Hyperion Research SC24 HPC/AI Market Update

November 2024

Earl Joseph, Bob Sorensen, Mark Nossokoff, Tom Sorensen, Melissa Riddle, Jaclyn Ludema, Mike Thorp and Kurt Gantrish

www.HyperionResearch.com
www.hpcuserforum.com

Welcome Message

- **Today's presentation will be sent to all registrants**
  - If unregistered:
    - Registration link is at HyperionResearch.com
    - Or drop your card or leave email at our registration table
- **For follow-up detail or discussions email: info@hyperionres.com**
- **A special thank you to CPC for sponsoring our Breakfast Briefing this morning!**
- **Presentations will end at 8:20am allowing time to attend SC24 keynote; Analysts will remain for Q&A until 9:00am**

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
- Melissa Riddle, Data Analyst
- Thomas Sorensen, Analyst

Executive

- Jean Sorensen, COO

Global Accounts

- Mike Thorp, Sr. Global Sales Executive
- Kurt Gantrish, Sr. Account Executive

Survey Specialist

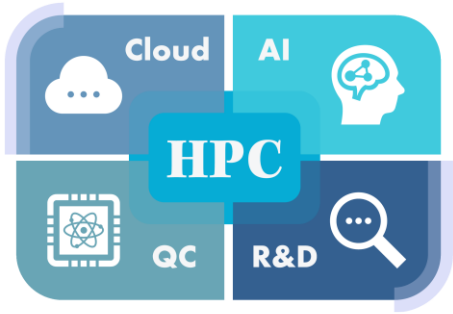
- Cary Sudan, Principal Survey Specialist

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
- Supply Chain Issues
- Sustainability
- Data Center Assessment

Today's Agenda

- **Earl Joseph, CEO**
  - HPC and AI Market Update
  - A New Way of Measuring Value of Leadership Computing
  - Tool for Deeper Understanding of Surveys Results
- **Bob Sorensen, SVP, Chief AI & QC Analyst**
  - Successfully Navigating the Changing Advanced Computing Landscape
- **Mark Nossokoff, Research Director, Chief Cloud & Storage Analyst**
  - Perspectives on HPC-AI Storage and Interconnects
  - HPC-AI Cloud Update
- **Innovation Award Winners Announcement**
- **Conclusions**

HYPERION RESEARCH

HPC/AI Market Update

2024 Looks Like a Strong Growth Year

Hopefully, supply chain issues won't impact installations too much

- **Across the HPC/AI on-premises market, buyers are expecting to increase their purchases by over 22% in 2024**
  - AI-focused on-premises servers are growing at high rate of close to 40% in 2024
  - **We are now tracking non-traditional server suppliers**

- **The HPC cloud market will see strong growth in 2024 -- End user spending on public cloud resources to run HPC/AI workloads is projected to grow over 20% in 2024**
  - Cloud computing is becoming more useful to a larger set of HPC workloads
  - Access to the latest hardware and the ability to quickly setup AI workloads are key drivers
  - This strong growth reflects the heavy work that the cloud service providers (CSPs) have done to make clouds more HPC friendly
  - Users have also gone through extensive work to profile and evaluate where clouds make the most sense

The Overall HPC/AI Market in 2024

2024 HPC/AI Spending is projected to reach $51.9 billion ($US)

- **$25.4 billion in on-premises servers**
- **$8.7 billion in spending to run HPC/AI workloads in the cloud**

The Overall HPC/AI Market in 2024

2024 HPC/AI Spending is projected to reach $51.9 billion ($US)

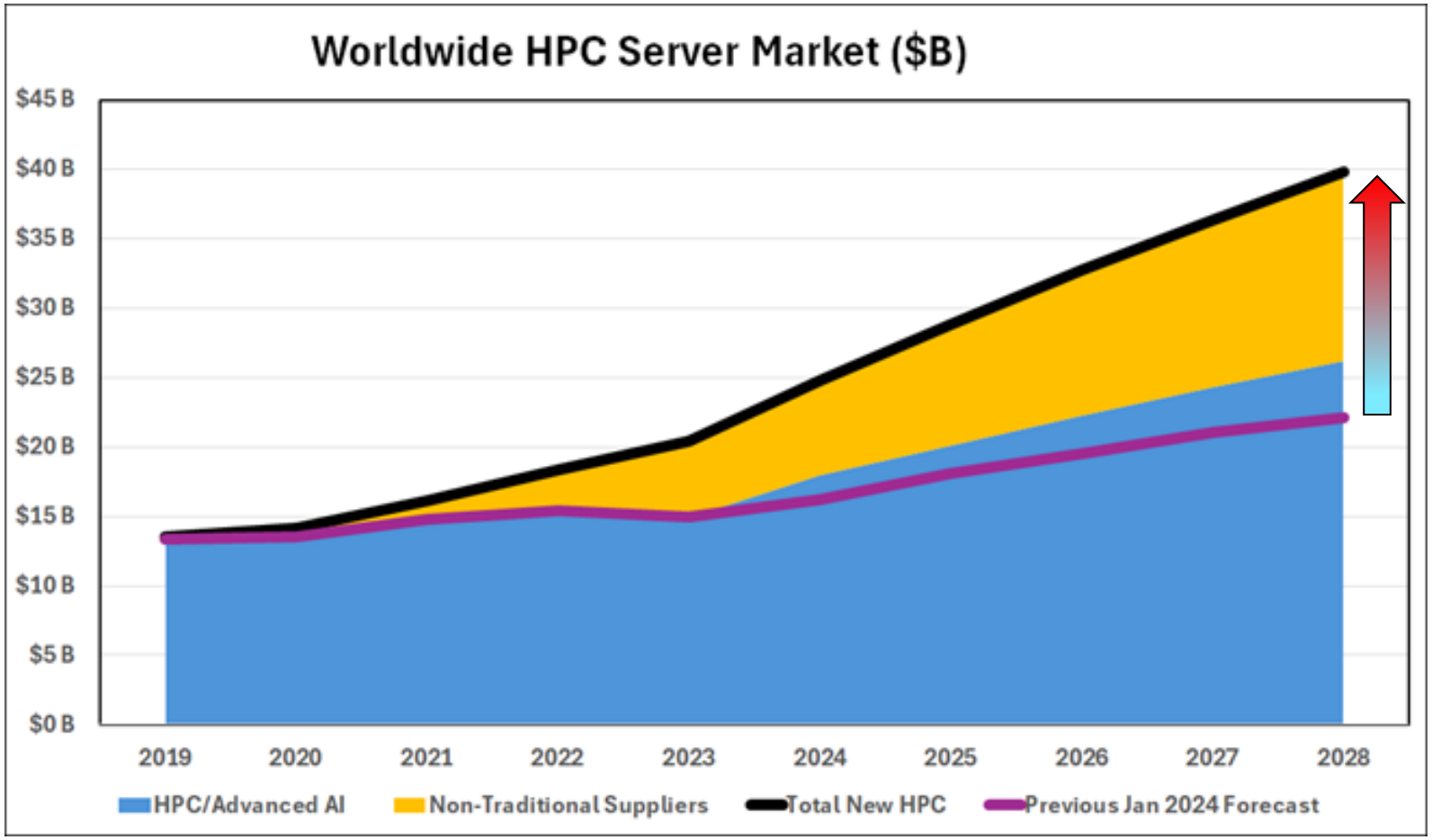
2024 HPC/AI Total Spending

Worldwide Technical Computing/HPC Spending	
	2024
Traditional HPC/AI Suppliers	$17.9
Non-Traditional Suppliers	$7.5
Storage, Software, Service	$17.8
HPC Cloud Spending	$8.7
Total HPC/AI	**$51.9**
Source: Hyperion Research, Oct. 2024	

- **$25.4 billion in on-premises servers**
- **$8.7 billion in spending to run HPC/AI workloads in the cloud**

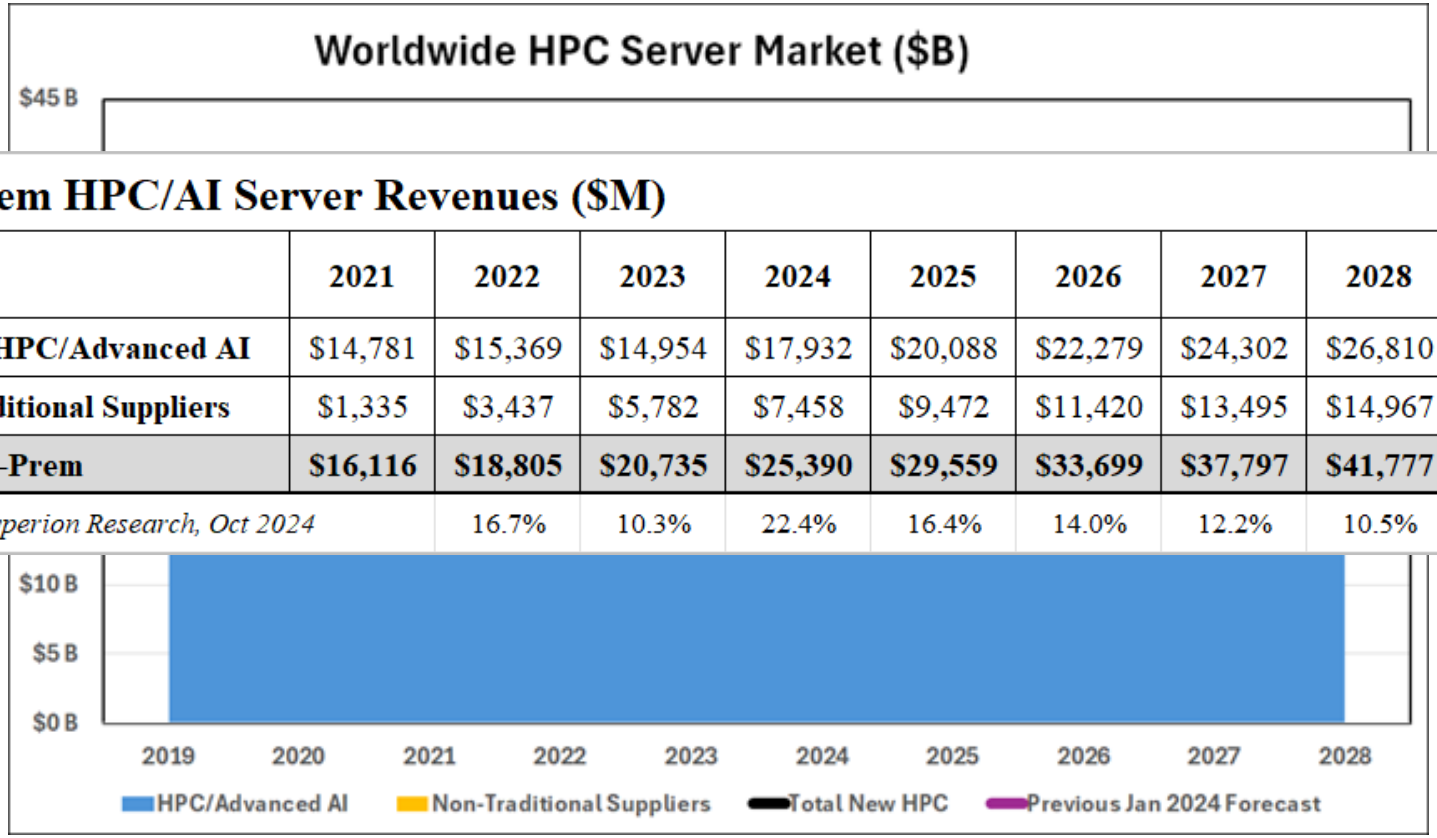
Updated View of the On-Prem Sever Market

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



Updated View of the On-Prem Sever Market

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



On-Prem HPC/AI Server Revenues ($M)

	2021	2022	2023	2024	2025	2026	2027	2028	CAGR 23-28
Historic HPC/Advanced AI	$14,781	$15,369	$14,954	$17,932	$20,088	$22,279	$24,302	$26,810	12.4%
Non-Traditional Suppliers	$1,335	$3,437	$5,782	$7,458	$9,472	$11,420	$13,495	$14,967	21.0%
Total On-Prem	**$16,116**	**$18,805**	**$20,735**	**$25,390**	**$29,559**	**$33,699**	**$37,797**	**$41,777**	**15.0%**
Source: Hyperion Research, Oct 2024		16.7%	10.3%	22.4%	16.4%	14.0%	12.2%	10.5%	

HYPERION RESEARCH

Tipping Point Examples

Tipping Points in HPC

A "Tipping Point" is when an event or threshold is reached that causes the market to change

In HPC, tipping points often happen when:

- Compute capabilities reach a point where they are capable of doing something that wasn't possible before

They are often driven by:

- An increase in computing power that allows computers to do an important task
- In some cases, the computer cost drops so much that a larger set of users can afford to apply them to important problems
- It also requires a strong data set and software as well as experts who create and apply the applications

Tipping Points: Early 1980's -- Successful Crashing Cars via Digital Simulation

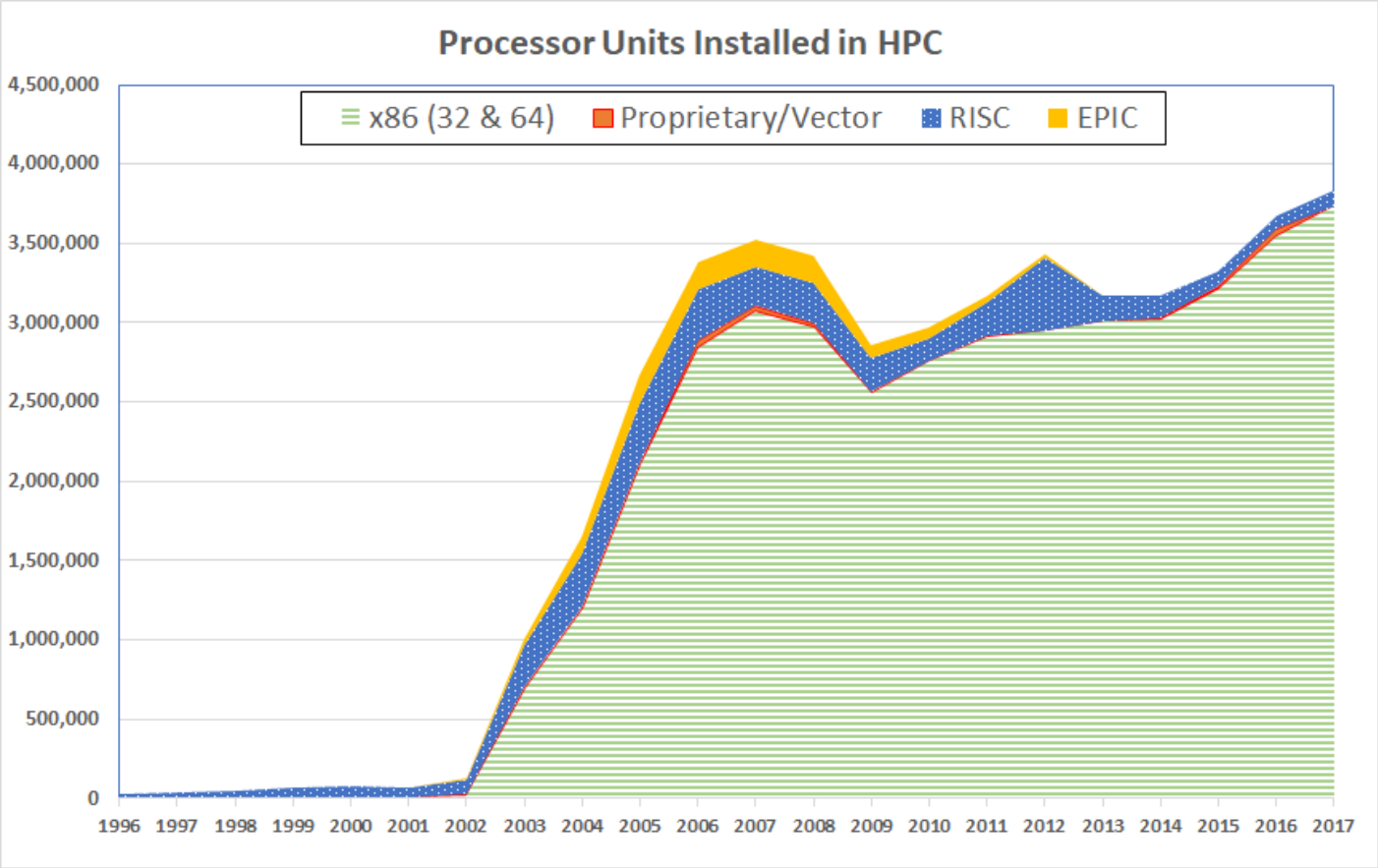
All major manufactures switched to HPC in just a few years

- **No more building an actual prototype for each design test**
- **New designs could be developed and tested in a fraction of the time that physical models required**
- **Provided major cost savings AND the testing of orders-of-magnitude more design models**
  - **Resulted in dramatically better car designs**

Play videos 1 & 2

Tipping Points: Early 2000's -- When x86 Became "Good Enough" for Many Jobs

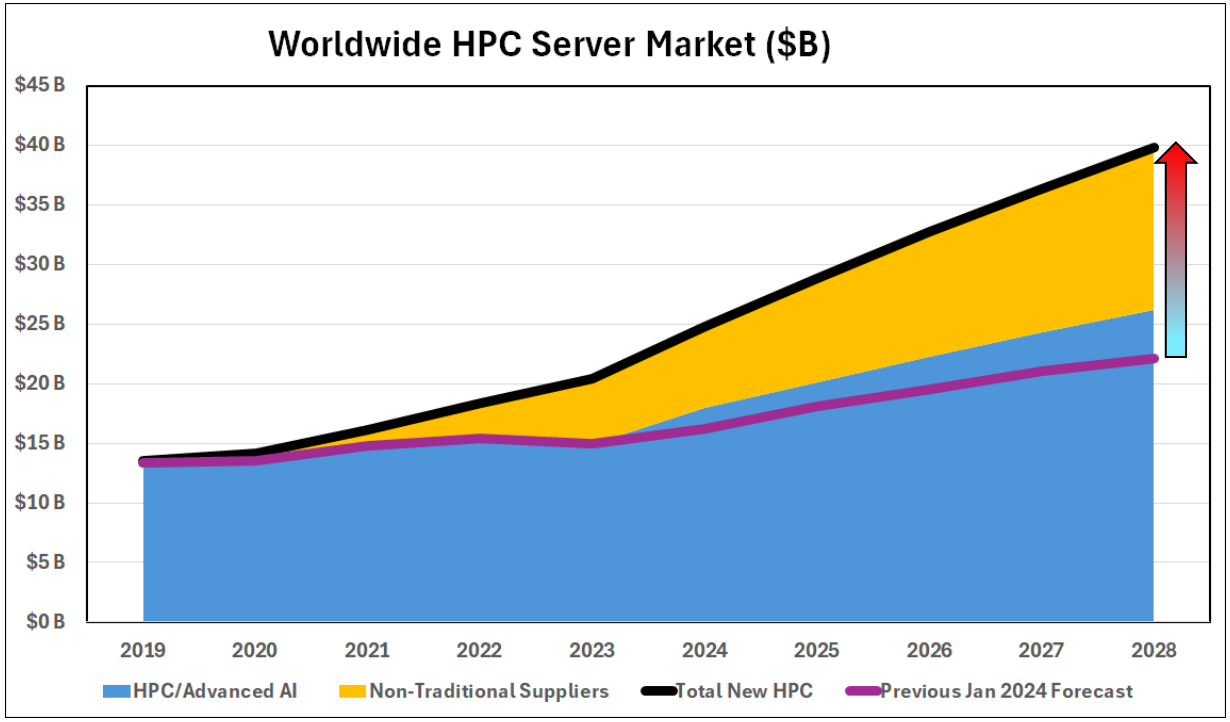
It both replaced other CPUs and launched major market growth



- **And it was ~100x cheaper**

Tipping Points: November 2022 -- When ChatGTP Showed the World That AI Works

On-Prem HPC/AI servers went from a $15 billion market growing at 7%, to being a $21 billion market growing at 15%



- **It required:**
  - **Computing at scale**
  - **Major data sets and data capabilities**
  - **The work of some very insightful people**

HYPERION RESEARCH

Vendor Shares

2023 On-prem Vendor Shares

Total HPC Market: $20.6 billion (US$), $15 billion from traditional suppliers, $5.5 billion from non-traditional suppliers

	Full Year 2023 Revenues ($M)	Traditional HPC Suppliers Market Shares	New HPC 2023 Market Shares
HPE	4,712	31.3%	22.9%
Dell Technologies	3,659	24.3%	17.8%
Lenovo	1,268	8.4%	6.2%
Inspur	1,041	6.9%	5.1%
Sugon	580	3.9%	2.8%
Atos	528	3.5%	2.6%
IBM	405	2.7%	2.0%
Penguin	385	2.6%	1.9%
Fujitsu	218	1.4%	1.1%
NEC	197	1.3%	1.0%
Other HPC	2,075	13.8%	10.1%
Non-Traditional Suppliers	5,489	--	26.7%
Total	20,557		100.0%

Hyperion Research, October 2024

HYPERION RESEARCH

Cloud Spending

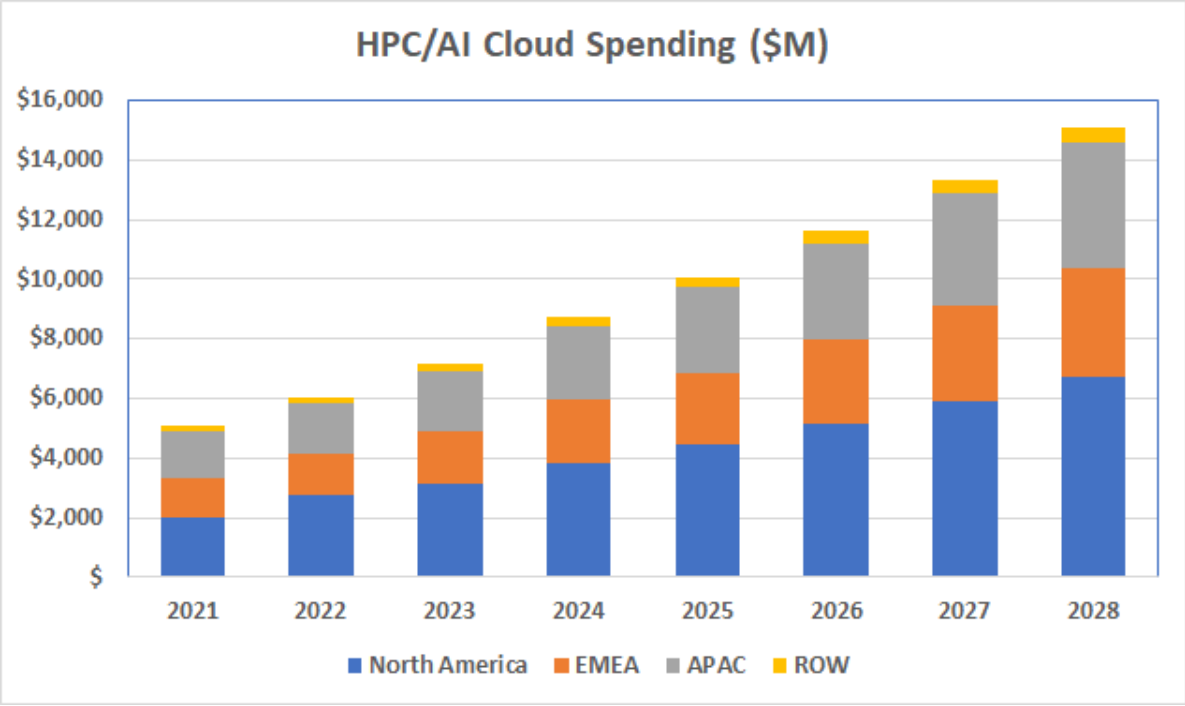
Cloud Study Findings

Based on overall runtime, approximately what percentage of all your HPC workloads are run on external clouds TODAY?	Response	Count of Response	%
	None	6	7.1%
	1% to less than 5%	7	8.3%
	5% to less than 10%	9	10.7%
	10% to less than 15%	16	19.0%
	15% to less than 20%	9	10.7%
	20% to less than 25%	8	9.5%
	25% to less than 35%	7	8.3%
	35% to less than 50%	5	6.0%
	50% to less than 75%	8	9.5%
	75% to less than 95%	4	4.8%
	95% or more	3	3.6%

