component failures...
...plus diverse range of profiles and requirements to keep GPUs fed with data



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?**

Data Platforms

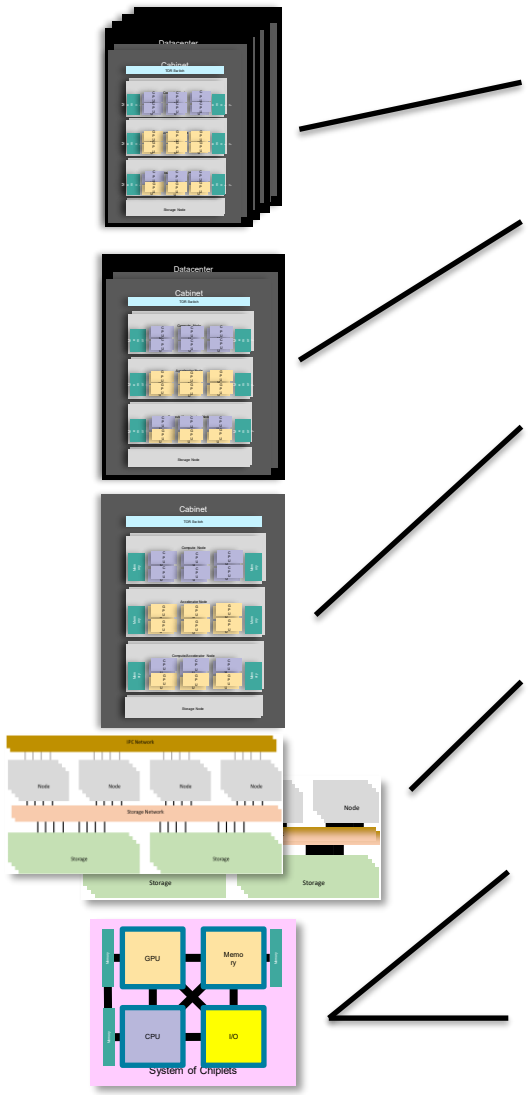
Growing number of vendors adopting the market segment

- DDN Infinia
- Hammerspace
- HPE
- Huawei
- IBM
- NetApp data platform
- Oracle data platform
- Pure Storage
- VAST AI Operating System
- VDURA
- Weka

Interconnects

Hierarchy of Interconnects

Scale Across
Scale Out
Scale Up



Use Case	Definition	Interconnect
Data center to outside world	Connection between geographically dispersed data centers, and between remote users and the data center	Ethernet
Rack to rack	Connection between racks within a data center	Bxi Ethernet InfiniBand OmniPath Slingshot
Node to node (aka, server to server)	Connection between compute elements either between modules in a rack or within a shelf; typically carries traffic associated with Interprocessor communication such as MPI In many cases this interconnect is also used for rack to rack	Bxi Ethernet InfiniBand OmniPath Slingshot
Node to storage (aka, server to storage)	Connection between compute elements and storage elements either between modules in a rack or within a shelf	Ethernet InfiniBand OmniPath PCIe Slingshot
xPU to xPU xPU to memory	Connection between compute elements within a node	InfinityLink NVLink PCIe
Chiplet to chiplet in-package	Connection between functional elements within a package	InfinityLink NuLink NVlink UCIe

Source: Hyperion Research, 2025

Emerging (Proliferation?) Standards

- **Ultra Ethernet Consortium (UEC)**
  - Scale-out
  - Contributions to Linux kernel
  - Released Rev 1.0
  - NVIDIA joined
  - Much of UEC is Slingshot
  - Now under the Linux Foundation
- **Ethernet Scale-Up Networking (ESUN)**
  - Scale-up
  - Announced at 2025 OCP Summit
- **UltraAccelerator Link (UALink) Consortium**
  - Released version 1.0 of spec
  - NVIDIA absent
- **InfiniBand**
  - Incumbent but adoption may have peaked
  - Quasi-standard; sole-sourced by NVIDIA
- **NVLink Fusion for 3rd party integration**
  - More than serdes
  - NVIDIA's response to UALink
  - Intel added in conjunction with corporate investment from NVIDIA

Interconnect Standards Leadership

Multiple companies support all

- **Founding members are also board members**
- **Certain board members were elected after the standard's formation**
- **Broadcom is no longer on the UAL board but still a member**
- **HPE includes Juniper leadership**
- **NVIDIA participating but no leading**

Company	Ultra Ethernet Consortium (UEC)	Ethernet Scale-Up Networking (ESUN)	Ultra Accelerator Link (UALink)
AMD	Founding Member	Founding Member	Founding Member
Alibaba	Member		Board Member
Apple			Board Member
Arista	Founding Member	Founding Member	
ARM		Founding Member	
Astera Labs			Founding Member
AWS			Board Member
Broadcom	Founding Member	Founding Member	Founding Member
Cisco	Founding Member	Founding Member	Founding Member
Eviden (Atos)	Founding Member		
Google			Founding Member
HPE	Founding Member	Founding Member	Founding Member
Intel	Founding Member		Founding Member
Marvell	Member	Founding Member	
Meta	Founding Member	Founding Member	Founding Member
Microsoft	Founding Member	Founding Member	Founding Member
NVIDIA	Member	Member	
OpenAI		Founding Member	
Oracle	Board Member	Founding Member	
Synopsys			Board Member

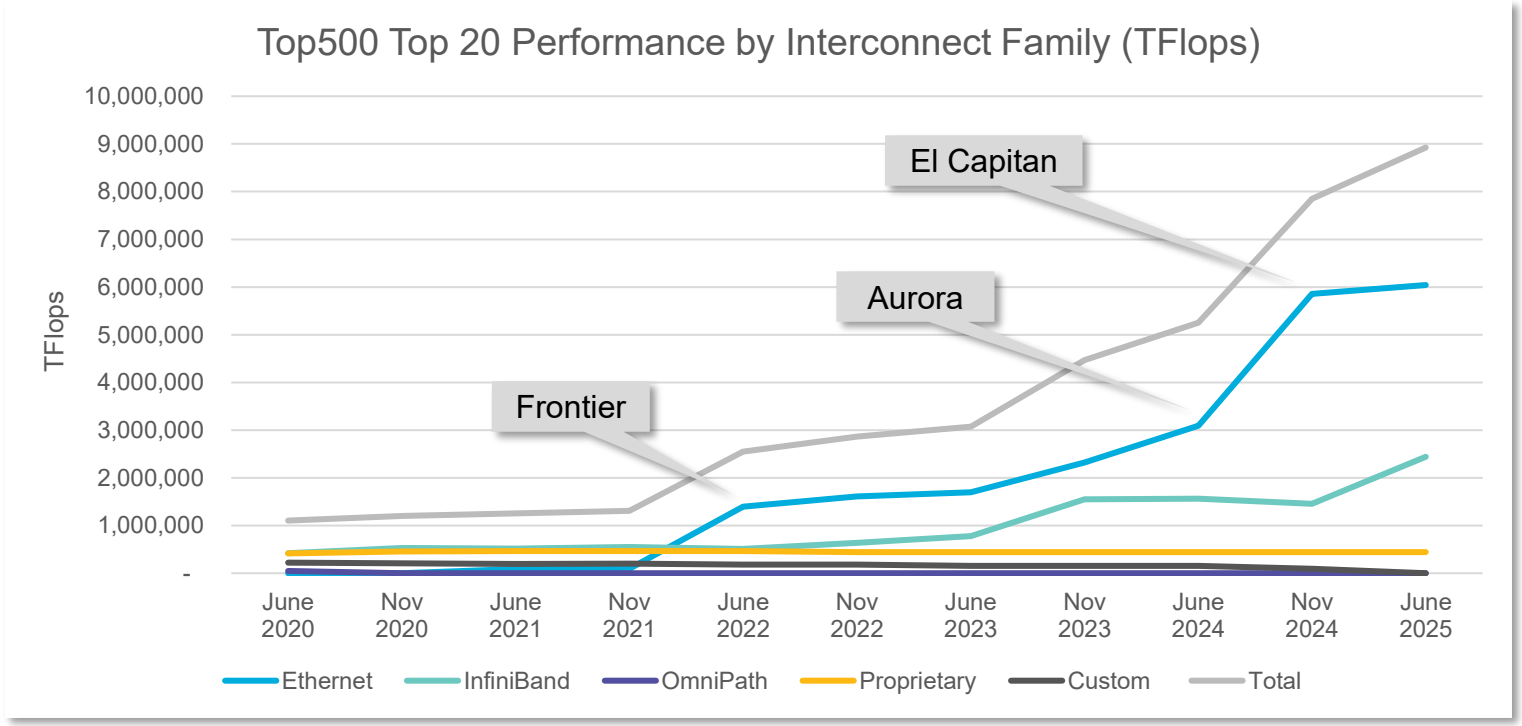
Other Interconnect Considerations

- **Other interconnects**
  - HPE Slingshot
    - Increasing line rates
    - Heavy contributions to UEC
    - Increasing promotion and visibility within the market (e.g., Slingshot workshop at SC25)
  - Eviden Bxi
    - Increasing line rates
    - Roadmap to intercept UEC
  - Cornelis OmniPath
    - Increasing line rates
    - Roadmap to intercept UEC
  - Huawei
    - UB-Mesh
    - Challenging NVLink
    - Open source the spec
- **CSPs**
  - Oracle Zettascale10 Acceleron RoCE networking
  - AWS EFA sidecar
  - Google Falcon and optical switching
- **Technology**
  - Optical adoption

Top500 Top 20 Performance by Interconnect Family

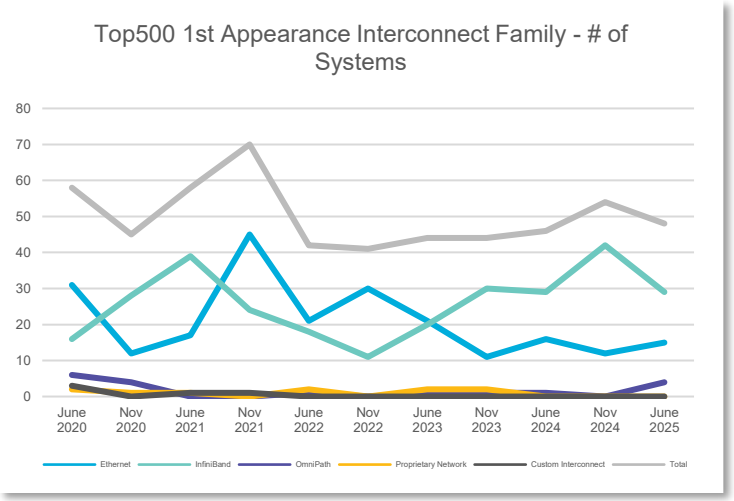
Ethernet emerged in 2021 and established itself with Frontier 2022