Please distribute your total HPC/AI/HPDA cloud resource spending between the following categories. Must sum to 100%	Response		Average %
	% Compute instances:		46.6%
	% Ephemeral storage:		13.8%
	% Persistent storage:		16.2%
	% System SW (e.g., file systems, databased):		11.0%
	% Application SW:		11.0%
	% Other:		1.3%

Cloud 5-year Forecast by Region

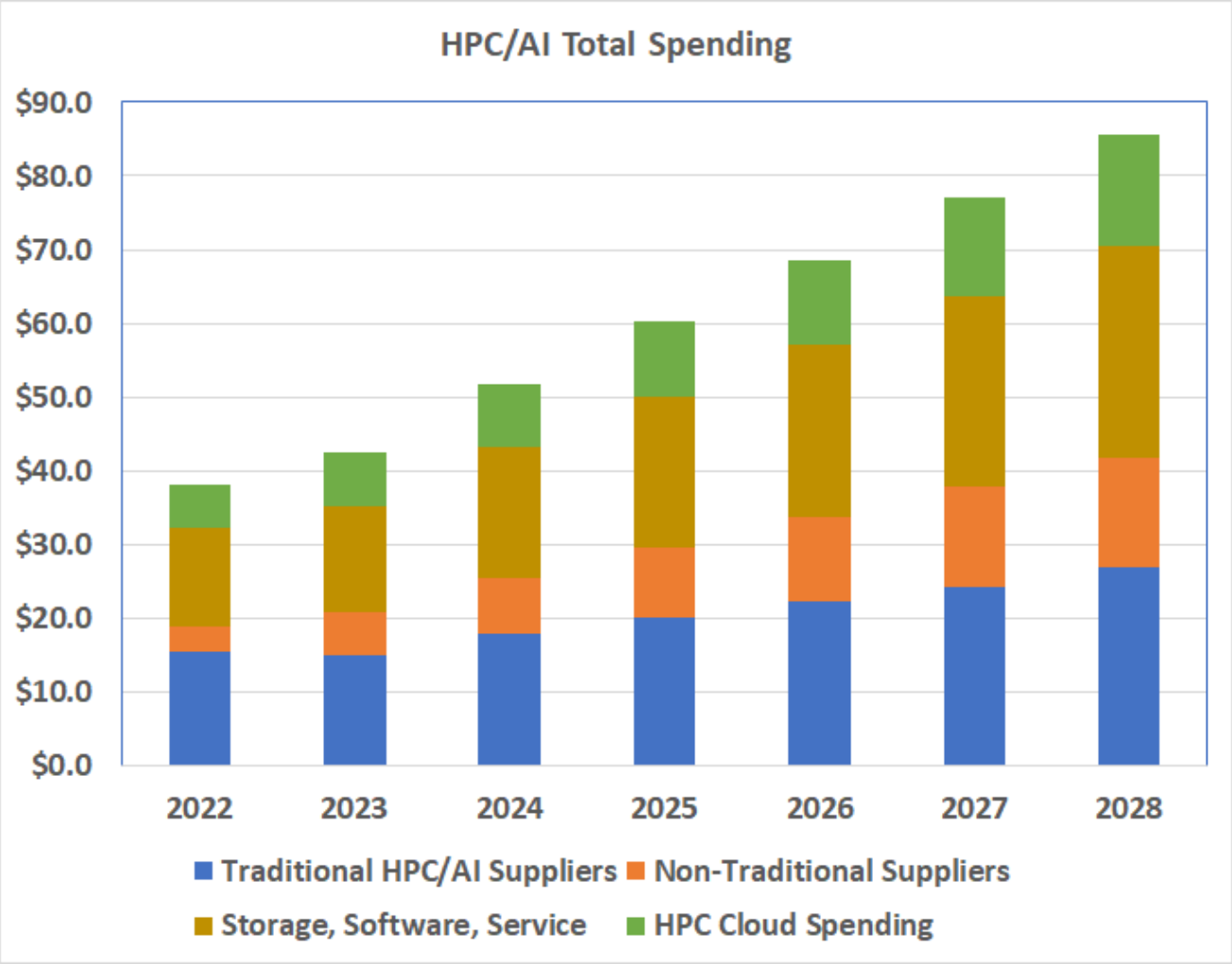
Projected to reach $15.1 billion (US$) by 2028 (16% CAGR)



HPC/AI Cloud Spending by Region ($M)									
	2021	2022	2023	2024	2025	2026	2027	2028	CAGR 23-28
North America	$2,031	$2,770	$3,150	$3,852	$4,454	$5,170	$5,926	$6,734	16.4%
EMEA	$1,315	$1,401	$1,727	$2,093	$2,420	$2,779	$3,186	$3,620	16.0%
APAC	$1,568	$1,671	$2,059	$2,471	$2,857	$3,265	$3,742	$4,252	15.6%
ROW	$186	$198	$243	$295	$342	$391	$448	$509	15.9%
Total HPC Cloud Spending	**$5,100**	**$6,040**	**$7,180**	**$8,711**	**$10,072**	**$11,605**	**$13,302**	**$15,115**	**16.1%**
Source: Hyperion Research, Oct. 2024									

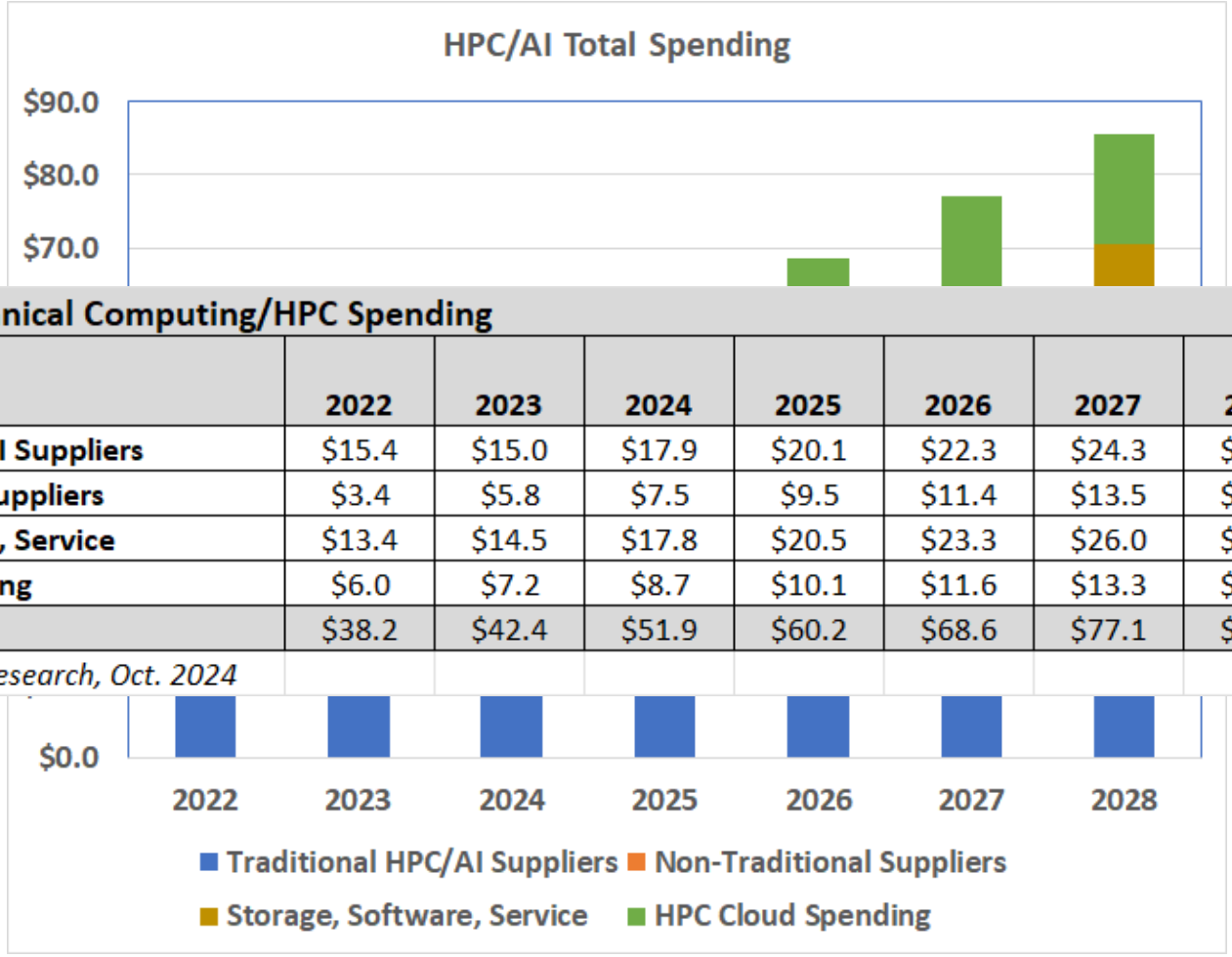
The HPC/AI 5-year Forecast: On-Prem Plus Cloud Spending

Projected to reach $85.5 billion (US$) by 2028



The HPC/AI 5-year Forecast: On-Prem Plus Cloud Spending

Projected to reach $85.5 billion (US$) by 2028



Worldwide Technical Computing/HPC Spending								
	2022	2023	2024	2025	2026	2027	2028	CAGR 23-28
Traditional HPC/AI Suppliers	$15.4	$15.0	$17.9	$20.1	$22.3	$24.3	$26.8	12.4%
Non-Traditional Suppliers	$3.4	$5.8	$7.5	$9.5	$11.4	$13.5	$15.0	21.0%
Storage, Software, Service	$13.4	$14.5	$17.8	$20.5	$23.3	$26.0	$28.6	14.5%
HPC Cloud Spending	$6.0	$7.2	$8.7	$10.1	$11.6	$13.3	$15.1	16.1%
Total HPC/AI	$38.2	$42.4	$51.9	$60.2	$68.6	$77.1	$85.5	**15.0%**

Source: Hyperion Research, Oct. 2024

The Exascale Market (System Acceptances)

Over 45 systems and over $13 billion in value

Exascale and Near-Exascale Leadership Systems (2020 to 2028)

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	1 exascale system ~$350M	1 or 2 pre-exascale systems ~$150M each	1 near-exascale system ~$150M	Remaining 1/3 of Frontier system	--	4-5	~$1.0B
2024	1 exascale system ~$350M	2 or 3 pre-exascale systems ~$150M each	?	1 exascale system ~$600M	1 pre-exascale system ~$125M	5-6	~$1.3B
2025	1 or 2 exascale systems ~$300M each	2 exascale systems ~$350M each	?	2 exascale system ~$600M	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 exascale system ~$200M	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	8-11	$1.7B - $2.6B
Total	**12-13**	**14-19**	**5-6**	**8-11**	**7-10**	**47-61**	**$12.5B - $15.9B**

* 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, October 2024

HYPERION RESEARCH

Measuring The Value Of Leadership Computing

Showing the Value of Leadership Computing

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

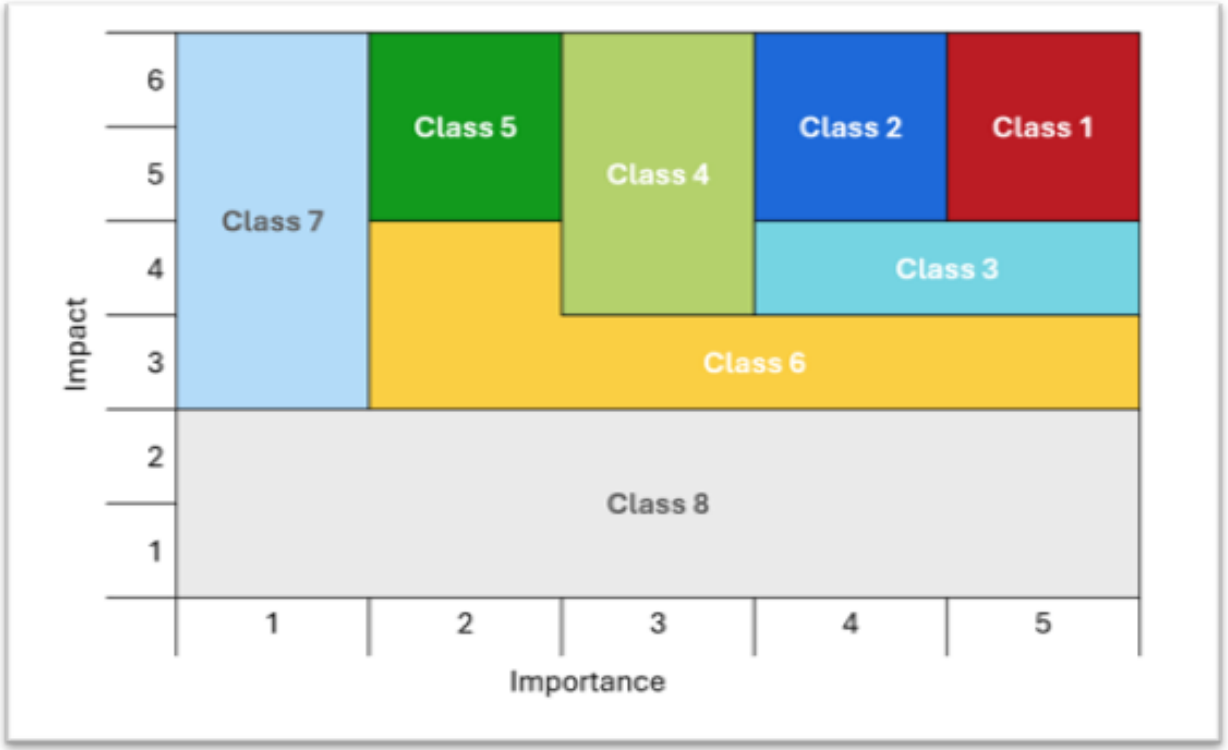
- Class 1 innovations – One of the top 1-3 innovations in a field over the last ten years PLUS useful to over 10 organizations
- Class 2 innovations -- One of the top 5 innovations in a field over the last ten years PLUS useful to over 10 organizations
- Class 3 innovations – One of the top 5 innovations in a field over the last ten years PLUS useful to over 5 organizations
- Class 4 innovations – One of the top 10 innovations in a field over the last ten years PLUS useful to over 5 organizations
- Class 5 innovations – One of the top 25 innovations in a field over the last ten years PLUS useful to over 10 organizations
- Class 6 innovations – One of the top 25 innovations in a field over the last ten years PLUS useful to at least 2 organizations
- Class 7 innovations – One of the top 50 innovations in a field over the last ten years PLUS useful to at least 2 organizations
- Class 8 innovations – All other innovations

A New Way to Show the Value of Leadership Computing

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

FIGURE 1

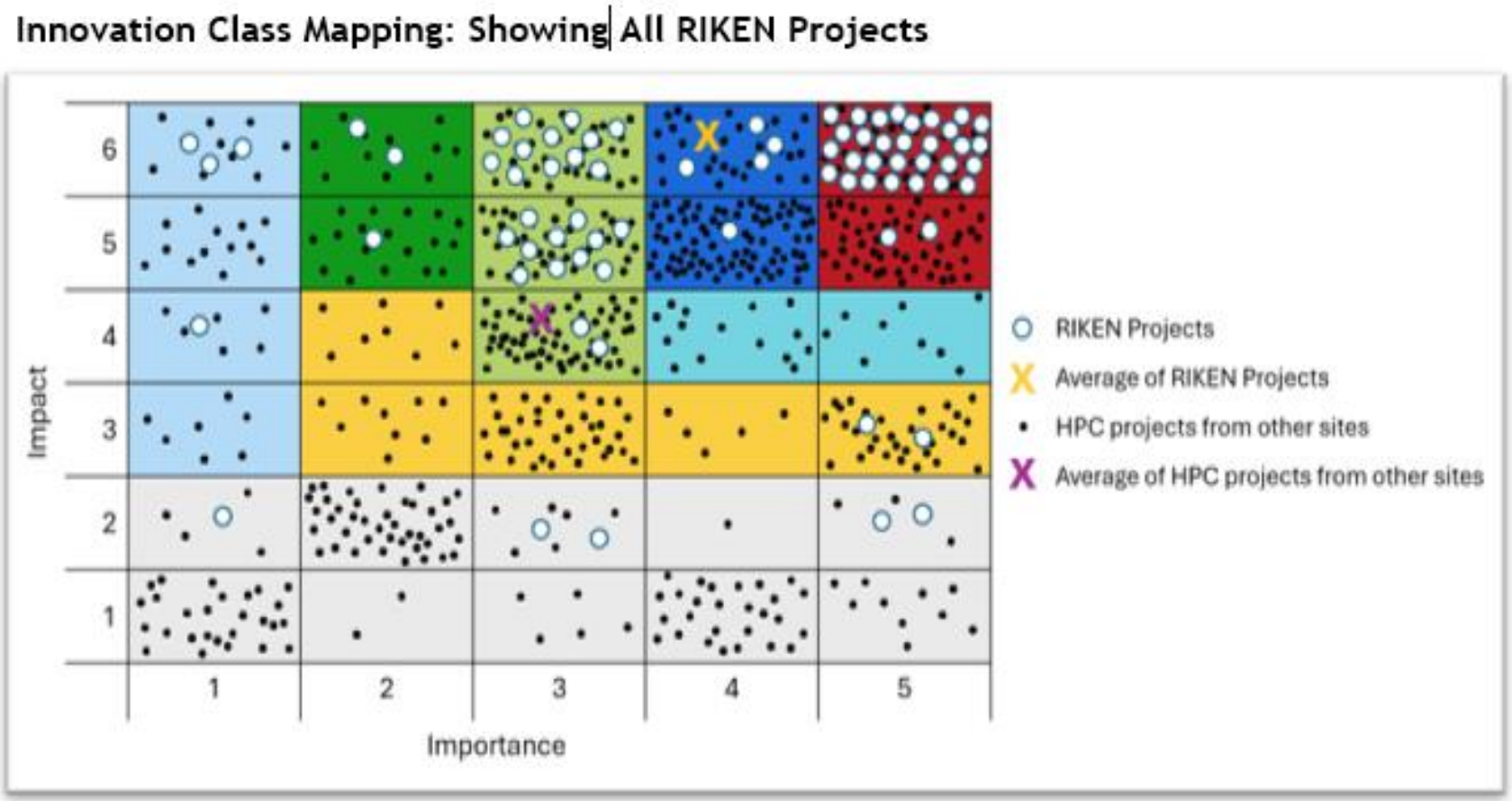
Innovation Class Map



Source: Hyperion Research, 2024

A New Way to Show the Value of Leadership Computing

An example from a 2024 study compared to 650 other projects



Source: Hyperion Research, 2024

HYPERION RESEARCH

The Hyperion Research AI Advisory Committee

We Invite You to Join the Hyperion Research AI Advisory Committee

AI Advisory Committee Mission:

- To support the worldwide AI/HPC community by helping to answer key questions about how AI is evolving
  - Guiding Hyperion Research AI-related studies by identifying the most important questions to be explored and topics to be researched
  - A major portion of study findings will be shared with the broader AI/HPC community
- To share ideas, best practices, and areas of concern between Committee members

Overview of the AI Advisory Committee:

- Members are from all areas of the AI ecosystem: users, vendors, CSPs, academic experts, etc.
  - The Committee is worldwide
  - There is no cost for membership
- Members receive the ability to help guide Hyperion Research studies with questions that are important to Committee members
  - Members receive the results before the broader AI/HPC community

HYPERION RESEARCH

Some Recent AI Study Results

AI Study Findings

- ***Within AI workloads, training is using more CPU hours today***
- ***Clouds are heavily used for AI training and traditional HPC***

Please approximate the portion of your AI/ML/DL/LLM workload that is training compared to inferencing (based on CPU hours):	**Response**	**Count of Response**	**%**
	100% training and 0% inferencing	**5**	**4.9%**
	90% training and 10% inferencing	**20**	**19.4%**
	75% training and 25% inferencing	**38**	**36.9%**
	50% training and 50% inferencing	**24**	**23.3%**
	25% training and 75% inferencing	**11**	**10.7%**
	10% training and 90% inferencing	**4**	**3.9%**
	0% training and 100% inferencing	**1**	**1.0%**

Of all your workloads in your HPC/AI/HPDA cloud environment, please distribute utilization time by the following: Must sum to 100%.	**Response**		**Average %**
	% Traditional Modeling and Simulation		**25.4%**
	% Traditional Data Science (HPDA, Big Data) (excluding AI/ML/DL) Workloads:		**21.0%**
	% AI - training:		**25.5%**
	% AI - inferencing:		**17.6%**
	% AI – Other:		**3.2%**
	% Quantum:		**5.2%**
	% Other		**2.1%**