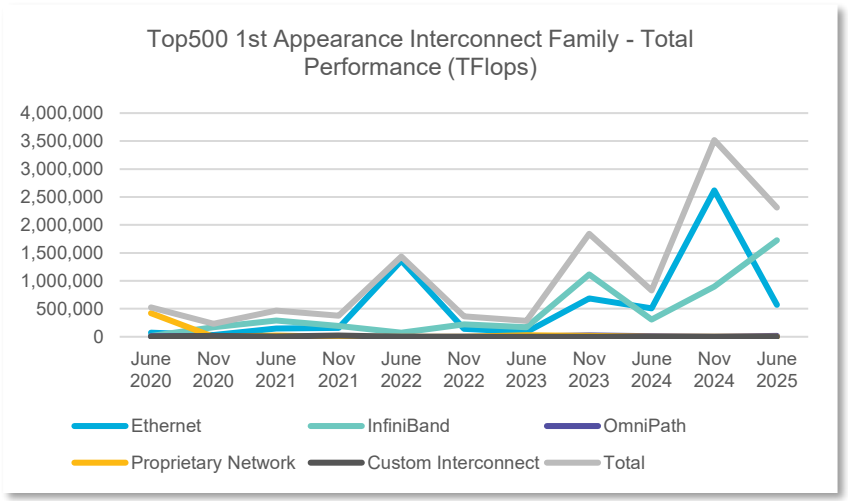
- **Top500 characterizes Slingshot as Ethernet**
- **June 2021-June2025 4-year Top20 performance CAGRs**
  - Ethernet: 191.3%
  - InfiniBand: 47.7%

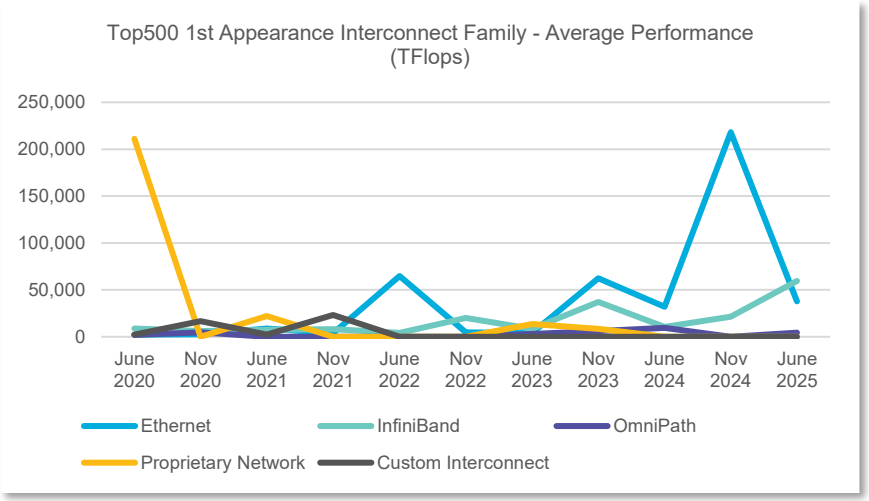


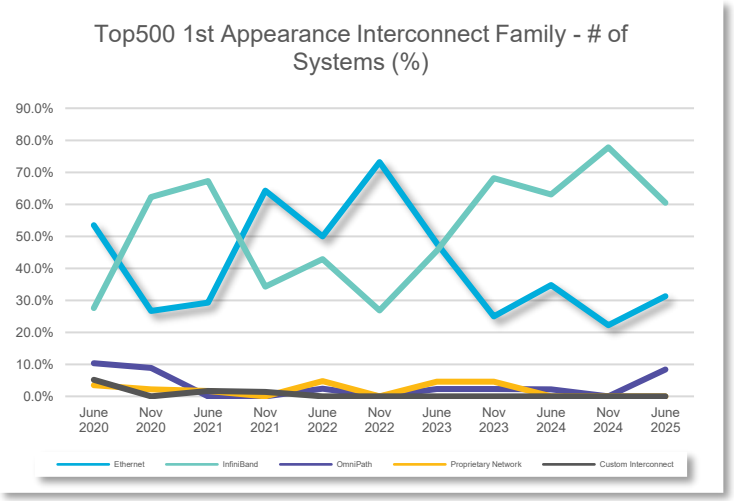
Source: Hyperion Research 2025

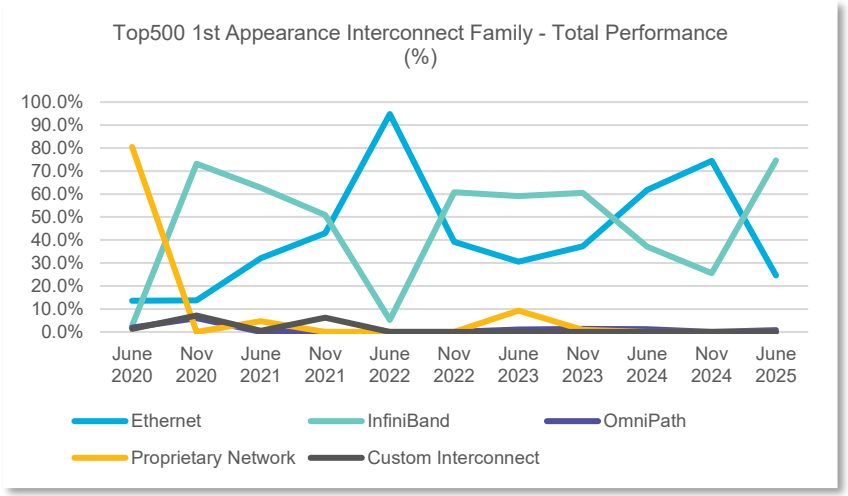
Top500 1st Appearance by Interconnect Family