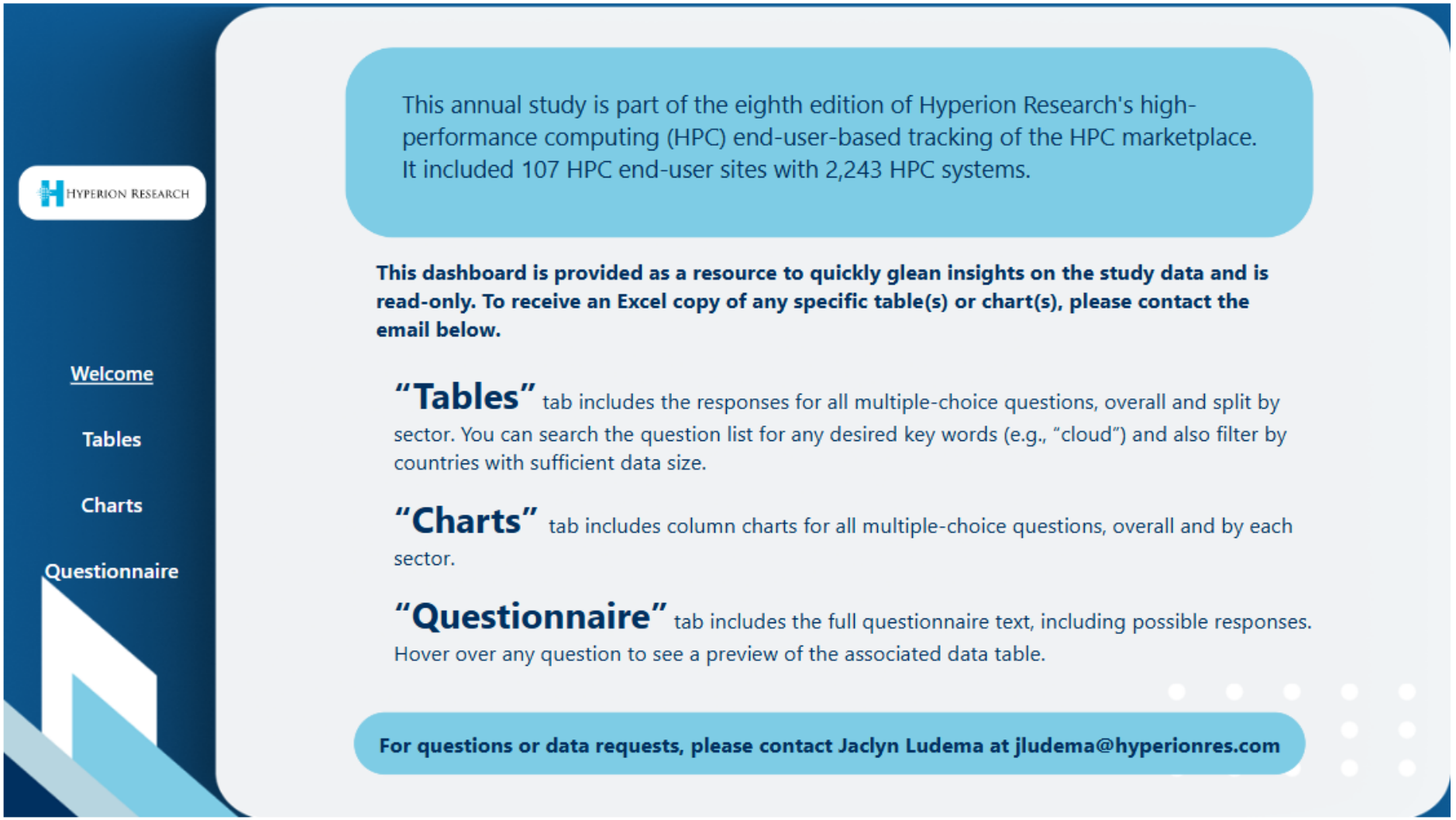
HYPERION RESEARCH

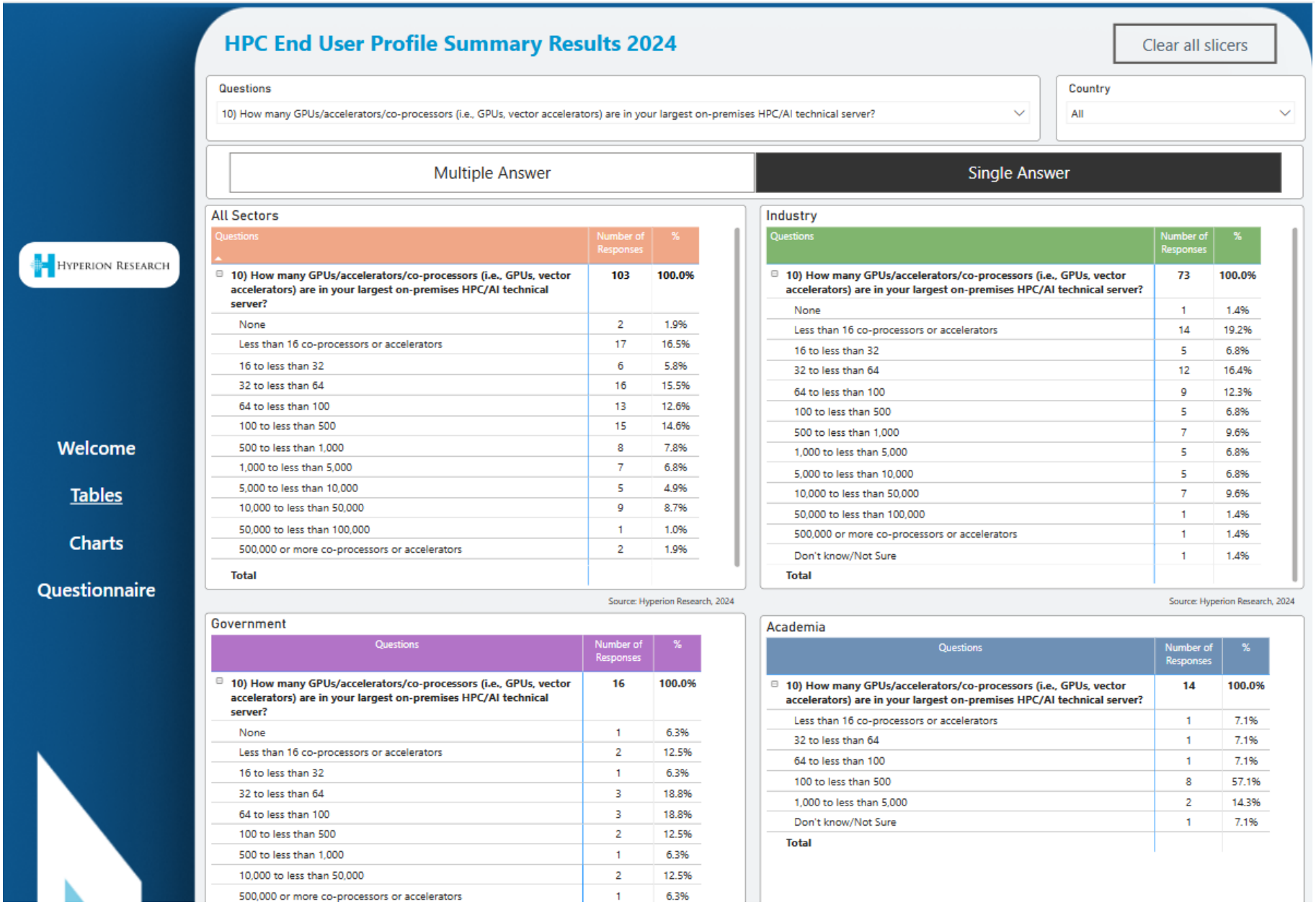
A New Tool
For Data Analysis:

For Our 2024 MCS Study

The New Hyperion Research Data Tool

To help find important findings more quickly





HPC End User Profile Summary Results 2024

Clear all slicers

Questions: 10) How many GPUs/accelerators/co-processors (i.e., GPUs, vector accelerators) are in your largest on-premises HPC/AI technical server?

Country: All

Multiple Answer | Single Answer

HYPERION RESEARCH

Welcome

Tables

Charts

Questionnaire

All Sectors

Questions	Number of Responses	%
10) How many GPUs/accelerators/co-processors (i.e., GPUs, vector accelerators) are in your largest on-premises HPC/AI technical server?	**103**	**100.0%**
None	2	1.9%
Less than 16 co-processors or accelerators	17	16.5%
16 to less than 32	6	5.8%
32 to less than 64	16	15.5%
64 to less than 100	13	12.6%
100 to less than 500	15	14.6%
500 to less than 1,000	8	7.8%
1,000 to less than 5,000	7	6.8%
5,000 to less than 10,000	5	4.9%
10,000 to less than 50,000	9	8.7%
50,000 to less than 100,000	1	1.0%
500,000 or more co-processors or accelerators	2	1.9%
Total		

Source: Hyperion Research, 2024

Industry

Questions	Number of Responses	%
10) How many GPUs/accelerators/co-processors (i.e., GPUs, vector accelerators) are in your largest on-premises HPC/AI technical server?	**73**	**100.0%**
None	1	1.4%
Less than 16 co-processors or accelerators	14	19.2%
16 to less than 32	5	6.8%
32 to less than 64	12	16.4%
64 to less than 100	9	12.3%
100 to less than 500	5	6.8%
500 to less than 1,000	7	9.6%
1,000 to less than 5,000	5	6.8%
5,000 to less than 10,000	5	6.8%
10,000 to less than 50,000	7	9.6%
50,000 to less than 100,000	1	1.4%
500,000 or more co-processors or accelerators	1	1.4%
Don't know/Not Sure	1	1.4%
Total		

Source: Hyperion Research, 2024

Government

Questions	Number of Responses	%
10) How many GPUs/accelerators/co-processors (i.e., GPUs, vector accelerators) are in your largest on-premises HPC/AI technical server?	**16**	**100.0%**
None	1	6.3%
Less than 16 co-processors or accelerators	2	12.5%
16 to less than 32	1	6.3%
32 to less than 64	3	18.8%
64 to less than 100	3	18.8%
100 to less than 500	2	12.5%
500 to less than 1,000	1	6.3%
10,000 to less than 50,000	2	12.5%
500,000 or more co-processors or accelerators	1	6.3%

Academia

Questions	Number of Responses	%
10) How many GPUs/accelerators/co-processors (i.e., GPUs, vector accelerators) are in your largest on-premises HPC/AI technical server?	**14**	**100.0%**
Less than 16 co-processors or accelerators	1	7.1%
32 to less than 64	1	7.1%
64 to less than 100	1	7.1%
100 to less than 500	8	57.1%
1,000 to less than 5,000	2	14.3%
Don't know/Not Sure	1	7.1%
Total		

HPC End User Profile Summary Results 2024

Clear all slicers

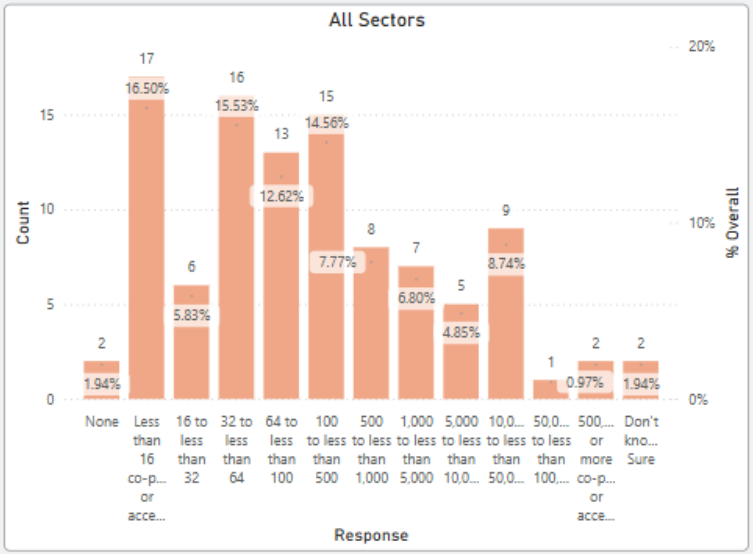
Questions

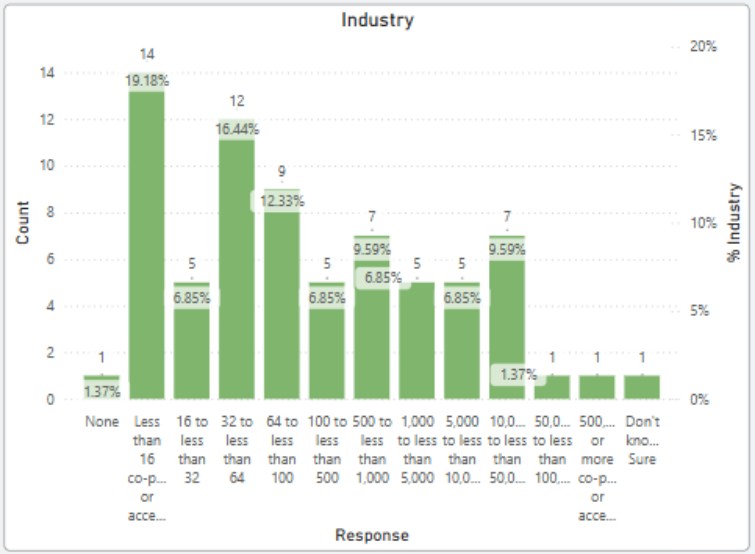
10) How many GPUs/accelerators/co-processors (i.e., GPUs, vector accelerators) are in your largest on-premises HPC/AI technical server?

Country

All

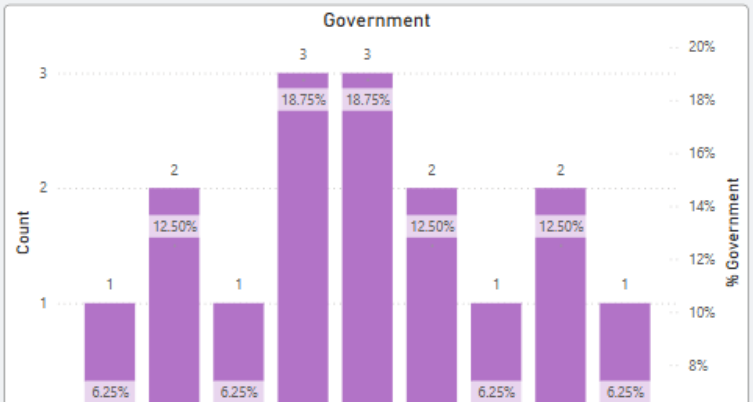
Multiple Answer | Single Answer

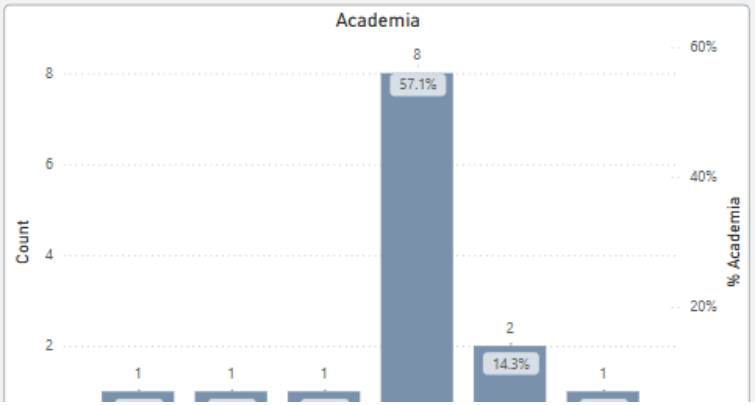




Source: Hyperion Research, 2024

Source: Hyperion Research, 2024





Welcome

Tables

Charts

Questionnaire

HPC End User Profile Summary Results 2024

HYPERION RESEARCH

Welcome

Tables

Charts

Questionnaire

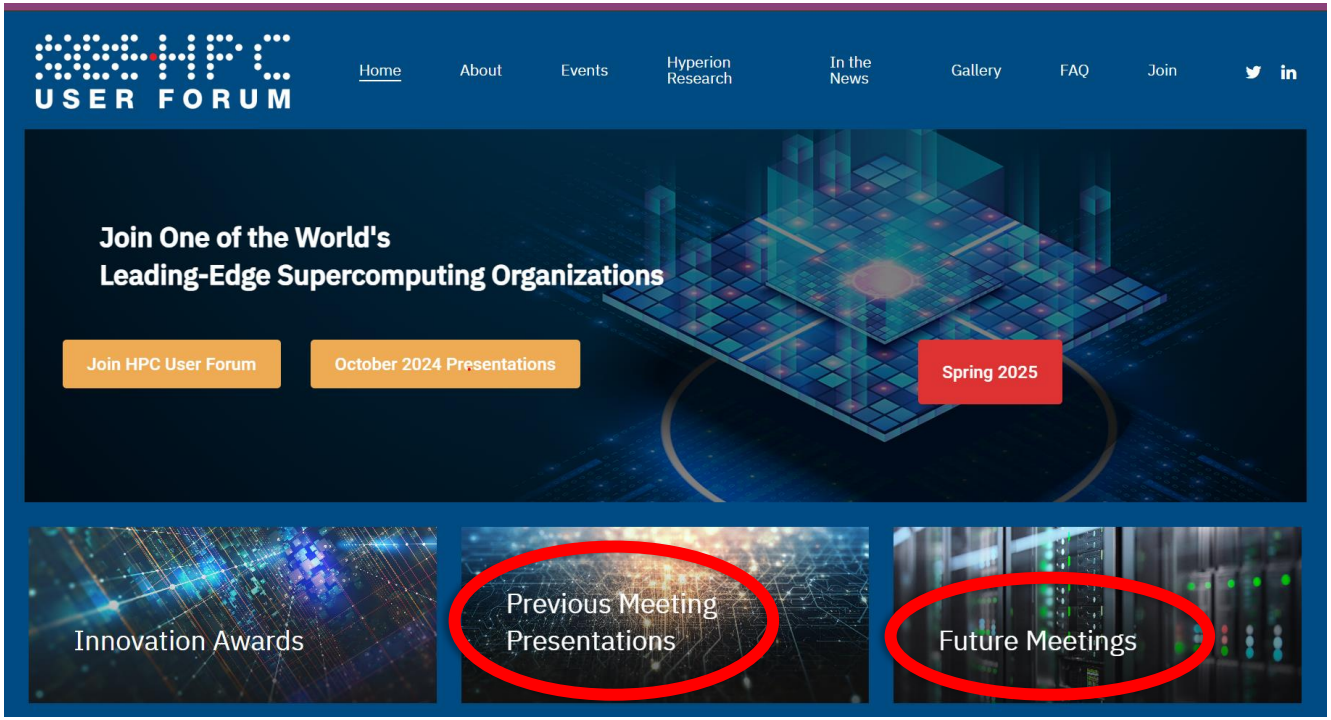
Questions

Search

Questions	Responses
1) In what sector is your organization in?	a) Government b) Academia c) Industry
1.b) Please specify which industry:	i. Bio & Life sciences, pharmaceutical, biological, life sciences, healthcare, drug discovery, bioinformatics, genomics, etc. ii. CAE, manufacturing, e.g., aerospace, automotive, consumer products, etc. iii. Chemical engineering, chemical design, development, and production iv. Mechanical design e.g., CAD v. DCC, entertainment, digital content creation, 3D animation, advanced graphics, gaming, visualization, etc. vi. Financial or economic modeling, pricing, risk management, modeling, business intelligence, etc. vii. EDA, electronic design and analysis viii. IT, computers, HPC systems, IT services, ISV, software company, cloud provider, etc. ix. Geosciences, energy, petroleum, oil and gas, seismic, reservoir simulation, alternative energy, power distribution, etc. x. Weather/climate xi. Transportation and logistics, traffic management, pattern recognition, linear programming, etc. xii. Retail, marketing, and related BI xiii. Telecommunications xiv. Other (please specify)
2) How many on-premises HPC/AI technical server systems (number of clusters, SMP systems, etc.) does your organization have? Please include all HPC/AI systems/servers. By "server" we mean a full system, not the individual nodes.	a) None – we only use external clouds b) 1 c) 2 to 4 d) 5 to 7 e) 8 to 10 f) 11 to 12 g) 13 to 15 h) 16 to 20 i) 21 to 25 j) 26 to 30 k) 31 to 40 l) 41 to 50 m) 51 to 75 n) 76 to 100 o) More than 100 HPC/AI systems
3) What is the approximate Peak Performance (in PF) of your LARGEST SYSTEM:	a) Less than 0.5 petaflops b) 0.5 to less than 1 petaflop c) 1 to less than 5 petaflops d) 5 to less than 10 petaflops e) 10 to less than 25 petaflops f) 25 to less than 50 petaflops g) 50 to less than 100 petaflops h) 100 to less than 250 petaflops i) 250 to less than 500 petaflops j) 500 to less than 750 petaflops k) 750 to less than 1,000 petaflops l) 1,000 to less than 2,500 petaflops m) 2,500 petaflops or greater

HPC User Forum: Recent Presentations and Upcoming Meetings

- **Find previous meeting presentations at:**
  https://www.hpcuserforum.com/hpc-user-forum-presentations/



- **The next HPC/AI User Forum meeting will be:**
  - April 8-9, 2025
  - At La Fonda on the Plaza, Santa Fe, New Mexico

Today's Agenda

- **Earl Joseph, CEO**
  - HPC and AI Market Update
  - A New Way of Measuring Value of Leadership Computing
  - Tool for Deeper Understanding of Surveys Results
- **Bob Sorensen, SVP, Chief AI & QC Analyst**
  - Successfully Navigating the Changing Advanced Computing Landscape
- **Mark Nossokoff, Research Director, Chief Cloud & Storage Analyst**
  - Perspectives on HPC-AI Storage and Interconnects
  - HPC-AI Cloud Update
- **Innovation Award Winners Announcement**
- **Conclusions**

Successfully Navigating the Changing Advanced Computing Landscape

SC24 Breakfast Briefing

Bob Sorensen
Senior Vice President for Research
Hyperion Research, LLC

www.HyperionResearch.com
www.hpcuserforum.com

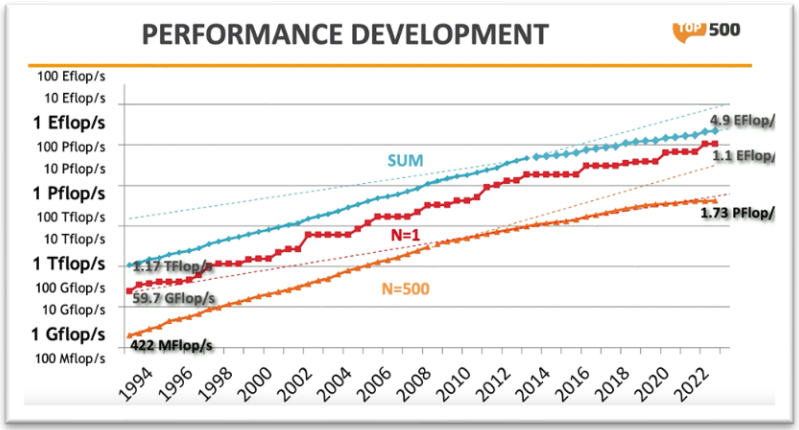
The HPC Landscape is Undergoing Significant Changes...Again

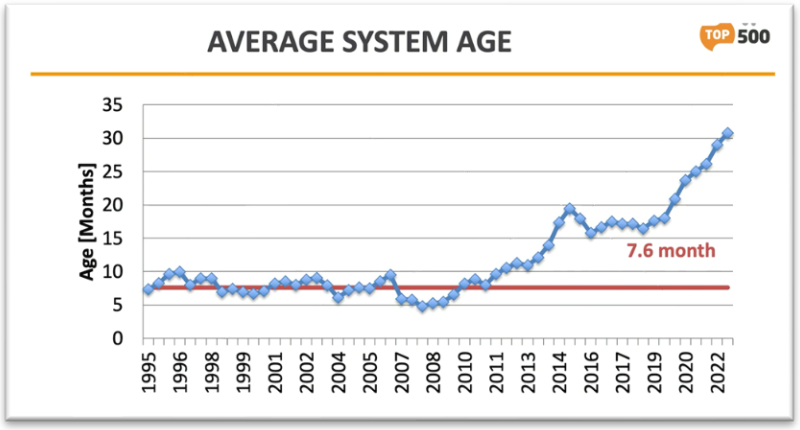
HPC becomes advanced computing in many forms

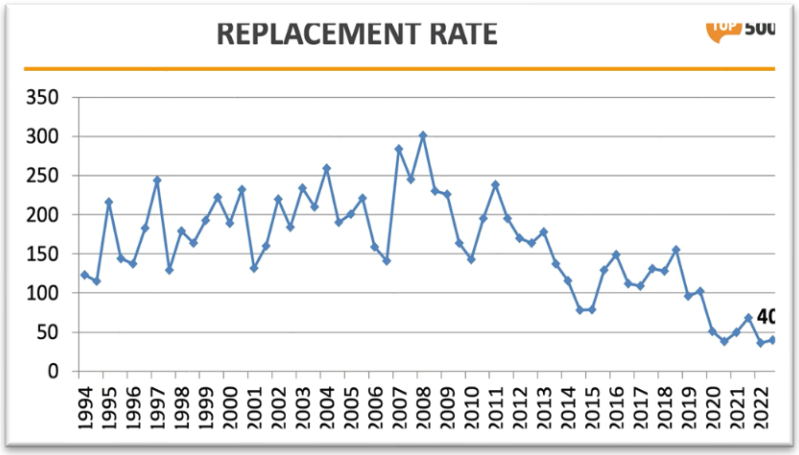
- Multiple AI developments are significantly altering the overall HPC landscape
  - Generative AI is a near-term game changer (for some)
  - ML/DL growing relevance (influence) in S&E world
  - AI-centric GPUs and accelerators dominate processor design mindshare
- Computing at the edge
  - Supports real-time decision making and processing at the data collection point
- CSP's HPC-related impact
  - Offers instance variants vs on-prem permanence
  - Increasingly drives advanced computing ecosystem (and supply chain)
    - Redefining traditional HPC supplier base
- End user decisions for on-premises vs. cloud vs. hybrid
- Dark HPCs on the horizon
- Quantum computing making strides
  - 3-4 years out, but many exploring options now
  - Quantum/HPC integration is the next big thing
- Traditional modeling and simulation remain crucial...but issues loom

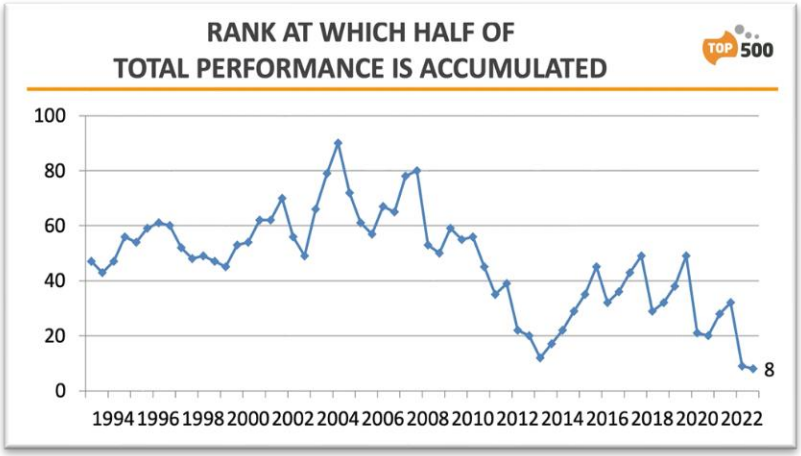
Traditional HPC: Reality Behind the Numbers

Slower progress, less turnover, a sectoral schism



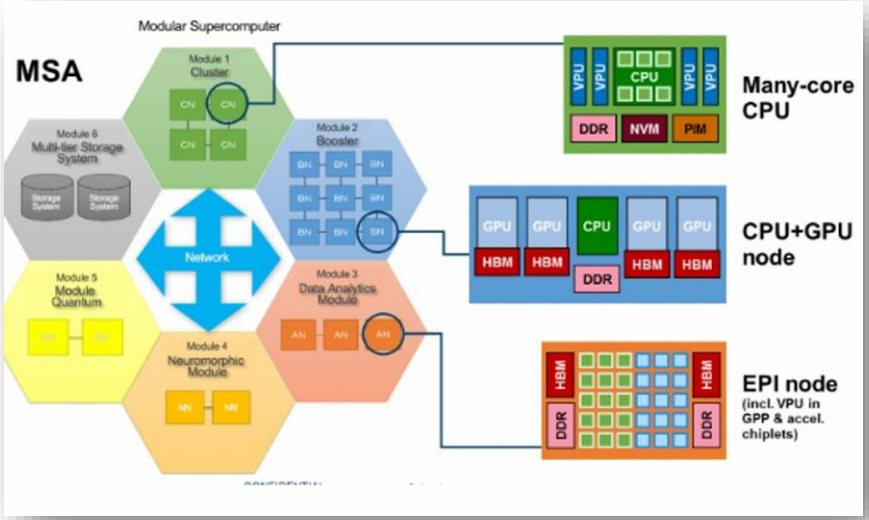






Two Visions For Exascale Systems

Is Jupiter the prototype for neo exascale systems?

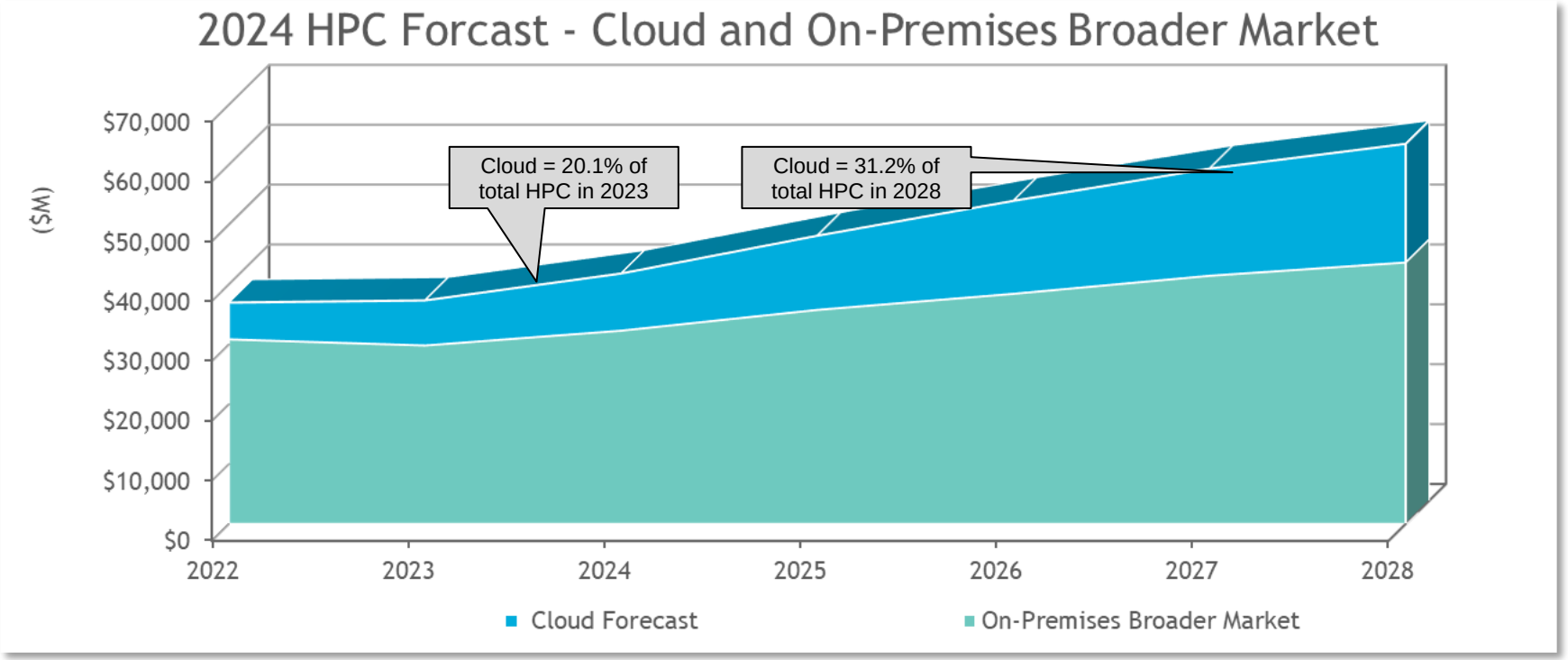


- Argonne National Lab's Newest Exascale System
- 1.98 EFlops peak performance
- Intel was prime on project
- 10,624 nodes: each node:
  - 2 Intel Xeon Max Series (Sapphire Rapids)
  - 6 Intel Data Center GPU Max (Ponte Vecchio)
- 9,264,128 cores total

- Jülich Supercomputing Centre, Germany
- Eviden's Bull Sequana XH3000 technology
- Partnered with ParTec AG's Module System Architecture
- Cluster module: 1300 nodes SiPearl Arm-based Rhea Processors
- Booster Module: 600 nodes Grace-Hopper superchips
- EPI Node: SiPearl Cronos and others?

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

The cloud market approaching 1/3 of total market in 2028



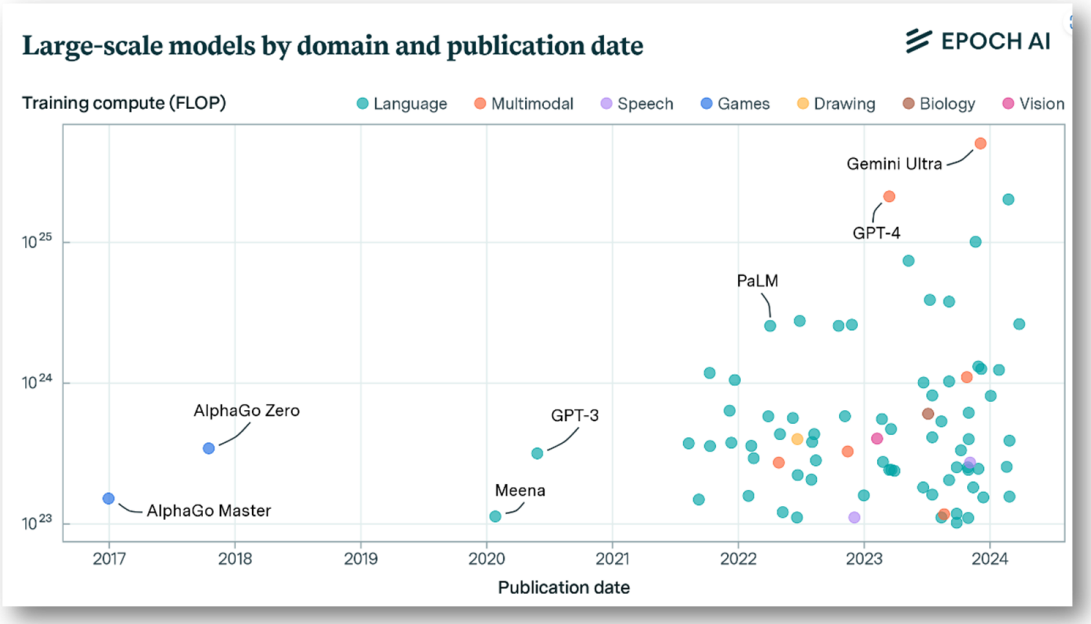
Large Language Models (LLMs) and HPC

Focus on the most demanding, recent, and provocative AI space

- Language–centric class of foundation models, so called to underscore their critically central yet incomplete character
  - language ~= text
- AI writ large is undergoing a paradigm shift with the rise of LLM models (e.g., Claude, Ernie, Gemini, DALL-E, GPT-4++) trained on broad data sets (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks
- LLMs have demonstrated uses in language, vision, robotic manipulation, reasoning, human interaction
- LLMs are based on standard deep learning and transfer techniques where knowledge learned in one realm that transfers to another
- Their scale results in new and sometimes unexplainable emergent capabilities
  - Hallucinations (confabulation)

LLMs Consume Significant FLOPs

LLM flops growth eclipses Top 500 growth



- Currently, highest end training session require on the order of 10^{25} FLOPs (tens of YottaFLOPs)
- LLM training FLOPs requirements doubling every 5.6 months
- Roughly 11X faster than HPC Top One Linpack performance growth rate

- Fastest HPC in June 2024, Frontier at Oakridge National Lab, DOE peaks at 10^{18} floating point operations *per second* -> 10^{25} /10^{18} = 10^{7} or ~ten million seconds or ~115 days
  - This assumes 100% efficiency: reality is much less

LLM Training: Primarily Limited to a Handful of Specialized Organizations

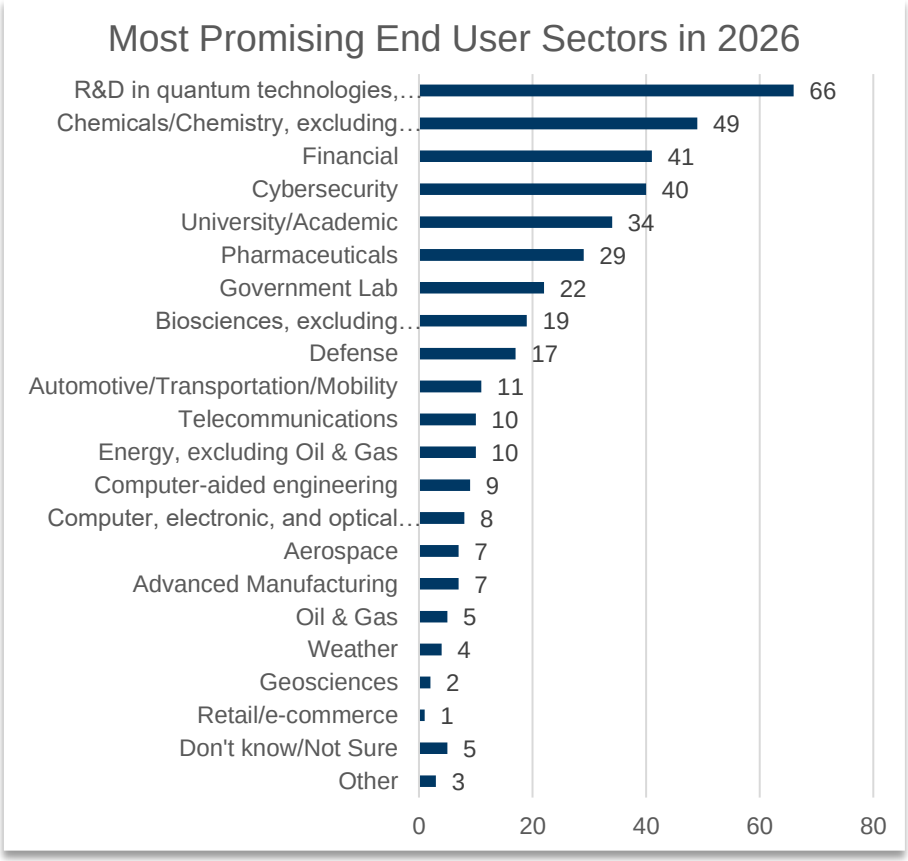
- Only a handful of the largest and best funded organizations can commit to train LLMs at this scale
  - Examples include LLM suppliers such as Google, Hugging Face, OpenAI, and Anthropic
    - The bulk of Microsoft's 2023 $10 billion investment in OpenAI was for cloud compute credits and not cash
  - Likewise, a few aggressive generative AI users train LLMs in-house with their own purpose-built systems such as Telsa and Meta
    - Telsa's Colossus, arguably the most power LLM training machine in the world, today incorporates 100,000 Nvidia H100 GPUs at an estimated cost of about $3 billion
- Mainstream gen AI users typically train at 3-4 orders of magnitude lower flop counts
  - Most conduct multiple in-house LLM training sessions, trading off model size for model precision and custom understanding
  - The bulk of these LLMs are open-source versions, typically downloaded from Hugging Face
  - LLM end users are looking to implement efficient, targeted small language models to
    - Reduce computational complexity
    - Ease the requirements for large and sometimes unverified data sets
    - Produce more focused sector, disciplined, or company specific LLMs

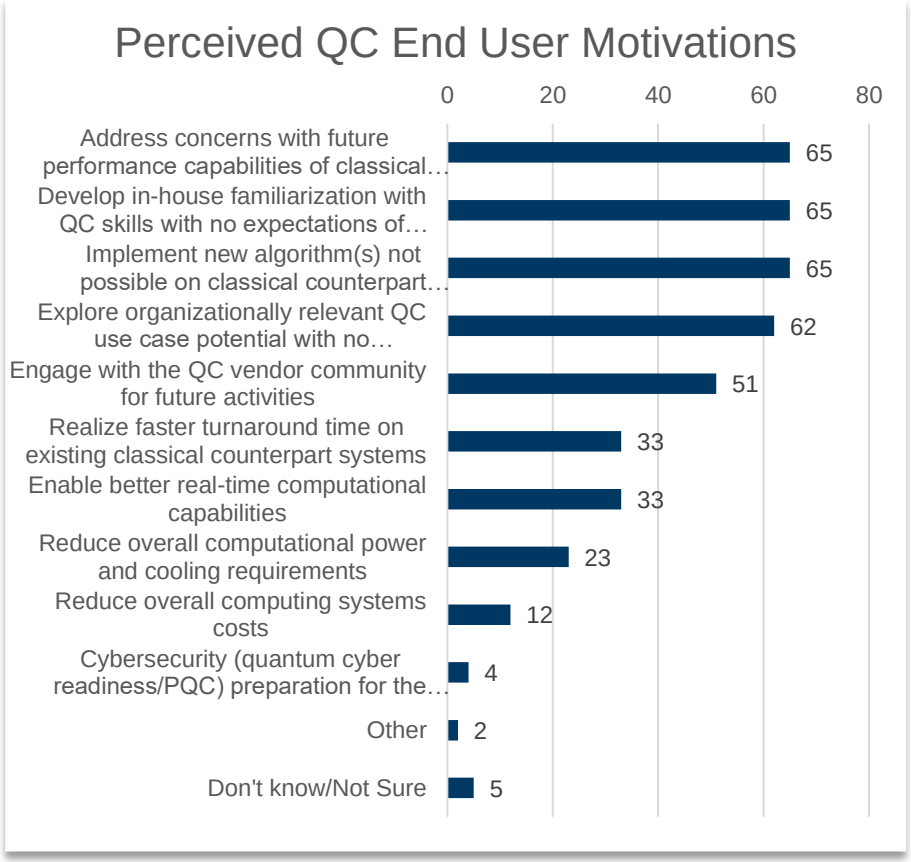
Global Quantum Computing Market Highlights

Continued strong and steady progress for the global QC sector

- The global quantum computing (QC) market is estimated to be approximately $1 billion in 2024
  - A projected annual growth rate of 22.1% brings the global QC market to $1.5 billion in 2026
- Supply-side growth driven by a collection of factors:
  - Continued revenue growth by traditional QC suppliers
  - First revenue appearances by new-to-market players
  - Expanding base of domestic suppliers in nascent markets
  - Increasing sophistication and specialization of QC stack
- Demand-side interest on the rise:
  - Widespread interest in accelerating critical compute jobs
  - More end use case exploration within the overall HPC community
  - Sustained government programs, and related government procurements, fostering sales and increasing credibility to potential end users

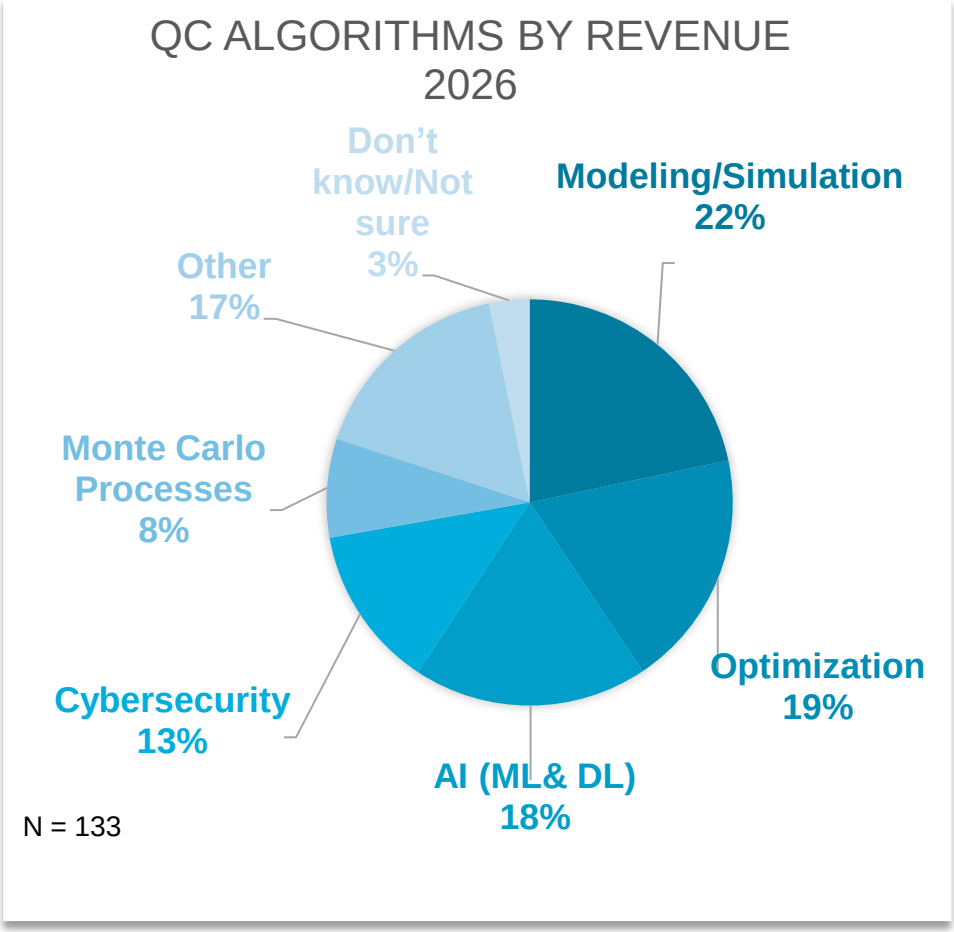
Currently, the Promise of QC is Substantial



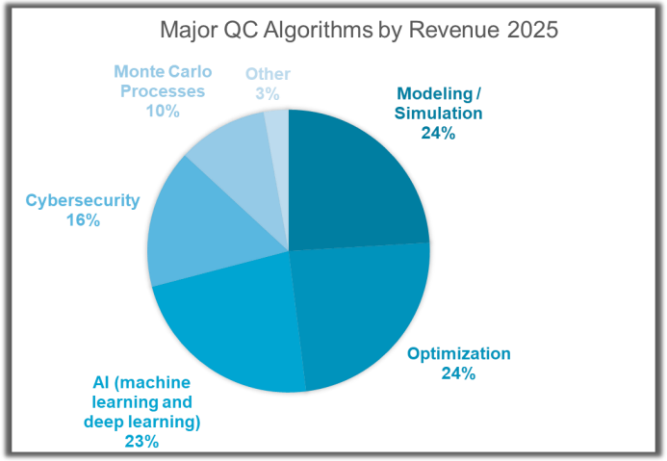


QC Market 2026: Major Algorithms by Revenue

Mod/sim, optimization, and AI still the three major algorithms areas



- Some changes from previous years' studies
  - The overall share of the big three is slipping due to major gains in the "Other" category
  - But there was little detail offered about the new alternatives
  - Suggests trends towards mixed algorithms use



Different Workloads/Different Solutions

HPC Use Case	Modeling/Simulation	Big Data/Data Science	AI: Large Language Models	Cloud-based HPC
Data Format	64-bit floating point numerical formats	64-bit floating point or integer data format	Low, mixed, or AI-specific precision formats	Variable formatting
General Code Characterization	Mix of both parallel and serial codes	Primarily parallel codes suited to cluster architectures	Distributed parallel codes, tightly coupled compute engines	Favors either small serial or large task parallel codes, loosely clustered systems
Processor and Accelerator Configuration	High core count CPU-based, GPUs support/augment CPU computation	High GPU counts, CPU-managed data flows	Emphasizes GPU or related AI-specific accelerators, strong GPU-GPU interactions	Flexible node configurations of CPU/GPU/accelerators, virtual and bare metal options
Data Storage	Consistent, uniform storage formats, large file size and consistent storage access patterns	Varying data formats: text, semi-structured data, structured data, binary, random data access patterns	Large numbers of small read-only files, multiple re-read iterations during training, high rate of data reuse	Multiple virtual compute instances with varying storage configurations, geographically distributed
Job Characterization	Small data input, compute intensive, large data output	Large data input, processes data at high velocity, small data output	Significant aggregate Flops count (10^{25}), large data input, matrix operation intensive	Wide job specifics addressable by both virtual and bare metal options
Exemplar Software	C++, Fortran, MPI, OpenMP	Python, Java, R, Scala, MATLAB	BERT, GPT, Megatron-Turing	Docker, Fargate, Kubernetes
Data Characterization	Program accuracy dependent on empirical validation/verification process	Verifiable data with strong statistical underpinnings	Dependent on existing training data availability/validity	Data physical location affects data size, access, performance, and price

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

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

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

What Will It Take to Get There?