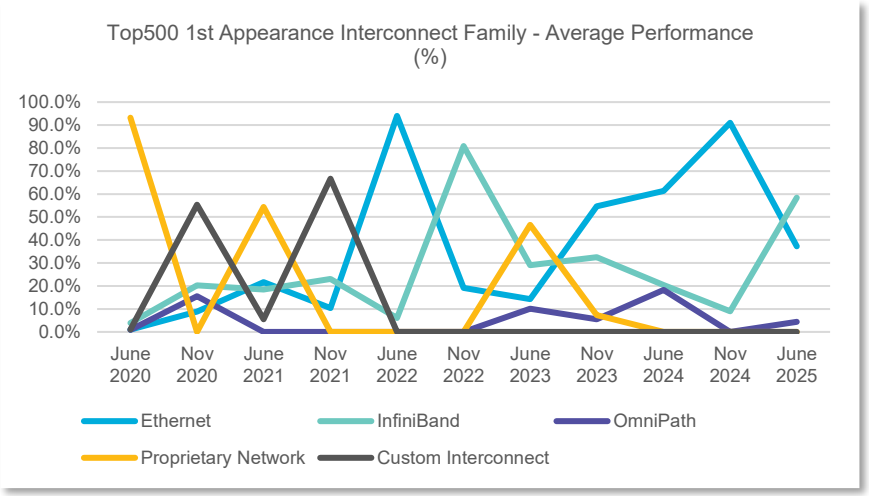












What Does This Mean for Future Interconnects?

Hetergenous AI and HPC workloads driving diverse set of requirements

- **Scalability**
  - Exascale
    - Capable of millions of endpoints
    - Multiple topologies
  - Terascale/Petascale
    - SME (Small Medium Enterprise)
- **High Performance**
  - High bandwidth
  - Low latency
- **Diversity of workloads**
  - HPC (FP64)
  - AI (lower precision, large scale out radix)
  - Random, sequential
  - Read, write
  - Small block, large block
  - File, block, object
- **Security**
  - Encryption
  - Authentication
- **Transport media options**
  - Copper
  - Optical
- **Compatibility/Choice**
  - Generation to generation
  - Multi-vendor
  - Interoperability
- **Extensibility**
  - Differentiation
- **Cost effective**

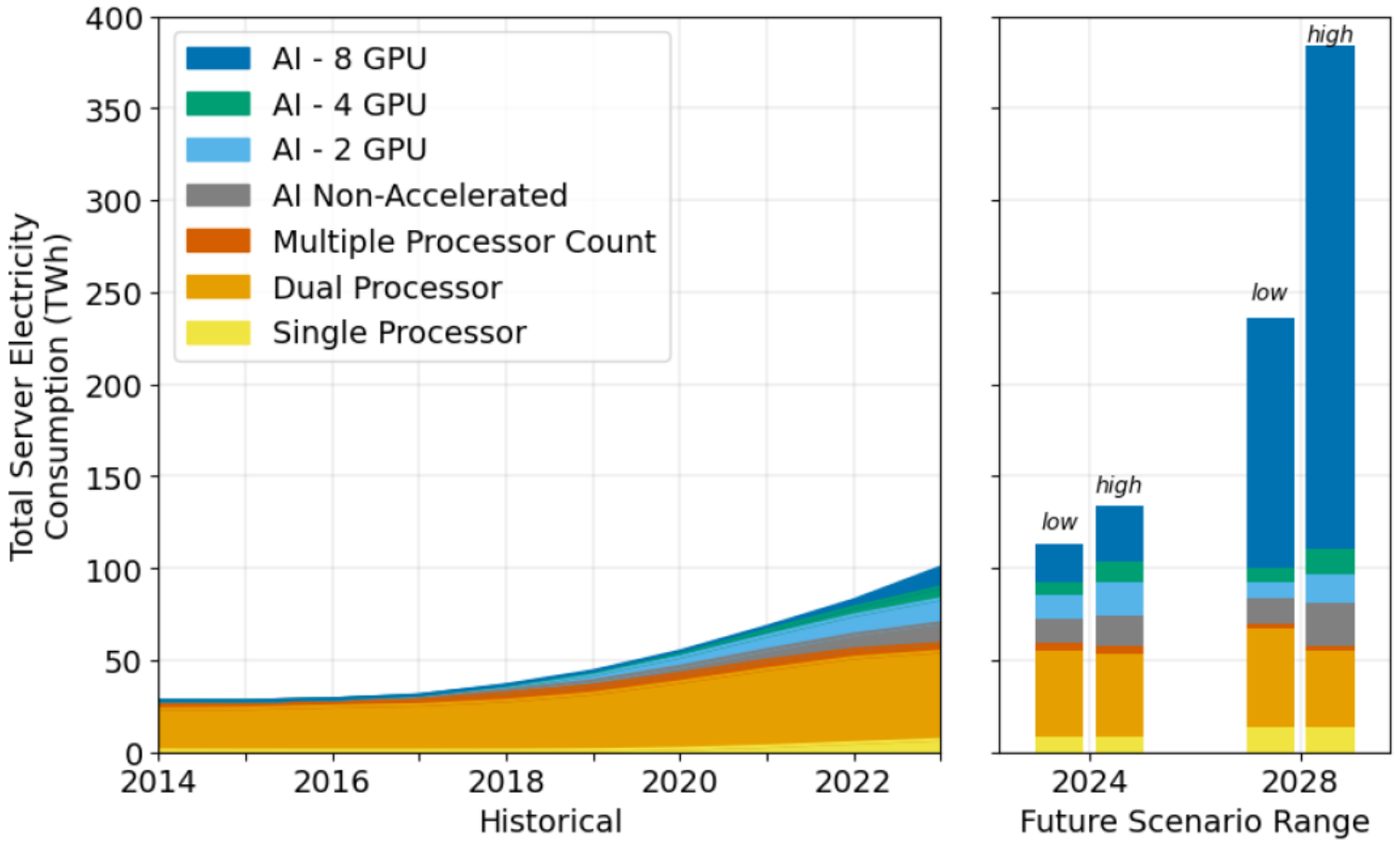
Sustainability

Sustainability Global Stage

- **In 2024, renewables made up about 92.5% of the total global power capacity expansion (585 GW)[1]**
  - 91% of new renewable projects are now cheaper than fossil fuel alternatives
  - 64% of new renewable projects were in China, just 7% in US
- **The United States has huge renewable energy capacity, but lacks transmission capacity**
  - In 2024, only 322 miles of high-voltage transmission lines were completed (steadily falling from a peak of 4,000 miles built in 2013)[2]
  - This slowdown may put semiconductor manufacturing and artificial intelligence at risk
  - Natural gas pipelines support significantly more energy transmission in the US than renewables[3], which may get expensive given new renewable affordability

1 https://www.irena.org/Publications/2025/Mar/Renewable-capacity-statistics-2025
2 https://cleanenergygrid.org/portfolio/report-fewer-new-miles-strategic-industries-held-back-by-slow-pace-of-transmission/
3 https://rextag.com/blogs/blog/how-huge-we-are-u-s-natural-gas-pipelines-infrastructure-2024-overview-by-rextag?srsltid=AfmBOopVmZRJw7VCI7ptOFt7UkbUx3s5eiq1nIjiqZAaXPNZTYp34UsN