Both sides need to cooperatively meet in the middle

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

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

Workloads drive architecture, and users drive workloads

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

HYPERION RESEARCH

QUESTIONS?

bsorensen@hyperionres.com

Float to the top or sink to the bottom. Everything in the middle is the Churn.
- Amos Burton, *The Expanse*

Today's Agenda

- **Earl Joseph, CEO**
  - HPC and AI Market Update
  - A New Way of Measuring Value of Leadership Computing
  - Tool for Deeper Understanding of Surveys Results
- **Bob Sorensen, SVP, Chief AI & QC Analyst**
  - Successfully Navigating the Changing Advanced Computing Landscape
- **Mark Nossokoff, Research Director, Chief Cloud & Storage Analyst**
  - Perspectives on HPC-AI Storage and Interconnects
  - HPC-AI Cloud Update
- **Innovation Award Winners Announcement**
- **Conclusions**

SC24 HPC-AI Market Update - Storage and Interconnects

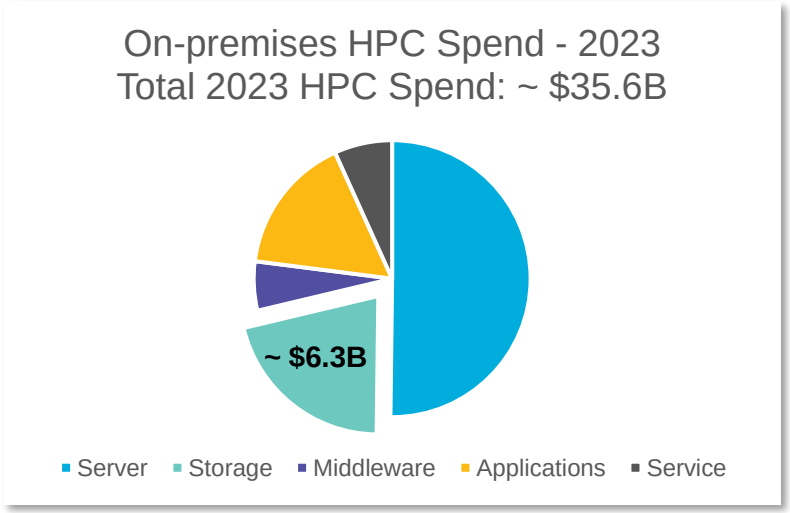
SC24 Breakfast Briefing
November 2024

Mark Nossokoff

www.HyperionResearch.com
www.hpcuserforum.com

HPC Storage Growth Continues

Demand increasing across all sectors and verticals



- **Storage continues as the highest growth HPC element**
- **Storage represents ~ 21% of on-premises HPC spending in 2023 and growing to ~22.4% in 2028**

Source: Hyperion Research, 2024

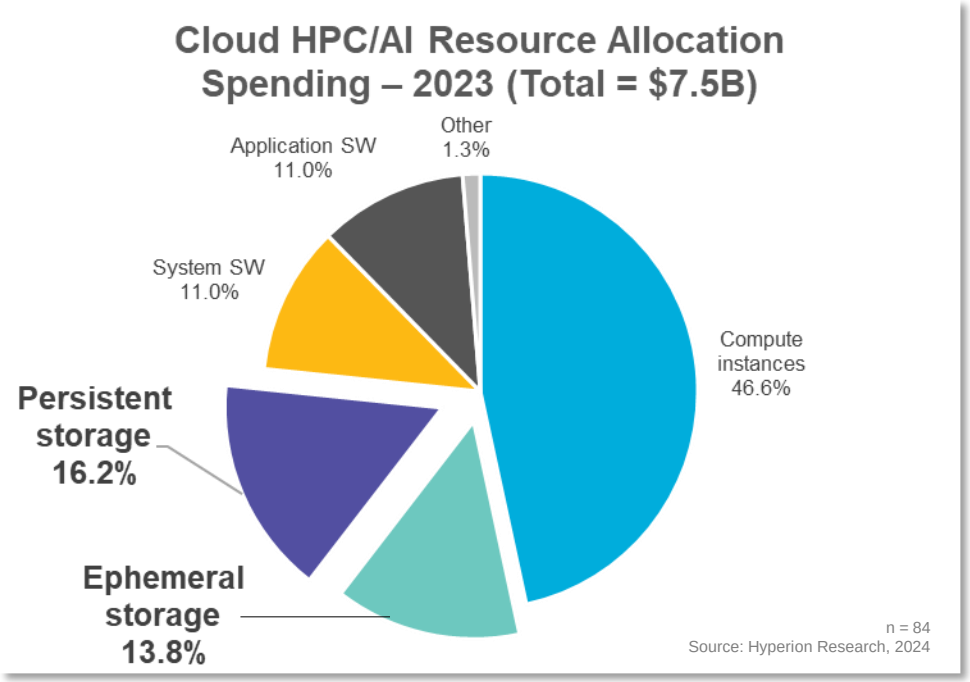
Area ($M)	2023	2024	2025	2026	2027	2028	CAGR 23-28
Server	$20,735	$25,390	$29,559	$33,699	$37,797	$41,777	15.0%
Add-on Storage	**$6,282**	**$7,692**	**$8,745**	**$9,771**	**$10,738**	**$11,846**	**13.5%**
Middleware	$1,711	$2,026	$2,241	$2,468	$2,691	$2,968	11.6%
Applications	$4,830	$5,684	$6,267	$6,878	$7,468	$8,240	11.3%
Service	$2,014	$2,262	$2,411	$2,498	$2,696	$2,973	8.1%
Total Revenue	$35,573	$43,054	$49,223	$55,315	$61,390	$67,805	13.8%

Source: Hyperion Research, 2024

HPC-AI Cloud Resource Allocation Spending - 2023

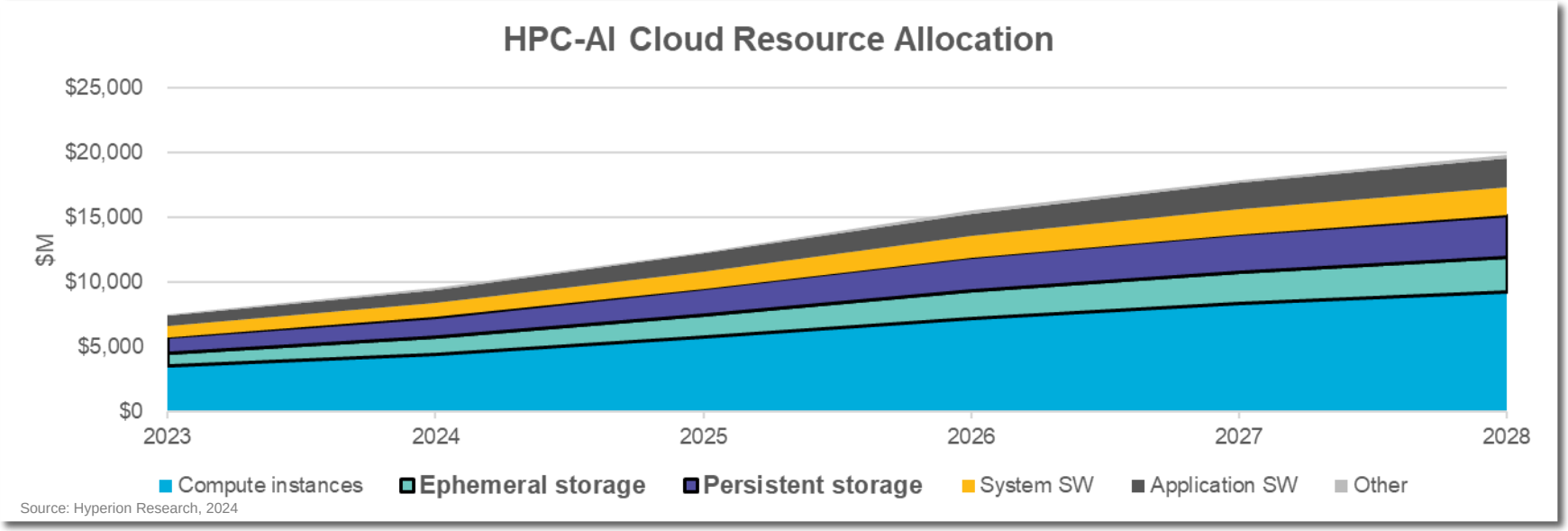
Cloud adoption for storage remains strong and growing

- **Storage ~ 1/3 total HPC spending in the cloud, consistent with prior study**
- **Spending on persistent, durable storage now only slightly greater than on ephemeral, temporal storage**
  - Persistent storage was 2x ephemeral storage in prior study



HPC-AI Cloud Resource Allocation Spending - Forecast

Cloud storage projected to approach $6B in 2028



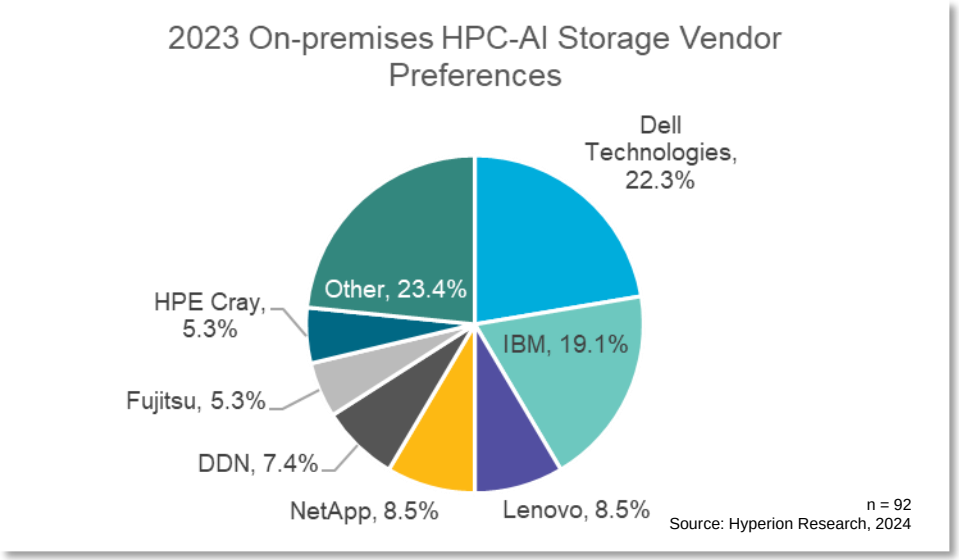
Source: Hyperion Research, 2024

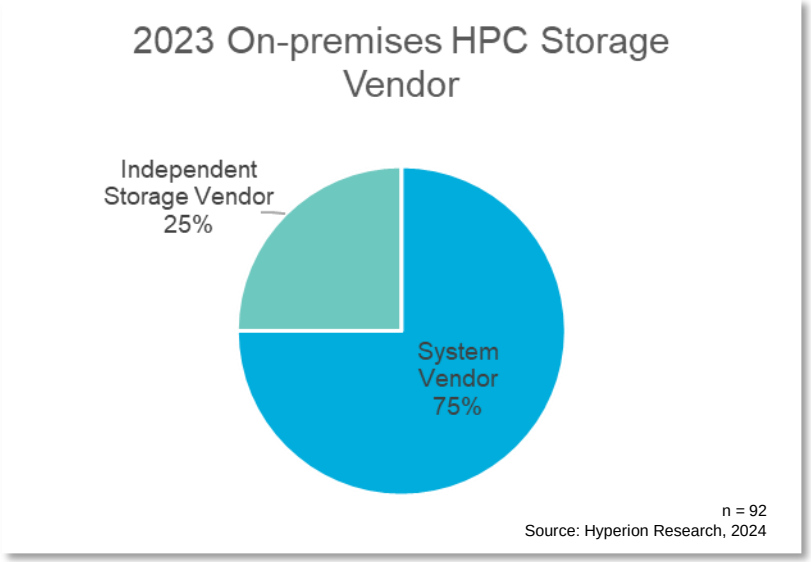
($M)	2023	2024	2025	2026	2027	2028
Compute instances	$3,503	$4,446	$5,768	$7,232	$8,338	$9,229
Ephemeral storage	**$1,041**	**$1,321**	**$1,714**	**$2,149**	**$2,477**	**$2,742**
Persistent storage	**$1,216**	**$1,544**	**$2,002**	**$2,511**	**$2,895**	**$3,204**
System SW	$828	$1,051	$1,363	$1,709	$1,970	$2,181
Application SW	$829	$1,052	$1,365	$1,711	$1,973	$2,184
Other	$98	$125	$162	$203	$234	$259
Total	$7,516	$9,540	$12,376	$15,519	$17,892	$19,804

Source: Hyperion Research, 2024

HPC Storage Vendor Preferences

Dell Technologies continues as top preferred storage vendor





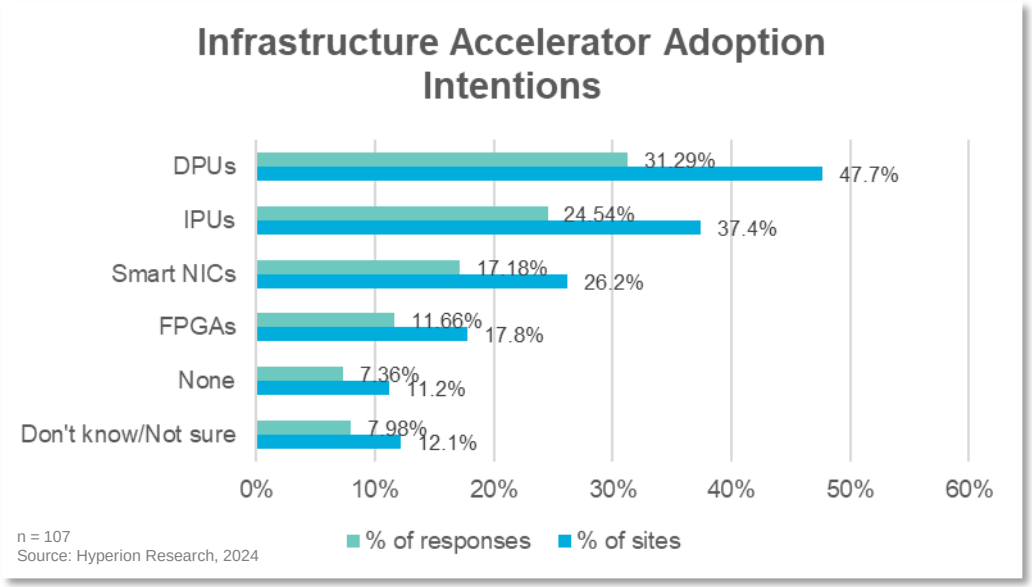
Q: Who is your primary storage system provider?

- **Dell remains the preferred overall and systems storage provider, followed by IBM and Lenovo**
- **HPE Cray appears low due to heavy industry makeup of survey and minimal participation of leadership sites**
- **NetApp edged past DDN as the most preferred independent storage provider as the 4th and 5th overall preferred providers.**
- **Others (@ >2% preference): VDURA, ATOS, Huawei, Inspur**

Infrastructure Accelerator Adoption Intentions

~82% of sites with an opinion intend to employ an infrastructure accelerator

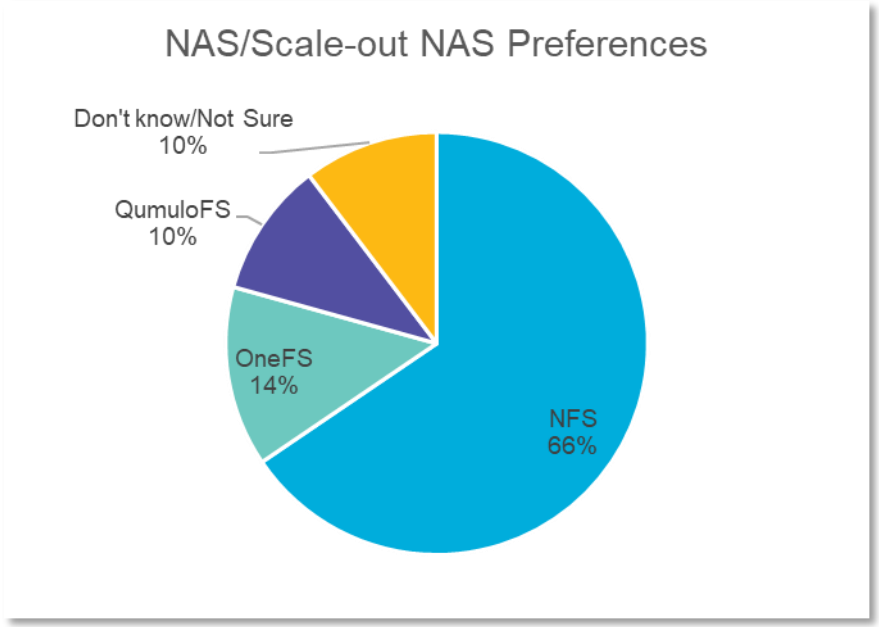
- **DPUs cited as the most popular infrastructure accelerator**



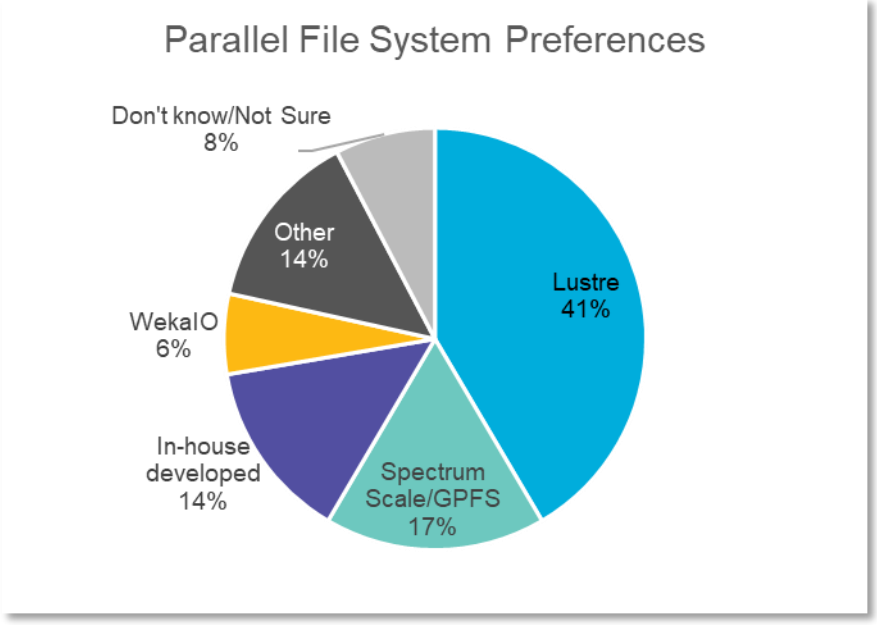
Q: Which types of infrastructure-oriented accelerators, if any, are you planning on using in any of your HPC systems over the next 1-2 years? Select all that apply.

File System Preferences

Lustre and NFS continue as preferred parallel file system and NAS/scale-out preferred file systems, respectively



n = 103 otal responses
% represents adoption at sites that indicated they deploy NAS/Scaleout on their largest system
Source: Hyperion Research, 2022



n = 103
% represents adoption at sites that indicated they deploy parallel file systems on their largest system
Source: Hyperion Research, 2022

Q: What primary file system type do you use on your largest HPC system?
Q: Who is your primary Parallel File System provider?
Q: Who is your primary NAS / Scale-out NAS file system provider?