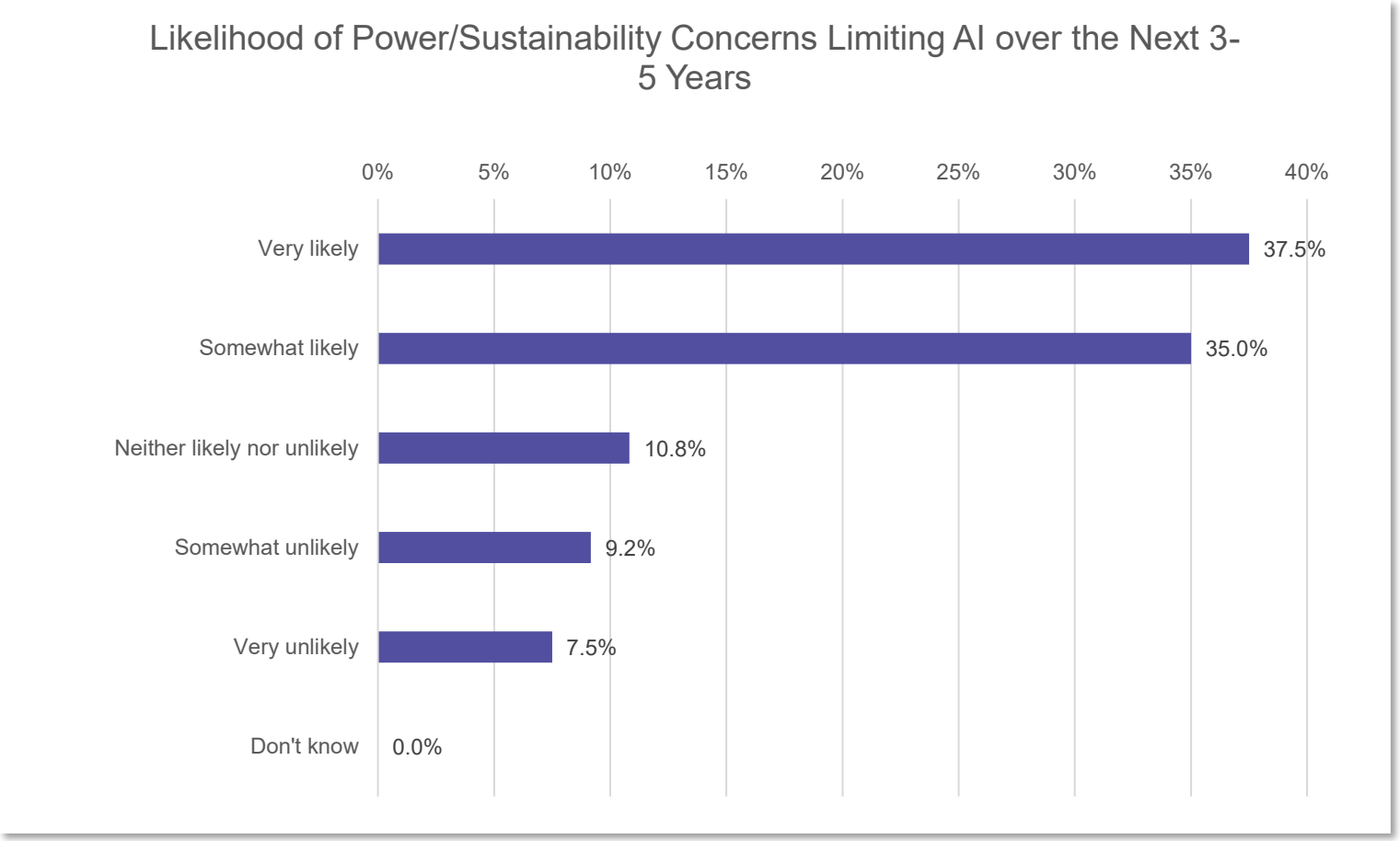
Server Annual Electricity Usage by Type



Source: Lawrence Berkeley National Lab, 2024

- **Traditional HPCs largely found in the "Multiple Processor Count" category, per the study methodology (orange bar)**
- **On the low and high range of future scenario ranges, AI is projected to dominate total server electricity consumption by 2028**

How likely is it that power/sustainability concerns will limit the use of AI over the next 3-5 years?

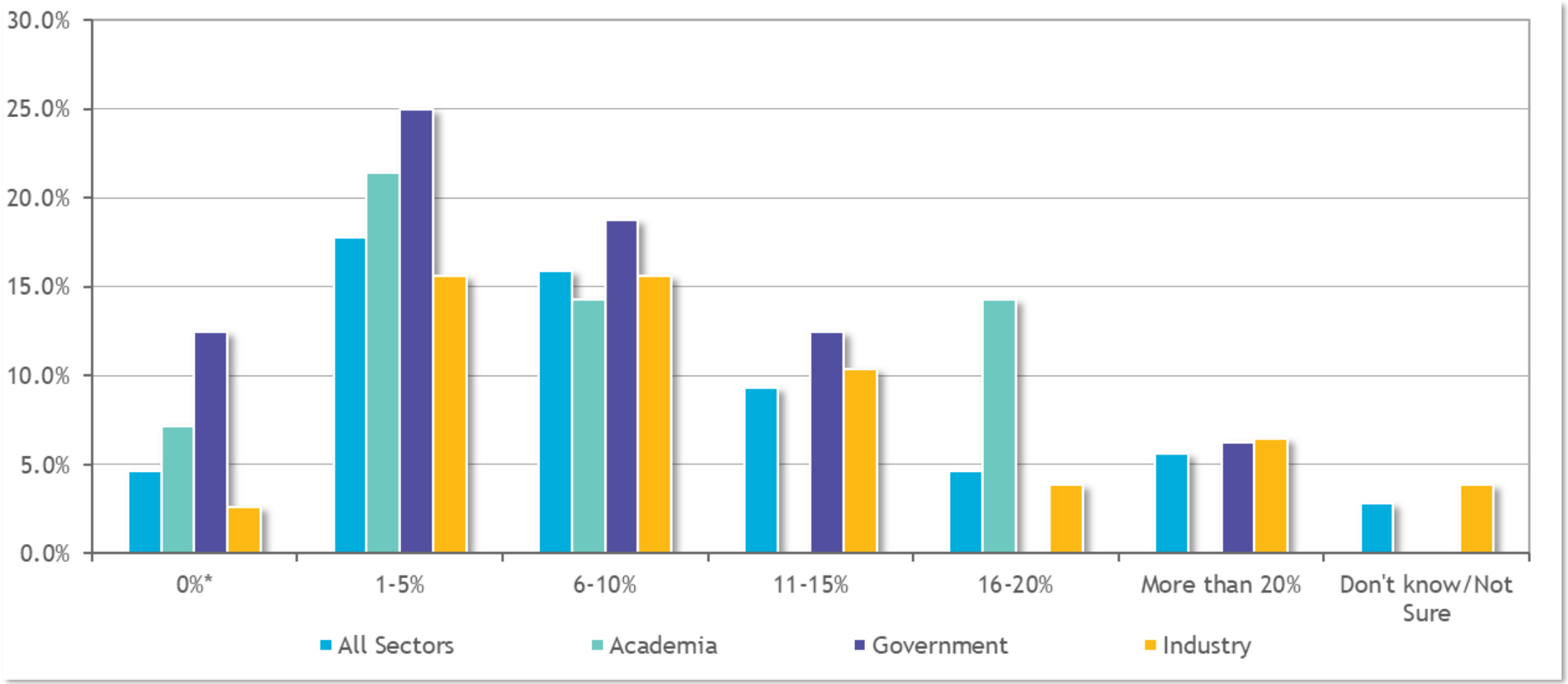


- **Strong consensus on impact: 72.5% of respondents think power/sustainability will likely (very or somewhat) limit AI use in the near term**
- **Few (16.7%) think it is unlikely (very or somewhat)**
- **Power and sustainability considerations are no longer peripheral concerns in the AI landscape. They are likely to be a limiting factor…soon**

N = 120, Source: Hyperion Research, 2025

Willingness to Exchange Performance for Energy Efficiency

Limited appetite to tradeoff performance for energy efficiency



n=65, no response: 42/107

New Sustainability Technology

- **Battery storage at datacenter scale**
  - Aligned Data Centers and Calibrant Energy: New 31MW/62MWh battery energy storage system (BESS) at their Hillsboro, Oregon campus
  - Stores during high production periods, and discharges during peak demand periods
  - Useful to regional power grids not equipped to keep pace with power requirements of advanced computing and AI
- **New liquid cooling system: Microsoft In-Chip Microfluidics**
  - Like capillaries in the body, tiny channels are etched in the back of individual chips, allowing liquid to pass right against heated silicon and remove heat
  - May increase rack density

Call to Action and Coming Attractions

- **New website:** High Performance Computing (HPC) Research | Hyperion Research
  - Check it out!
- **New global site survey**
  - Launched in 4Q25
  - Look for results to start rolling out in 1Q26
- **Continuum Computing TCO/Value/ROI Model & Tool**
  - Sponsored by PNNL
  - Assist in providing project-based guidance on TCO and ROI analysis between cloud and on-premises infrastructure
  - Based on direct research on TCO

HYPERION RESEARCH

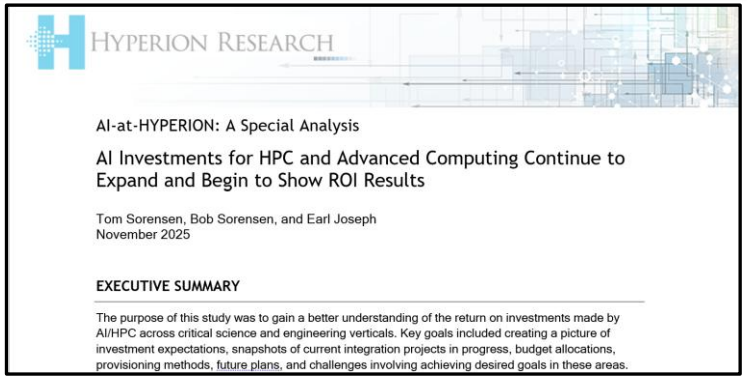
Questions?

mnossokoff@hyperionres.com
jludema@hyperionres.com

In Summary

Just Completed Study:

"AI Investments for HPC and Advanced Computing Continue to Expand and Begin to Show ROI Results"


HYPERION RESEARCH

AI-at-HYPERION: A Special Analysis

AI Investments for HPC and Advanced Computing Continue to Expand and Begin to Show ROI Results

Tom Sorensen, Bob Sorensen, and Earl Joseph
November 2025

EXECUTIVE SUMMARY

The purpose of this study was to gain a better understanding of the return on investments made by AI/HPC across critical science and engineering verticals. Key goals included creating a picture of investment expectations, snapshots of current integration projects in progress, budget allocations, provisioning methods, future plans, and challenges involving achieving desired goals in these areas.

Highlights from the study:

- **75.8% of the sites felt that their AI projects met or exceeded expectations**
- **74.9% of respondents indicated plans to moderately or significantly expand generative AI to support HPC workloads**
  - Less than 3% expect to contract their use of AI, none of which would characterize that contraction as significant
- **Roughly 40% of respondents are already using agentic AI models**
- **However, technical challenges continue to bring hesitation when it comes to broad adoption:**
  - Hallucinations, lack of explainability, and integration complexity are persistent concerns
  - This signals a transition from reactive adoption to more measured, application-specific onboarding

Conclusions

Projecting strong growth

- **2025 is grew will in the first half of the year at 22%**
- **2024 was a strong growth year at 24%**
  - AI for HPC, science and engineering is growing very fast
  - Clouds, GPUs, are high growth areas
  - QC systems are being installed around the world
- **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
- **Storage will likely see major growth driven by AI, big data and the need for much larger data sets**
- **There are growing concerns around system costs, power, talent and political changes**

A Concern: AI & HPC Expertise Shortage

The growing scarcity of HPC and AI experts to implement new technologies is the number one roadblock for many HPC sites

- **Three major trends:**
  1) A shrinking HPC workforce
  2) A massive increase in system complexity and new technologies
  3) AI requires new skills, and non-HPC AI sites are aggressively hiring talent

  - **HPC experts are an aging workforce**
    - The pipeline of new HPC staff entering the workforce does not match the outflow of retirees
    - Competition for HPC & AI staff is growing rapidly
  - **Increasingly complex workloads are more difficult to manage**
    - Increasing number of AI & HPC systems per site
    - Incorporating multiple processor types, co-processors, accelerators, and other specialized hardware
    - Balancing on-prem and cloud
    - Enterprise IT users are entering the AI/HPC space and need HPC expertise
  - **AI & HPC users will need major improvements in ease-of-use, ease-of-selection, and ease-of-optimization**

Another Concern: Systems are Moving Away From 64bit

HPC end users with major investments in legacy codes built on 64-bit floating-point data formats, will need to explore using the capabilities of mixed and lower precision hardware

- **Many AI applications do not need 64-bit floating-point formats**
  - They often require only 32-bit, 16-bit, 8-bit lower floating point or integer schemes
- **GPU designers are increasingly optimizing their chip and core designs to take advantage of this trend**
  - Configuring hardware to offer increased computational performance with lower memory overhead for these mixed and lower precision AI jobs
- **Creating opportunities/concerns for traditional HPC end users**
  - Performance on lower precision is growing when compared with counterpart gains for 64-bit floating point
  - Potentially leaving future processors underpowered for some traditional science and engineering applications or forcing major, if not complex, HPC end user rewrites of existing legacy codes