- **Preferences reflect half of respondents coming from Industry**
- **61% of the 94 respondents do not use the same file system in the cloud as on-premises**

Interconnect Architecture

Subtle shift expected in interconnect implementation

Network Architecture	Prior survey		2024 Survey	
	Then-current	Next	Current	Next
Single, converged network for both server-server and storage-server communication	45.9%	53.0%	41.7%	44.7%
Separate networks for server-server and storage-server communication	54.1%	47.0%	58.3%	55.3%

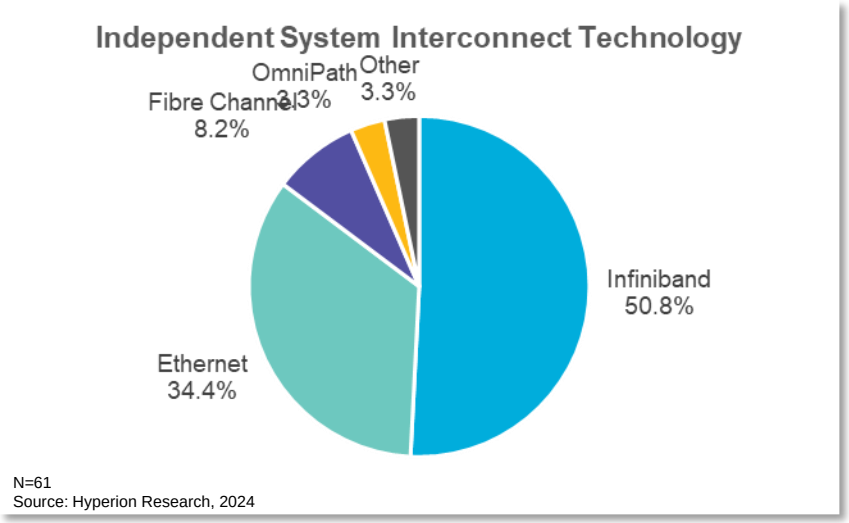
n = 103
Source: Hyperion Research, 2024

- **Prior survey (2022) indicated an intent for majority of next procurements to flip from independent to converged networks**
- **Current survey indicated higher existing adoption of independent networks with a dampened shift towards converged networks**

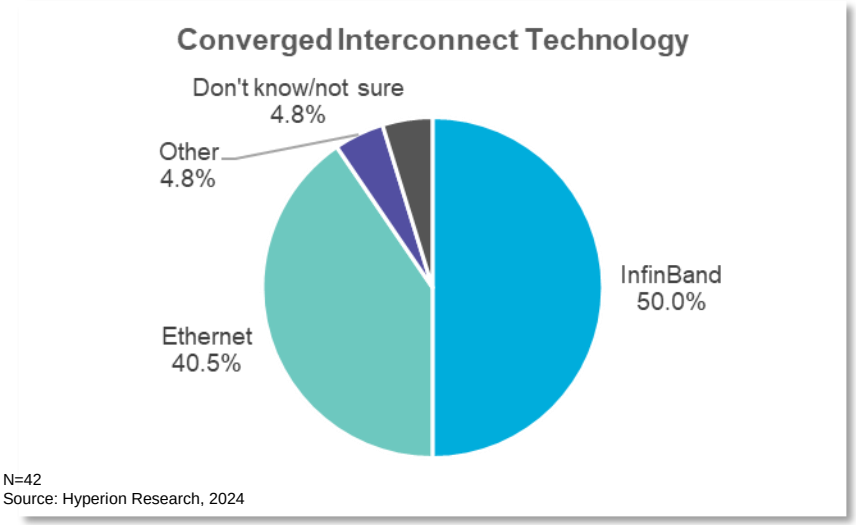
Interconnect Architecture

InfiniBand preferred overall

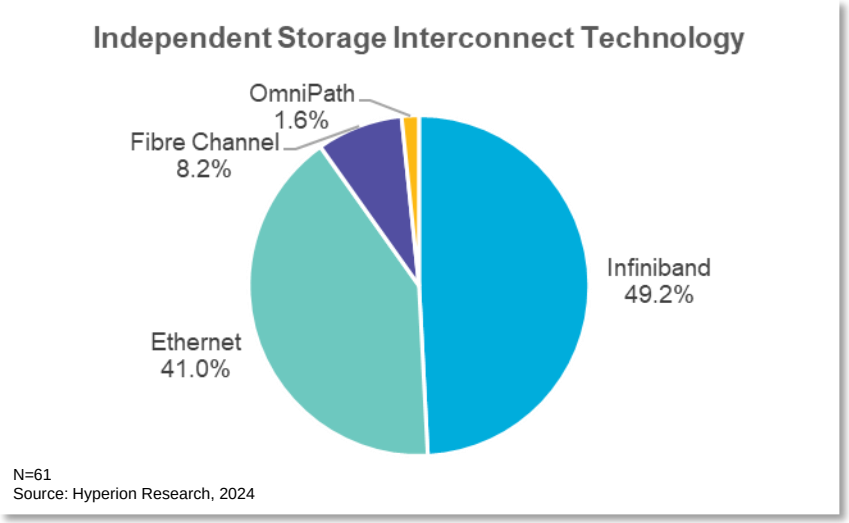
- **InfiniBand preferred in half of systems, regardless of architecture**
- **Ethernet 2nd preferred at between 35%-40%**
- **Other**
  - Converged: Slingshot
  - System: Cray Aries, HPE Slingshot



N=61
Source: Hyperion Research, 2024



N=42
Source: Hyperion Research, 2024

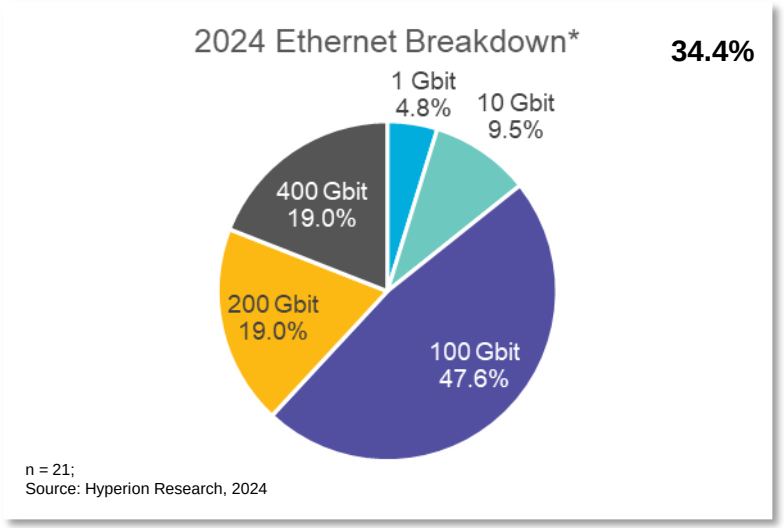


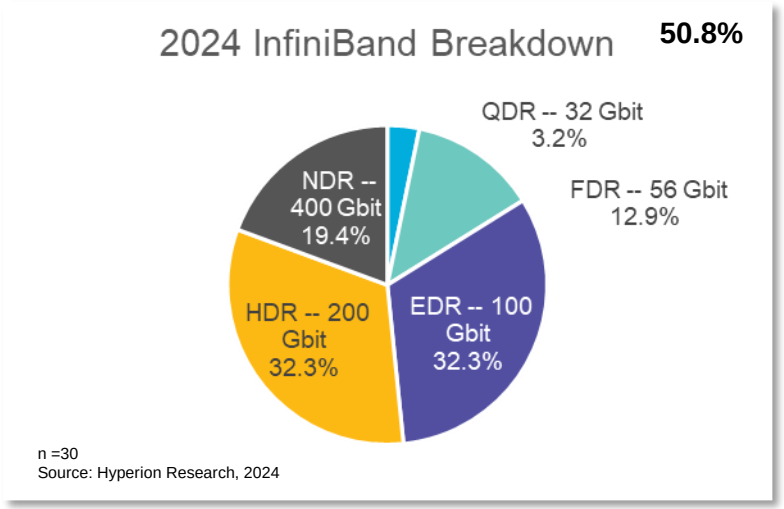
N=61
Source: Hyperion Research, 2024

Independent Networks: System-System

InfiniBand preferred at ~ 50% more sites than Ethernet

- **Ethernet**
  - 34.4% of sites surveyed with independent networks; down from 44.9% from prior survey
  - 10 Gbit halved, 100 Gbit flat, 200 Gbit doubled, 400 Gbit up from 15%
- **InfiniBand**
  - 50.8% of sites surveyed with independent networks; up from 35.9% from prior survey
  - QDR down, FDR up, EDR & NDR flat, NDR down somewhat
- **Other**
  - Omni-Path = 3.5%
  - Other = 3.5%
    - Cray Aries
    - HPE Slingshot



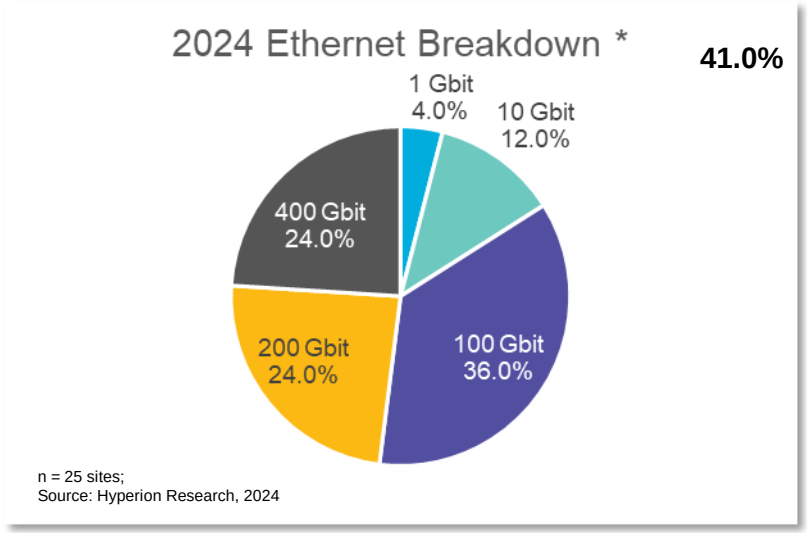


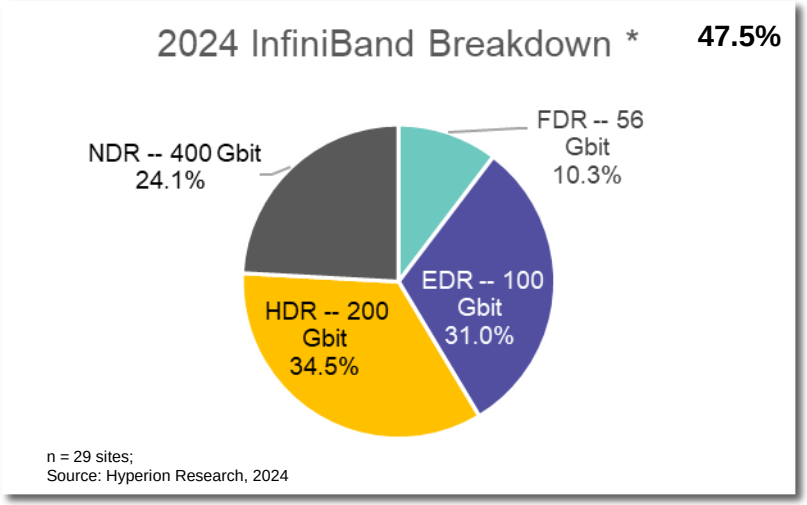
* % of site adoption Ethernet and Infiniband, respectively

Independent Networks: System-Storage

InfiniBand preferred over Ethernet for system-storage

- **Ethernet**
  - 47.8% of sites surveyed with independent networks
  - 100Gb most widely deployed, but down from 49%, with 200 Gbit growing from 10%
  - 400 Gbit up from 17%
- **InfiniBand**
  - 47.5% of sites surveyed with independent networks
  - HDR 200 Gbit most widely deployed at InfiniBand sites, overtaking EDR
- **Other**
  - Fibre Channel = 9.8%
  - Omni-Path = 1.6%



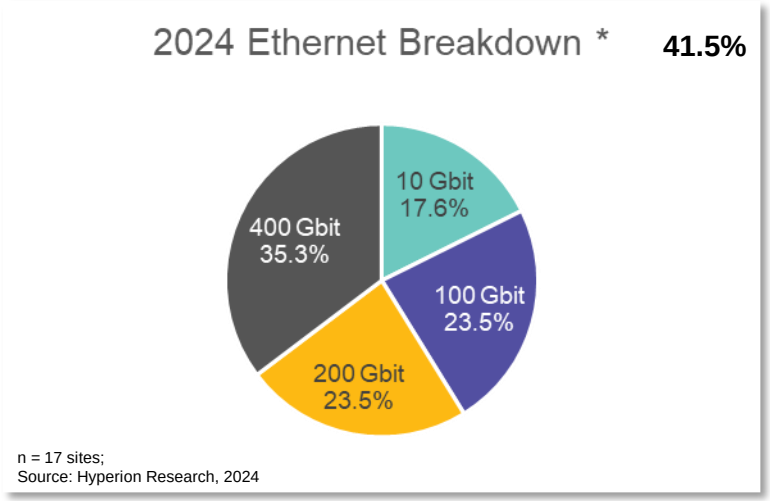


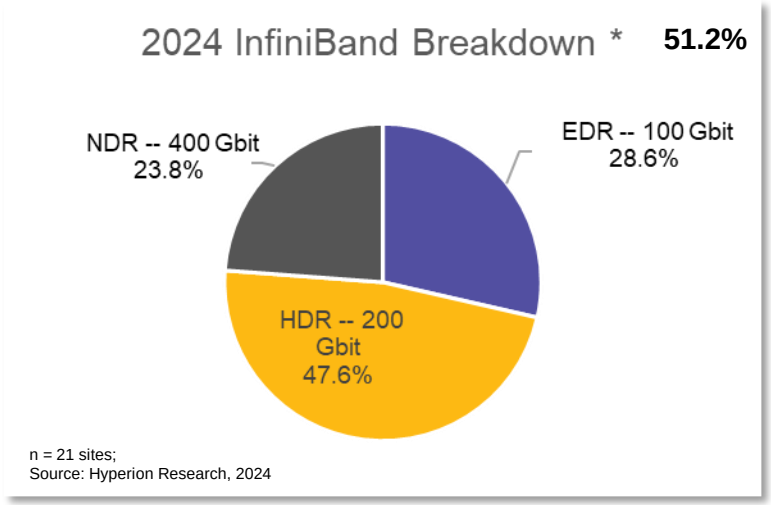
* % of site adoption Ethernet and Infiniband, respectively

Converged Networks

InfiniBand grew to half of respondents for converged networks

- **Ethernet**
  - 41.5% of sites surveyed with converged networks
  - 400Gb most widely deployed within Ethernet sites (was 100 Gbit)
- **InfiniBand**
  - 51.3% of sites surveyed with converged networks
  - HDR 200 Gbit most widely deployed at InfiniBand sites (was EDR)
  - NDR up from 9%
- **Other**
  - Slingshot = 2.4%

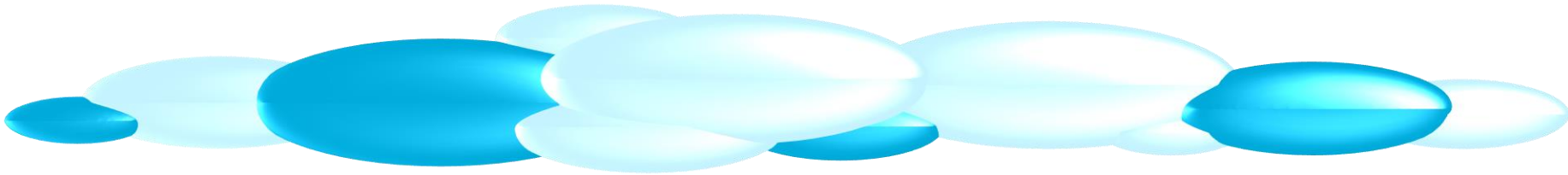




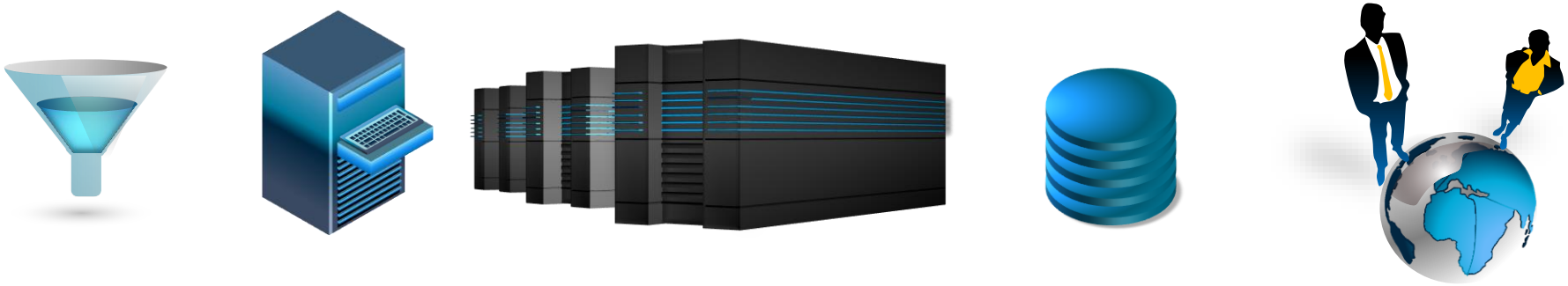
* % of site adoption Ethernet and Infiniband, respectively

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?)**

Future Research Direction

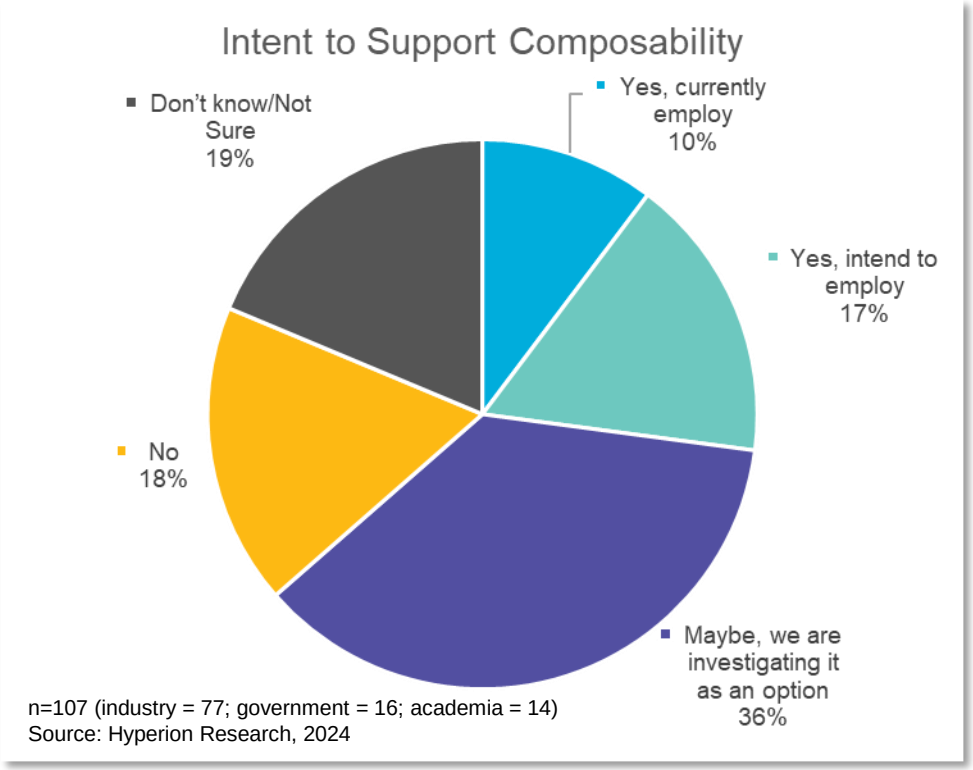
No shortage of industry and ecosystem activity

- **Impact and evolution of the AI data pipeline on storage architectures**
- **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)**

Intent to Support Composability

Broad range of intended adoption

- **27% currently employ or intend to employ**
- **36% investigating**
- **19% don't know**
- **18% have zero intent**

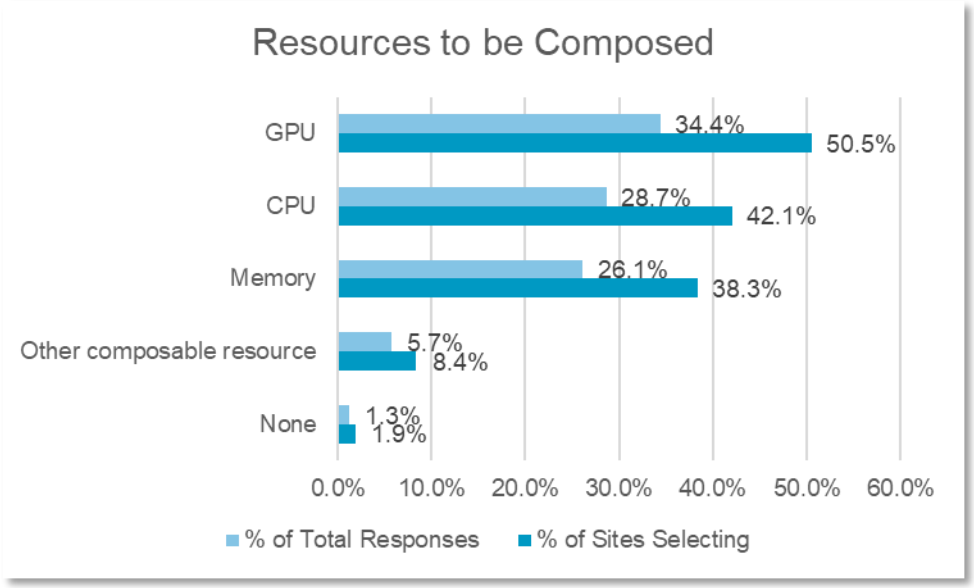


n=107 (industry = 77; government = 16; academia = 14)
Source: Hyperion Research, 2024

Elements Being Composed

GPUs selected by the most sites

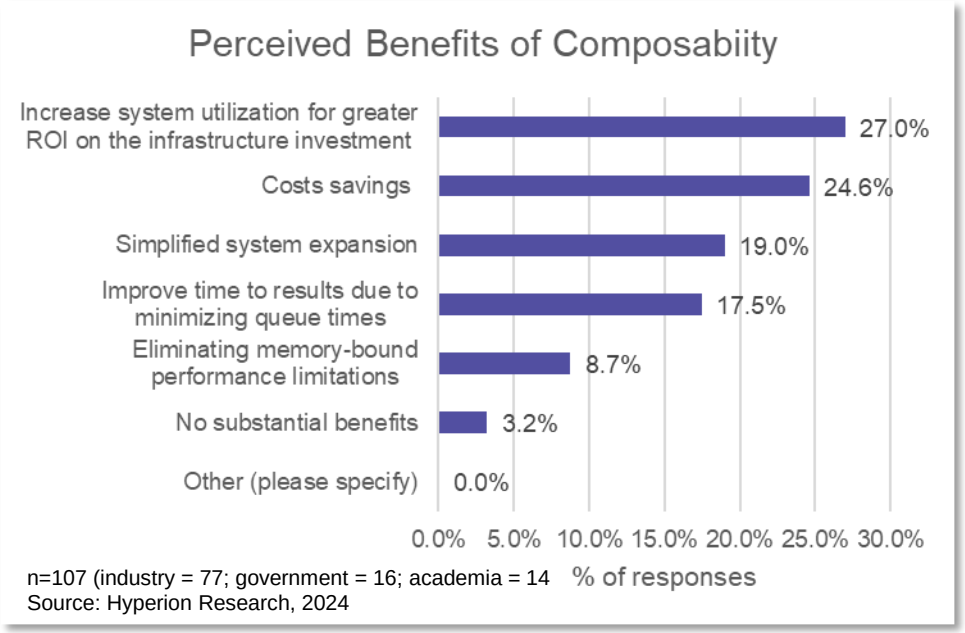
- **Relative to # of Sites**
  - GPUs: 50.5%
  - CPUs: 42.1%
  - Memory: 38.3%
- **Relative to Total Responses**
  - GPUs: 34.4%
  - CPUs: 28.7%
  - Memory: 26.1%



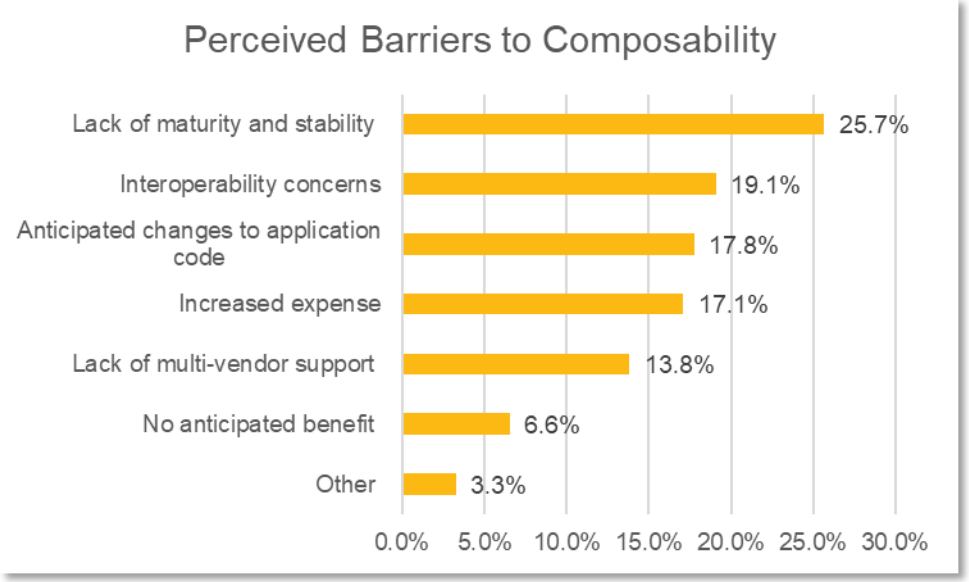
n=107 (industry = 77; government = 16; academia = 14
Source: Hyperion Research, 2024

Benefits / Barriers to Composability

- **Benefits**
  - Increased system utilization
  - Cost savings
  - Simplified scalability
- **Barriers**
  - Lack of maturity, stability
  - Interoperability concerns
  - Anticipated changes to application codes



n=107 (industry = 77; government = 16; academia = 14
Source: Hyperion Research, 2024

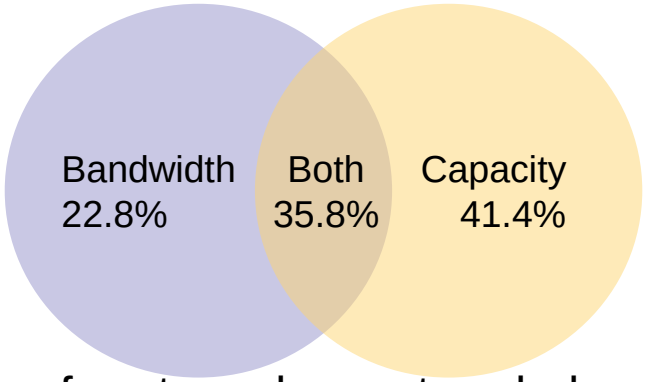


n=107 (industry = 77; government = 16; academia = 14
Source: Hyperion Research, 2024

Memory Considerations

Memory Considerations on Composability from Select Research

- **General:**
  - Almost 3/4s of surveyed sites have applications with memory-bound performance limitations
- **Applications with memory-bound performance limitations:**
  - Memory capacity is a limiting factor for more than 3/4s of applications at sites with memory-bound performance limitations



  - Large majority of systems have stranded memory capacity between 10%-50%
  - Traditional HPC modeling/simulation workloads more commonly had stranded memory capacity than AI or HPDA/data analytics workloads
  - More than half of surveyed sites require between 64GB/s and 256GB/s additional bandwidth to alleviate memory bandwidth limitations

CXL Sentiment

- **Over 3/4s of surveyed sites had at least a basic understanding of CXL**
- **At the time of the survey, participating sites with an opinion expressed a broad range of perception about market and ecosystem readiness for productions CXL/composable systems**
  - ~ 43% indicated 12-18 months
  - ~ 46% indicated 18-36 months or more

HYPERION RESEARCH

Questions? We look forward to hearing from you!

mnossokoff@hyperionres.com

SC24 HPC-AI Market Update - Cloud

SC24 Breakfast Briefing
November 2024

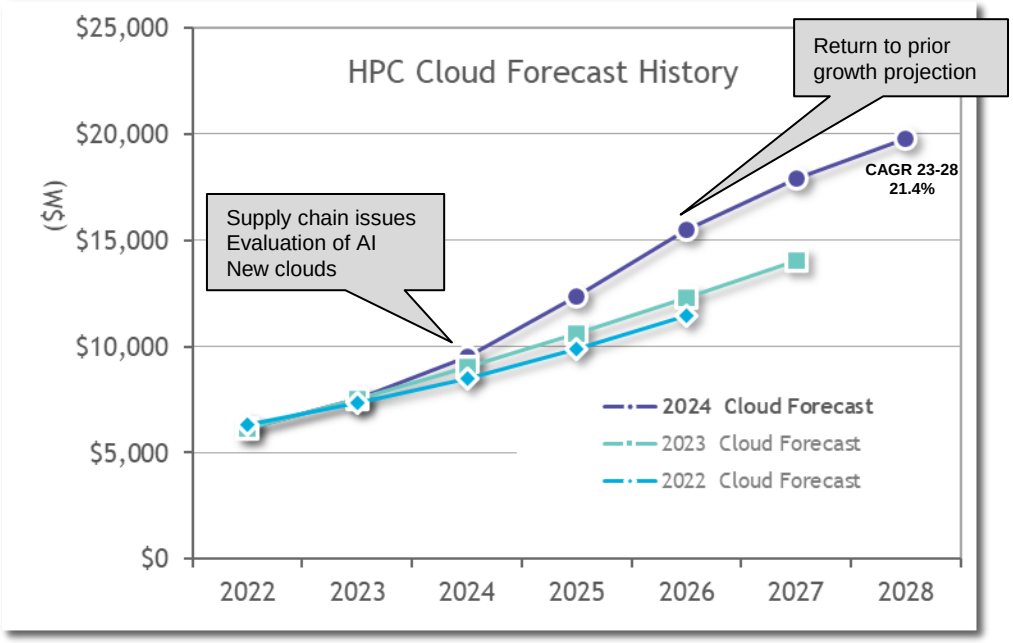
www.HyperionResearch.com
www.hpcuserforum.com

Mark Nossokoff

HPC/AI Cloud Forecast

Cloud revenue expected to approach $20B by 2028

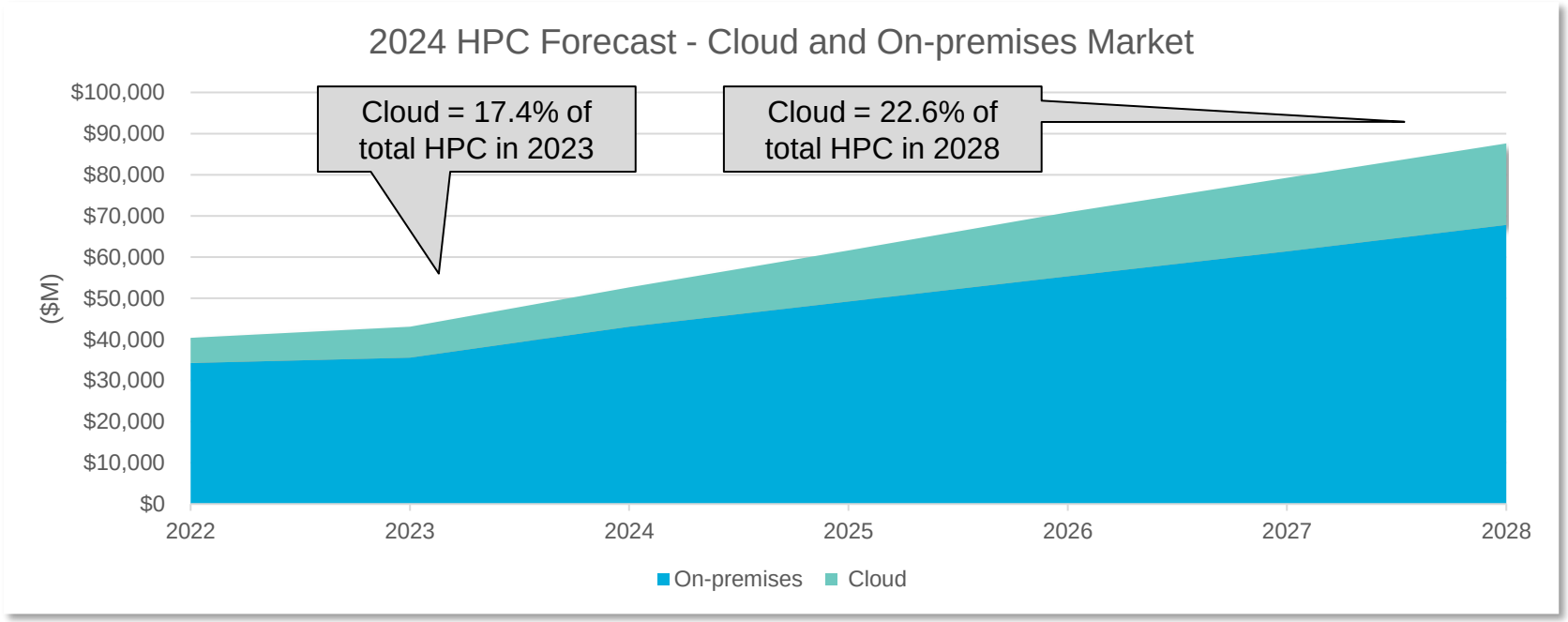
- **Global HPC/AI buyers around the world continue shifting portions of their on-premises budgets to spending in the cloud**
- **Increasing growth rate for 2024, returning to prior projected growth in 2026**
- **Primary growth drivers:**
  - Supply chain issues
  - Access to latest GPU technology
  - Experimentation and evaluation of AI workflow and infrastructure needs
  - Emergence of sovereign clouds



- **Forecast has steadily increased over previous forecasts**

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

Total HPC-AI exceeds $87B in 2028



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

Cloud Providers' Responses to Demand

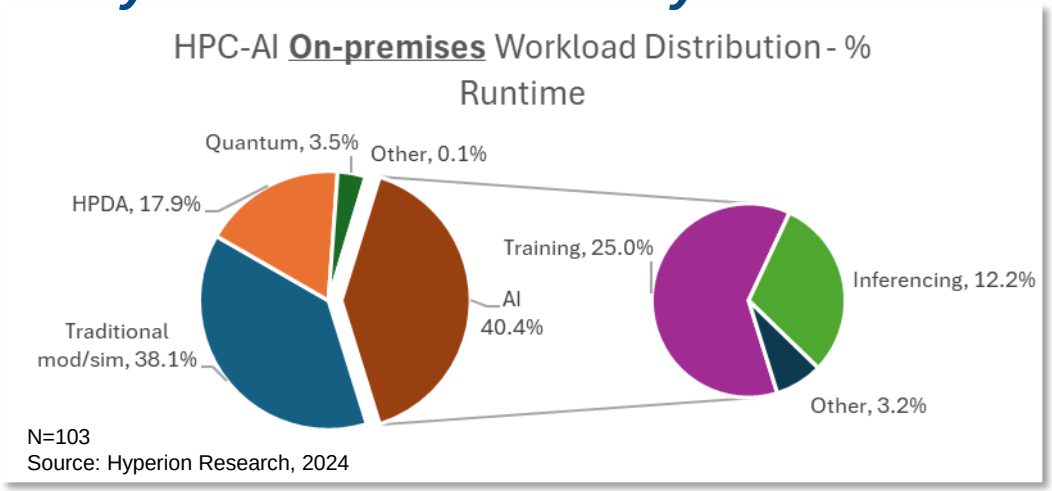
Rising cloud utilization driven almost entirely by AI adoption

- **Extensive investments by cloud providers**
  - Application-specific processors and accelerators
  - New data center buildouts
  - Expanded HPC focus from CSPs
    - AWS released Parallel Compute Service
    - Google GA's Parallelstore
- **AI and GPU clouds**
- **Sovereign clouds**
  - Driven by local legal requirements, cost constraints, and tax & privacy policies, a shift to smaller, more adaptable data centers could result
  - CSP's ability to dynamically shift load demand between geographic regions and national borders may be impacted
    - Local provisioning to become more critical
  - What is the correct metric to key in on to retain "sovereignty"?
    - Where data is created and where it can be moved, or…
    - …security and governance of users who can access it, wherever it may be stored?

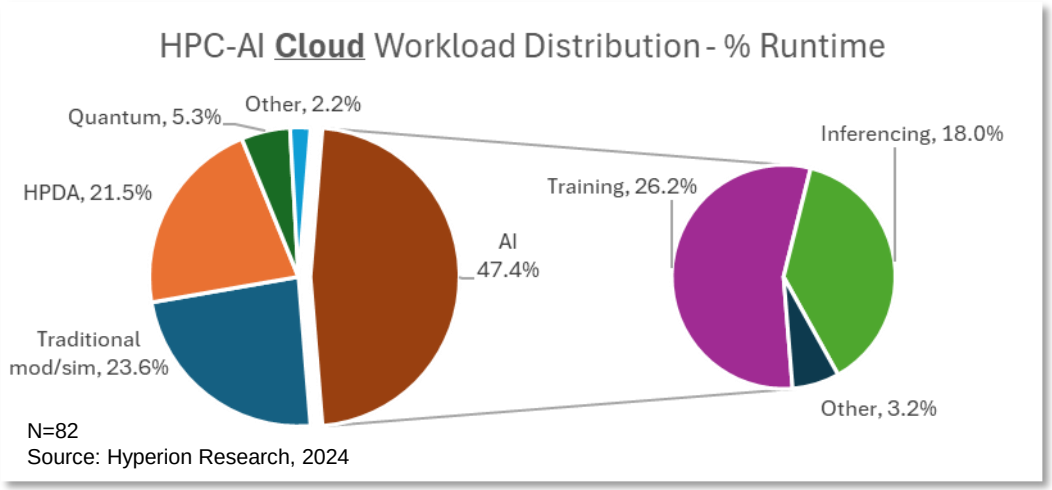
HPC-AI Workload Distribution by Environment - % Runtime

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

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



N=103
Source: Hyperion Research, 2024

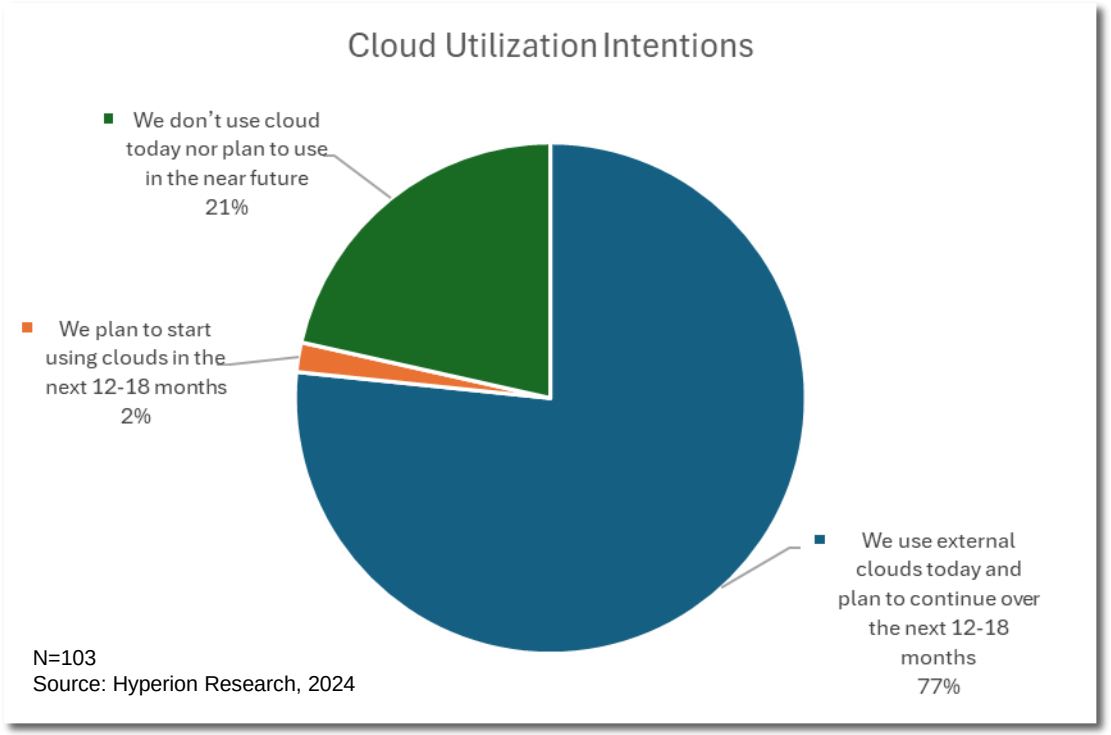


N=82
Source: Hyperion Research, 2024

Cloud Utilization for HPC-AI Workloads - Intentions

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

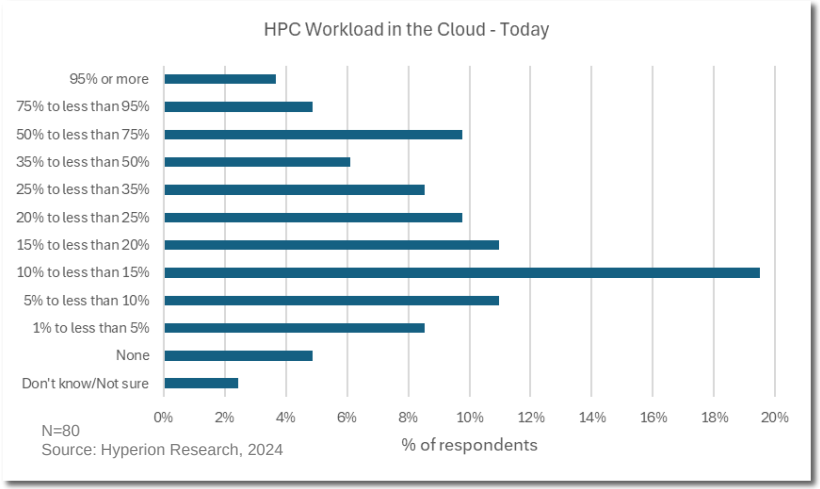
- **70% of respondents currently use or intend to use cloud within the next 12-18 months**
- **21% don't use the cloud or intend to use the cloud within the next 12-18 months**

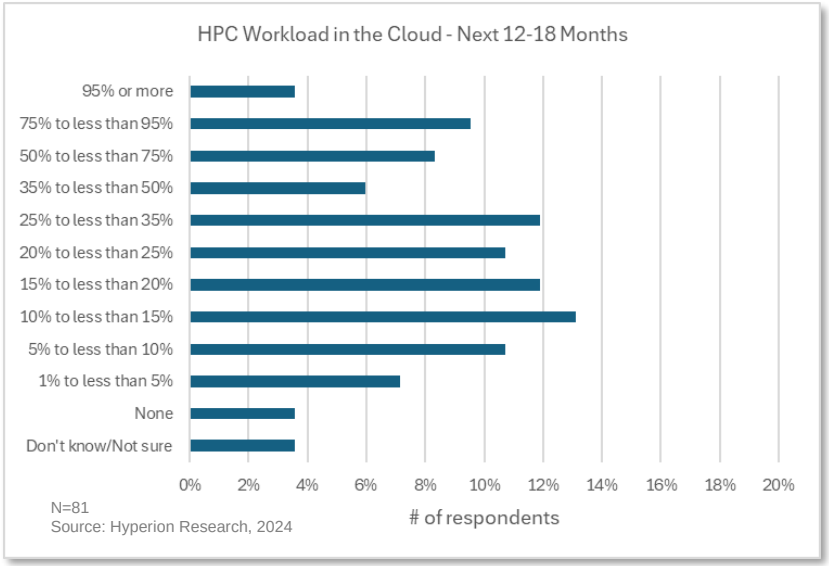


Cloud Utilization for HPC-AI Workloads - % Runtime

Based on overall runtime, approximately what percentage of all your HPC-AI workloads are run on external clouds TODAY/12-18 months?

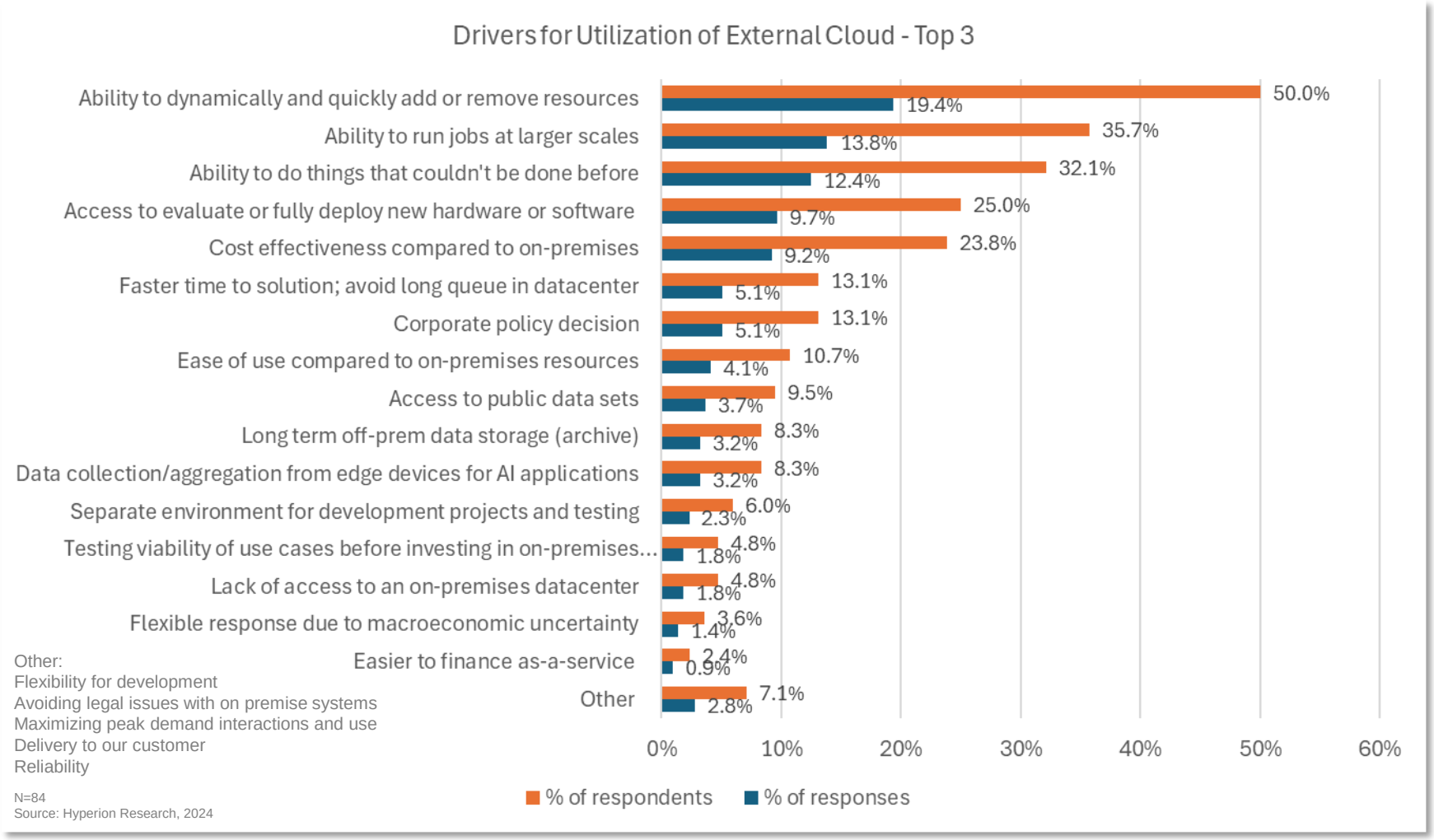
- **Respondents indicate an increase in application runtimes performed in the cloud (weighted average)**
  - Today: 27.2%
  - 12-18 months: 31.1%
- **Respondents running 50% or more of their application runtimes in the cloud growing**
  - Today: 18.8%
  - 12-18 months: 22.2%





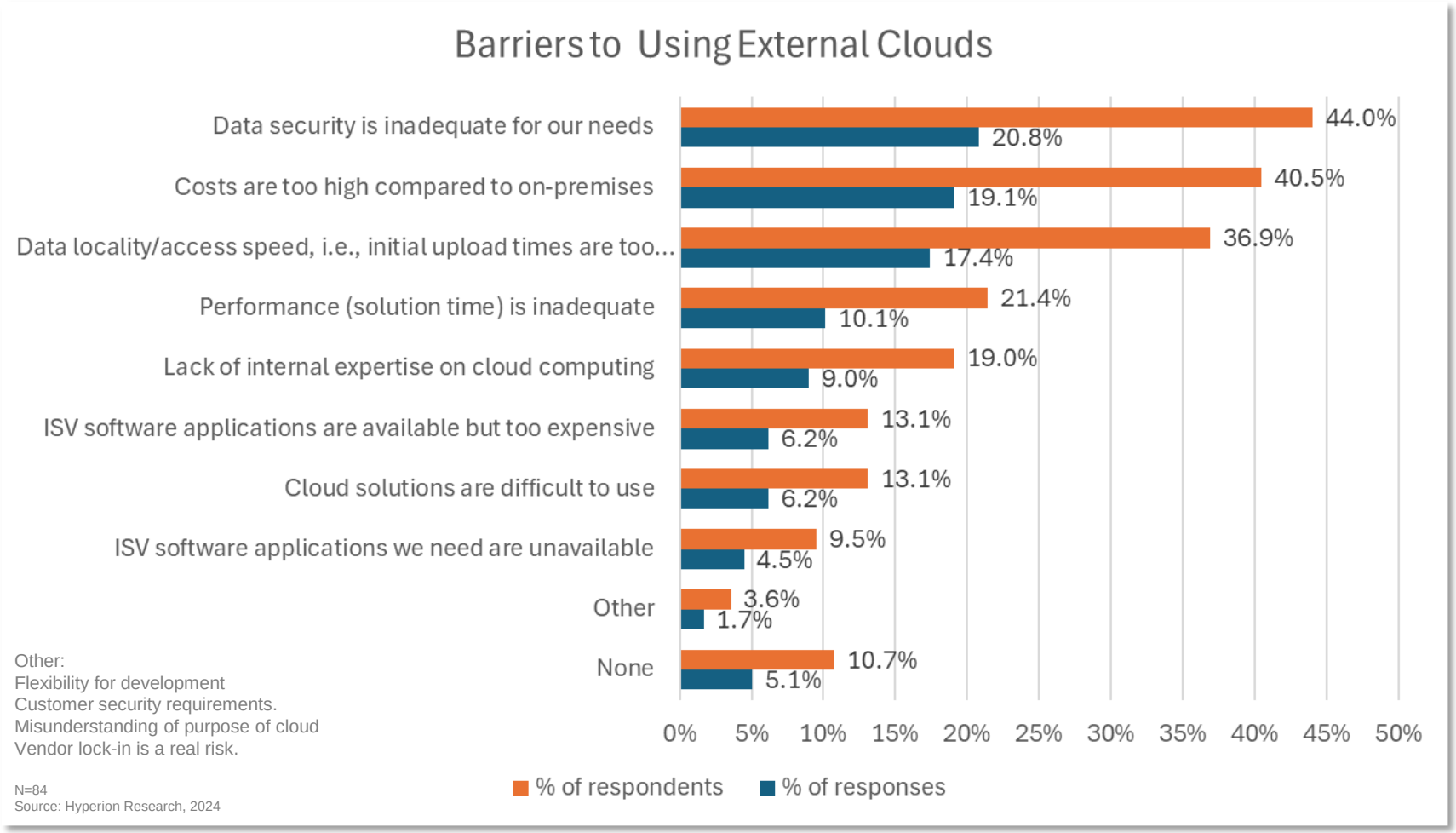
Drivers for External Cloud Adoption

Which of the following is a reason you use/plan to use external clouds for HPC? Please select up to 3.