HYPERION RESEARCH

HPC Innovation Excellence Awards

November 2025

www.HyperionResearch.com
www.hpcuserforum.com

Examples Of Previous Winners

The Trophy and Certificate For Winners

SC25 2025 Winners:

HPC User Innovation Excellence Awards

End User Award: HPC for Earthquakes Applied as a Service

Organizations: Barcelona Supercomputing Center (BSC) and Mexico's National Seismological Service (SSN)

- **Innovation: The participation of the Barcelona Supercomputing Center (BSC) and the ChEESE Centre of Excellence in Mexico's National Earthquake Drill marked a milestone in the global application of European high-performance computing.**
  - Urgent computing workflows and seismic simulation technologies developed within European flagship projects were deployed in a national-scale emergency exercise in collaboration with Mexico's National Seismological Service (SSN) and the Institute of Geophysics of the National Autonomous University of Mexico (UNAM).
  - The drill demonstrated the ability to deliver near real-time seismic simulations and hazard maps for a magnitude 8.1 earthquake scenario.
  - This showcased how European Tier-0 HPC resources can support rapid, life-saving decision-making and deliver actionable information within minutes after a seismic event.
  - This collaboration shows how exascale-ready infrastructures can strengthen societal resilience and international cooperation.

End User Award: Exploring Real-time Tsunami Warning System on World's Fastest Supercomputer

Organizations: Lawrence Livermore National Laboratory, University of Texas at Austin, and Scripps Institution of Oceanography (UC San Diego)

- **Innovation: Coastal communities often have only minutes to evacuate after an offshore earthquake.**
  - Now, a real-time tsunami warning framework can give people precious time to act.
  - By combining sensor data with full-physics modeling powered by the exascale El Capitan supercomputer at Lawrence Livermore National Laboratory (LLNL), researchers from LLNL, the University of Texas at Austin's Oden Institute, and Scripps Institution of Oceanography can predict the arrival and height of tsunami waves in seconds.
  - To help authorities issue earlier and more accurate warnings, this capability models 55.5-trillion-scaled acoustic-gravity wave propagations, reconstructs the seafloor motion caused by earthquakes, and solves billion-parameter-scale inverse problems in less than 0.2 seconds.
  - Together, these unprecedented computations make real-time tsunami forecasting a reality for the first time, incorporating confidence ranges into every forecast so responders can better assess risk with more certainty.

End User Award: Agentic AI for Biomanufacturing Optimization Using Hybrid Quantum–Classical HPC Systems

Organization: ORCA Computing

- **Innovation: ORCA Computing co-led a collaboration to develop Agentic AI, a next-generation artificial intelligence framework that fuses high-performance computing (HPC), photonic quantum processors, and agent-based automation to transform biomanufacturing.**
  - In partnership with the Technical University of Denmark (DTU), Novo Nordisk, and SiC Systems, the team created the Sense-Infer-Control (SIC) platform — a dynamic feedback and control architecture that links digital twin simulations with real-time execution across distributed process environments.
  - ORCA's photonic quantum hardware, integrated with GPU-accelerated HPC nodes via hybrid quantum-classical algorithms, enabled SiC's AI agents to autonomously analyze, infer, and optimize multiscale bioprocesses in real time, orchestrate fragmented sensor networks to maintain optimal conditions, and perform explainable, physics-based decision-making.
  - This approach delivers more consistent yields, lower costs, and improved reproducibility for biologics and chemical manufacturing.

End User Award: Gas Lift Optimization (GLO): Closed Loop, AI Driven Optimization at Permian Field Scale

Organization: Occidental Petroleum (OXY)

- **Innovation: The Gas Lift Optimizer (GLO) changed how Occidental Petroleum (OXY) manages gas lift operations across its oil fields.**
  - By introducing a closed-loop AI-driven system that autonomously adjusts gas injection rates in real time, OXY replaced manual workflows with intelligent automation.
  - GLO represents a first-of-its-kind deployment of scalable, self-optimizing control systems in upstream oil and gas.
  - GLO demonstrates how advanced analytics, digital twins, and real-time feedback can be integrated to make complex industrial systems smarter, safer, and more sustainable.
  - GLO v3 shifted the system from predictive network modeling to measurement-feedback control, cutting optimization run times, as well as onboarding times for new wells and compressors, to minutes.
  - Combined with pressure-aware flare mitigation logic, GLO maximizes production and prevents emissions events, reinforcing OXY's commitment to sustainability.

SC25 2025 Winner:

HPC Vendor Innovation Excellence Award

Vendor Award Winner: NVIDIA Blackwell GPU and Platform

Organization: NVIDIA

Innovation: NVIDIA developed a new GPU that is a major step forward in performance and capabilities.

- The combination of the NVIDIA Blackwell GPU and NVIDIA Grace CPU provides an important advancement for large scale AI applications and is dramatically faster than the previous NVIDIA Hopper GPU.
- It was developed and brought to market in a shorter timeframe than previous generations of GPUs.
  - GPU and CPU performance improvements are very time dependent, so the combination of increased performance and faster time-to-market makes it a major accomplishment.
- Users will be able to run larger AI problems than before and will be able to run models at a considerably faster rate.
  - This should provide a sizable improvement in applying AI to technical computing workloads.
- One measure of the value of the Blackwell breakthrough is the strong market adoption of the GPU, which forms the backbone of many AI systems around the world.

HYPERION RESEARCH

We Welcome Questions, Comments and Suggestions

Please contact us at:

info@hyperionres.com