Barriers for External Cloud Adoption

Which of the following do you consider to be barriers to increasing EXTERNAL cloud use for HPC workloads? Please select up to 3.



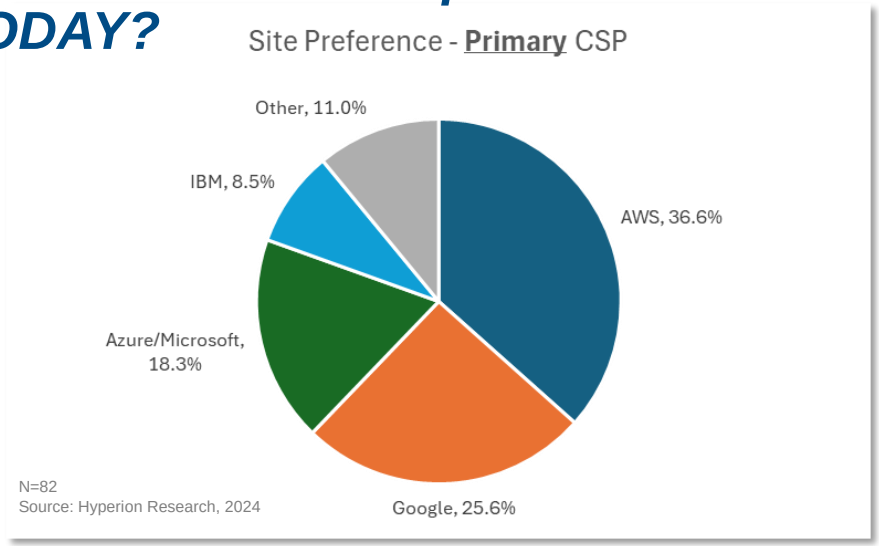
Other:
Flexibility for development
Customer security requirements.
Misunderstanding of purpose of cloud
Vendor lock-in is a real risk.

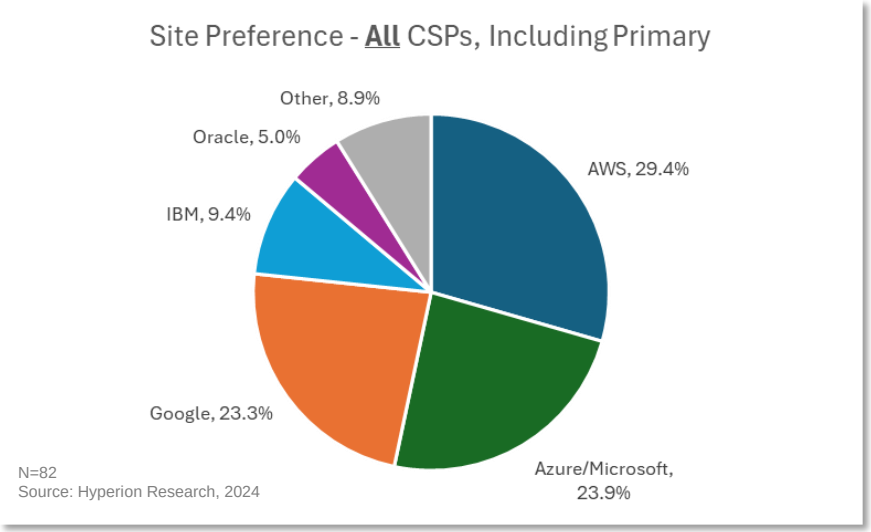
N=84
Source: Hyperion Research, 2024

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

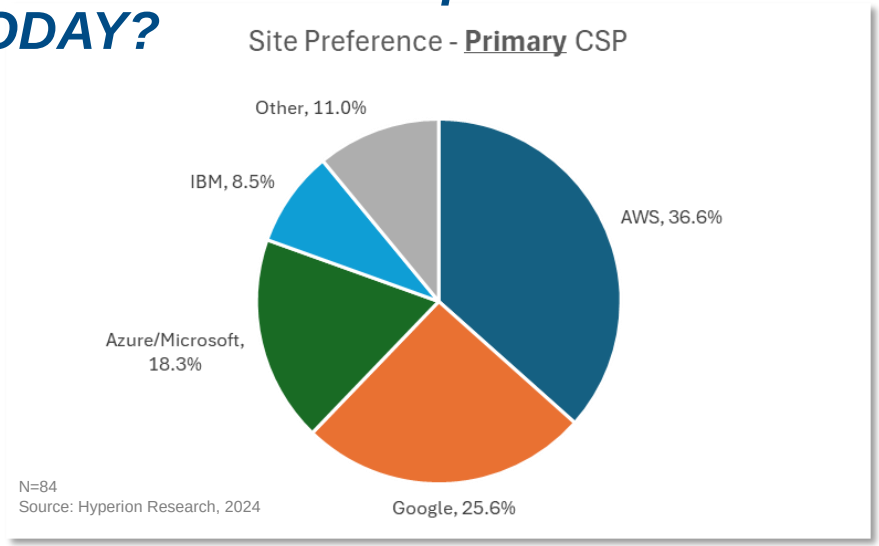


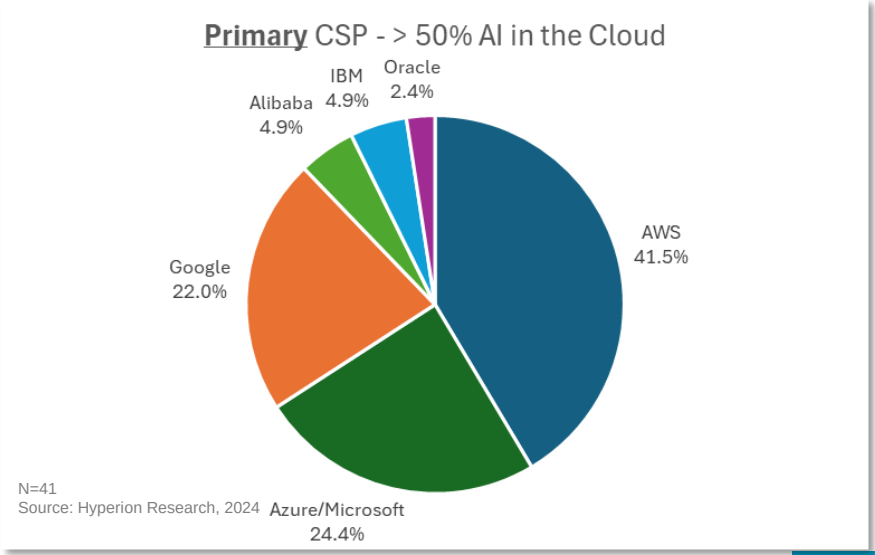


CSP Preferences – AI Workload Crosscut

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

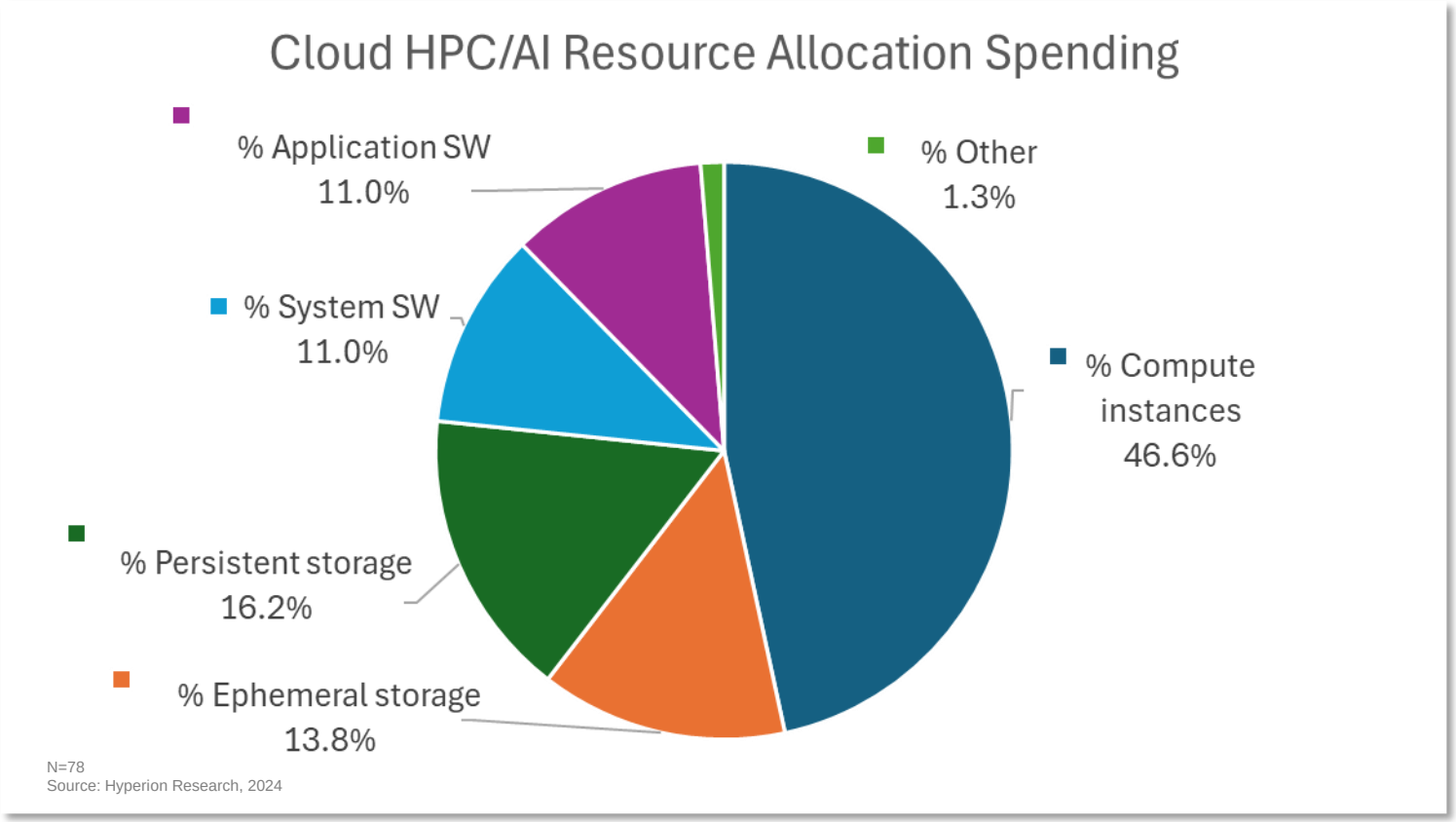
- **AWS the preferred primary CSP among respondents**
- **AWS as the primary CSP preference increases for sites who run >50% of their AI workloads in the cloud**
- **Microsoft moves to 2nd preferred primary preference for sites who run >50% of their AI workloads in the cloud**





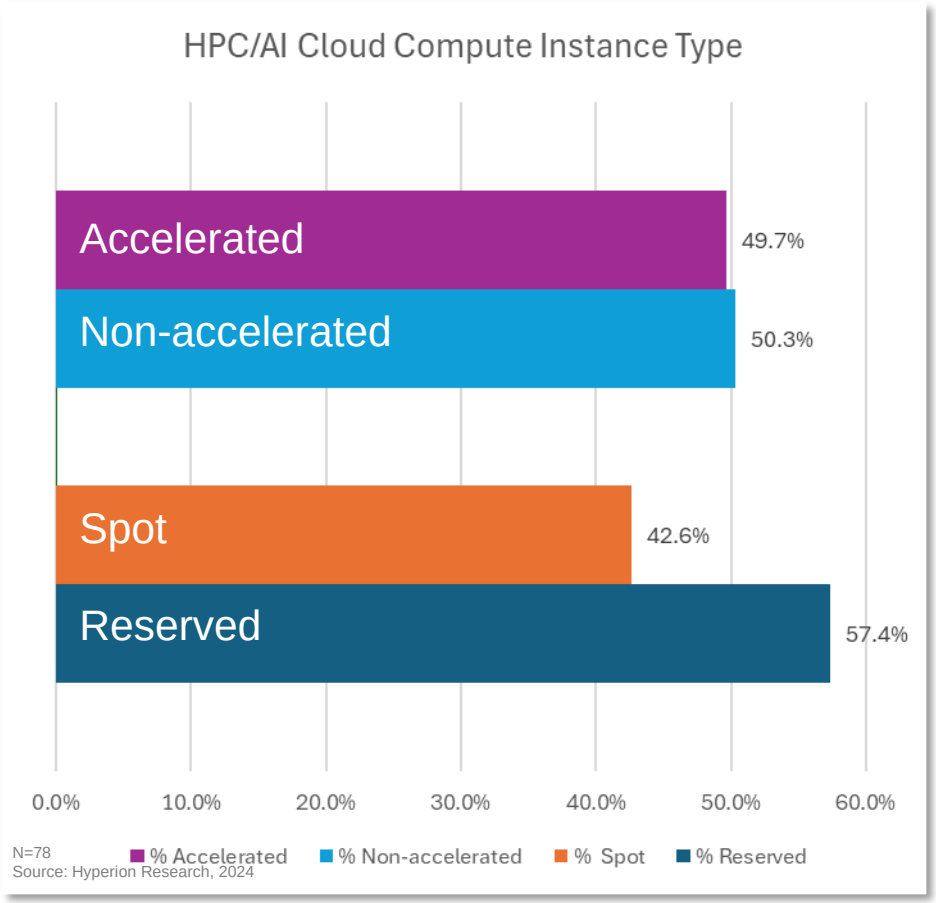
HPC-AI Cloud Resource Allocation Spending

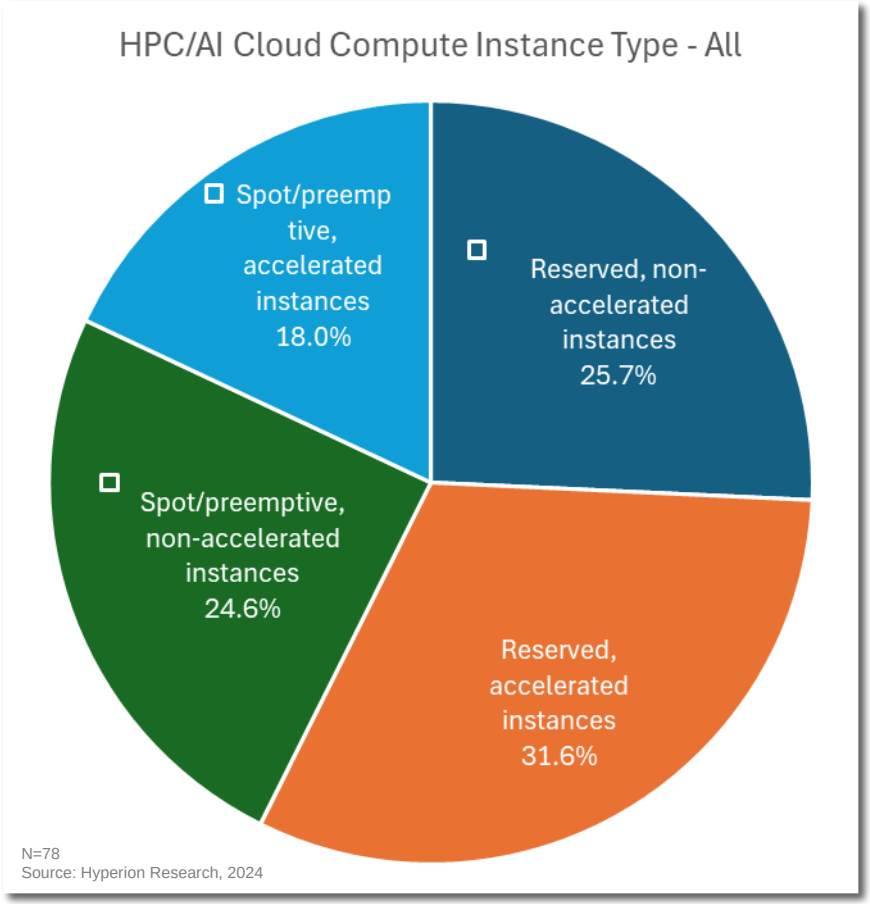
Please distribute your total HPC/AI/HPDA cloud resource spending between the following categories



HPC-AI Cloud Compute Instance Type

Please estimate your cloud instances for HPC/AI/HPDA workloads among the following categories:





What's Next in Clouds?

Dynamic environment anticipated for continuum computing and cloud adoption for the foreseeable future

- **Continued growth of user spending for HPC/AI advanced computing resources in the cloud from both new users and migration of workloads from current users**
- **More "specialization"**
  - Focused CSP accelerator and system designs for optimized performance and energy utilization
  - Capabilities (e.g., AI/GPU clouds) and outcomes (e.g., scope and complexity of workloads)
- **Build-out of new co-located hyperscaler data centers to support:**
  - New liquid cooling requirements of advanced architectures
  - Increasing number of "mid tier" service providers

HYPERION RESEARCH

Questions? We look forward to hearing from you!

mnossokoff@hyperionres.com

Today's Agenda

- **Earl Joseph, CEO**
  - HPC and AI Market Update
  - A New Way of Measuring Value of Leadership Computing
  - Tool for Deeper Understanding of Surveys Results
- **Bob Sorensen, SVP, Chief AI & QC Analyst**
  - Successfully Navigating the Changing Advanced Computing Landscape
- **Mark Nossokoff, Research Director, Chief Cloud & Storage Analyst**
  - Perspectives on HPC-AI Storage and Interconnects
  - HPC-AI Cloud Update
- **Innovation Award Winners Announcement**
- **Conclusions**

HYPERION RESEARCH

The SC24 HPC Innovation Award Winners

Examples Of Previous Winners

The Trophy For Winners

Users Must Submit the Value of the Accomplishment

Users are required to submit the value achieved with their HPC system, using 3 broad categories, following a very specific set of guidelines:

a) Dollar value of the HPC usage

- e.g., made $$$ in new revenues, saved $$$ in costs, made $$$ in profits, etc.

b) Scientific or engineering accomplishment

- e.g., discovered how xyz really works, develop a new drug that does xyz, etc.

c) Value to society as a whole

- e.g., ended nuclear testing, made something safer, provided protection against xyz, etc.

… and the investment in HPC that was required

HYPERION RESEARCH

SC24 Winners:

HPC Innovation Awards

AxoNN: Democratizing AI via Open-source Scalable AI Training

- **Organization: University of Maryland**
- **Contact: Abhinav Bhatele**
- **Innovation: AxoNN is a scalable, highly parallel AI training framework that uses a novel 4D hybrid parallel algorithm.**
  - It achieved over 620 Petaflop/s on NVIDIA A100 GPUs, 1423 Petaflop/s on H100 GPUs, and 1381 Petaflop/s on AMD MI250X GPUs.
  - As large language models (LLMs) grow, so do risks of privacy breaches due to data memorization. AxoNN highlights these concerns by exploring "catastrophic memorization," where models can memorize data in a single pass, and proposes mitigation strategies.
  - The framework also demonstrated fine-tuning a 405-billion parameter LLM on the Frontier supercomputer, addressing both performance and privacy challenges.

Closed-Loop Mars Entry Simulations with Distributed Exascale Computing

- **Organization: NASA**
- **Contact: Eric Nielsen**
- **Innovation: Human-scale Mars missions will require large landers using retropropulsion, as traditional parachutes are infeasible for the massive payloads needed.**
  - With no prior experience and limited testing possible on Earth, high-fidelity HPC simulations will be critical for developing and validating new technologies.
  - Researchers from NASA and Georgia Tech executed unprecedented, autonomous flight trajectory simulations using the Summit and Frontier platforms at Oak Ridge National Laboratory, demonstrating a viable approach for simulating future crewed Mars missions.
  - The software developed has broad applications across aerospace, defense, and academia.

Advancing Precision and Resilient Agriculture with AI, Reducing Costs and Boosting Crop Yields

- **Organization: DigiFarm**
- **Contact: Nils Helset**
- **Innovation: DigiFarm uses AI models, trained on LUMI, to accurately detect field boundaries and seeded acres, surpassing human accuracy.**
  - By super-resolving Sentinel-2 satellite imagery to 1-meter resolution and utilizing 4.7 million hectares of training data from 57 countries, DigiFarm achieves detection accuracies above 96%, which is 12-15% better than existing methods.
  - Its commercial SaaS/API solution, launched in 2022, now serves over 50 corporate clients, including Bayer, Syngenta, and various governments.
  - The solution reduces agricultural monitoring costs by over 25%, saving Lithuania, for instance, €1.5 million annually.

Conclusions

- **2024 is expected to be a strong growth year**
  - AI has fueled growth in both AI and traditional HPC
  - Non-traditional suppliers are growing fast
- **New technologies are showing up large numbers:**
  - Generative AI and LLMs
  - GPUs, 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 and the need for much larger data sets**
- **There are concerns about how the market can adapt to so many changes in GPUs and CPUs**
- **AI is scaling so fast that power and costs are redefining data centers**

HYPERION RESEARCH

We Welcome Questions, Comments and Suggestions

Please contact us at:

info@hyperionres.com